Developing a risk score for early detection of bowel cancer using routine blood test data



A thesis submitted for the degree of

Doctor of Philosophy

By

Pradeep Singh Virdee

Green Templeton College, University of Oxford

Michaelmas 2021

For Mum and Dad,

who always loved and supported me.

Contents

ACKNOWLEDGEMENTS IX

ABSTRACT XI

LIST OF FIGURES XIII

LIST OF TABLES XXIII

ABBREVIATIONS XXX

PUBLICATIONS & DISSEMINATION XXXII

STRUCTURE OF THESIS XXXV

1 INTRODUCTION 1

1.1 The problem addressed 1

1.1.1 Health problem 1

1.1.2 Impact on patients and health services 2

1.2 How this research can help 5

1.2.1 The full blood count for bowel cancer 5

1.2.2 A recent model for bowel cancer detection 7

1.2.3 The evidence gap 8

1.3 Aims & objectives 9

2 SYSTEMATIC REVIEW OF THE FBC BLOOD TEST FOR BOWEL CANCER DETECTION 11

2.1 Introduction 11

2.2 Methods 12

2.2.1 Search criteria 12

2.2.2 Study selection criteria 13

2.2.3 Screening of articles 14

2.2.4 Data extraction 14

2.2.5 Data analysis & synthesis 15

2.2.6 Risk of bias 17

2.3 Results 18

2.3.1 Description of studies 19
2.3.2 FBC for bowel cancer detection (Aim 1) 42
2.3.3 Appraisal of FBC-based prediction models (Aim 2) 62
2.3.4 Use of repeated measures 80
2.4 Discussion 81
2.4.1 FBC for bowel cancer (Aim 1) 81
2.4.2 Appraisal of FBC-based prediction models (Aim 2) 83
2.4.3 Summary of repeated measures analysis 84
2.4.4 Recommendations for future research 84
2.4.5 Strengths & limitations 88

3 PREPARATION AND QUALITY ASSESSMENT OF FBC DATA FROM CPRD 90
3.1 Introduction 90
3.2 The Clinical Practice Research Datalink 91
3.2.1 Source of data 91
3.2.2 Database structure & coding 92
3.3 The full blood count process 94
3.4 Methods 98
3.4.1 Step 1: extract FBC data 99
3.4.2 Step 2: compare medical and entity codes 100
3.4.3 Step 3: explore mismatches further 100
3.4.4 Step 4: rectify mismatches 101
3.4.5 Step 5: standardise FBC data to a single unit 101
3.4.6 Step 6: derive missing FBC levels using functional equations 101
3.4.7 Step 7: describe the resulting data 102
3.4.8 Discussions with haematologists to understand any issues 102
3.5 Results 103
3.5.1 Step 1: extract FBC data 104
3.5.2 Step 2: compare medical and entity codes 109
3.5.3 Step 3: explore mismatches further 112
3.5.4 Step 4: rectify mismatches 113
3.5.5 Step 5: standardise FBC data to a single unit 114
3.5.6 Step 6: derive missing FBC levels using functional equations 116
3.5.7 Step 7: describe the resulting data 117
3.5.8 Discussions with haematologists to understand any issues 117
3.6 Discussion 118
3.6.1 Summary of main findings 119

3.6.2 Causes of missing data 120
3.6.3 Conclusion 121
3.7 Handling missing data for subsequent analyses 122

4 TRAJECTORIES IN THE FBC TOWARDS BOWEL CANCER DIAGNOSIS 125
4.1 Introduction 125
4.2 Methods 126
4.2.1 Study design 126
4.2.2 Participants 127
4.2.3 Clinical outcome 128
4.2.4 Demographic and FBC variables 128
4.2.5 Statistical analysis 129
4.2.6 Missing data 131
4.2.7 Sensitivity analysis 131
4.3 Results 132
4.3.1 The FBC test 134
4.3.2 Bowel cancer and FBC trajectories 135
4.3.3 Bowel cancer and microcytic anaemia 140
4.3.4 Bowel cancer and FBC reference ranges 145
4.3.5 Bowel cancer staging and FBC trajectories 145
4.3.6 Sensitivity analyses 146
4.4 Discussion 147
4.4.1 Summary of main findings 147
4.4.2 Comparison with existing literature 147
4.4.3 Strengths & limitations 149
4.4.4 Conclusion and further research 151

5 A JOINT MODEL USING FBC TESTS FOR BOWEL CANCER DETECTION 153
5.1 Introduction 153
5.2 Joint modelling of longitudinal and time-to-event data 154
5.2.1 Cox proportional hazards model: 154
5.2.2 Mixed-effects model: 156
5.2.3 Joint models - linking mixed-effects and Cox models: 156
5.2.4 Dynamic predictions from joint models 158
5.2.5 Why joint models? 160
5.3 Methods 161
5.3.1 Data 161

5.3.2 Study design 161
5.3.3 Participants 165
5.3.4 Outcome 165
5.3.5 Predictors 167
5.3.6 Missing data 168
5.3.7 Model development 168
5.3.8 Model performance 170
5.3.9 Internal validation 172
5.3.10 Sample size 175
5.4 Results (development) 176
5.4.1 Summary of patient data 176
5.4.2 Joint model for two-year risk of bowel cancer 180
5.4.3 Dynamic predictions from the joint model 184
5.4.4 Performance of the joint models 188
5.5 Results (split sample internal validation) 189
5.5.1 Summary of patient data 189
5.5.2 Summary of predictions from the joint model and ColonFlag algorithm 191
5.5.3 Performance of the joint models and ColonFlag 194
5.5.4 Performance of the joint models and ColonFlag by age group 197
5.5.5 Performance of the joint models and ColonFlag by number of FBCs 202
5.6 Discussion 207
5.6.1 Summary of main findings 207
5.6.2 Comparison with existing literature 209
5.6.3 Implications for clinical practice 212
5.6.4 Strengths & limitations 213
5.6.5 Conclusion and further research 216

6 DISCUSSION 218
6.1 Summary of the thesis 218
6.1.1 Motivation 218
6.1.2 A review of the literature: is the FBC helpful and are existing models adequate? 219
6.1.3 New methodology to help assess EHR data quality 219
6.1.4 Changes in FBC levels over time are relevant for bowel cancer detection 220
6.1.5 Developing a dynamic prediction model for bowel cancer detection 221
6.2 Novelties of this research 221
6.3 Future work 223

6.4 How this research benefited me: a personal note 227
6.5 Conclusion 227

BIBLIOGRAPHY 228

APPENDIX A 245
A1. Search strategy used in the systematic review 246
A2. Data extraction forms in the systematic review 257
A3. Red blood cell count (Aim 1) 273
A4. Haemoglobin (Aim 1) 277
A5. Haematocrit (Aim 1) 295
A6. Mean corpuscular volume (Aim 1) 296
A7. Mean corpuscular haemoglobin (Aim 1) 303
A8. Mean corpuscular haemoglobin concentration (Aim 1) 304
A9. Red blood cell distribution width (Aim 1) 306
A10. Platelet count (Aim 1) 311
A11. Mean platelet volume (Aim 1) 316
A12. White blood cell count (Aim 1) 319
A13. Basophil count (Aim 1) 324
A14. Basophil proportion (Aim 1) 326
A15. Eosinophil count (Aim 1) 327
A16. Eosinophil proportion (Aim 1) 329
A17. Lymphocyte count (Aim 1) 330
A18. Lymphocyte proportion (Aim 1) 333
A19. Monocyte count (Aim 1) 334
A20. Monocyte proportion (Aim 1) 336
A21. Neutrophil count (Aim 1) 337
A22. Neutrophil proportion (Aim 1) 340
A23. Combination of parameters (Aim 1) 341
A24. Development details of FBC-based prediction models (Aim 2) 347

APPENDIX B 349
B1. Permission for re-use of work 350
B2. Functional equations for FBC parameters 352
B3. FBC entity codes with corresponding medical code =0 355
B4. Details of rectified mismatches 356
B5. Number of FBC parameters for each unit of measure 359
B6. Summary of quantiles to identify extent of extreme values per FBC parameter 379

APPENDIX C 391
C1. Biologically plausible FBC ranges 392
C2. Trajectories & FBC reference ranges 393
C3. Influence of number of FBCs on LOWESS trends 408
C4. Assessing non-linearity of time to index (years) 409
C5. Assessing non-linearity of age to index (years) 431
C6. Mixed-effects models 450
C7. Trajectories (LOWESS and modelled) in each FBC parameter 464
C8. Individualised trajectories 478
C9. Trajectories by Dukes' tumour stage 480
C10. Mixed models for Dukes' tumour stage 494
C11. Sensitivity analysis 1 – cohort study design 508
C12. Sensitivity analysis 2 – use study exit as index date in controls 511
C13. Sensitivity analysis 3 – matched design 514

APPENDIX D 517
D1. Justification for choice of candidate predictors 518
D2. Justification for choice of final longitudinal predictors 520
D3. Correlation between FBC parameters in the joint model 521
D4. Justification for using joint modelling 527
D5. Comparing functional forms for age in the Cox sub-model 530
D6. Diagnostic test accuracy in the split sample internal validation cohort 533
D7. Calibration of the ColonFlag machine-learning algorithm 540
D8. Calibration by number of FBCs for males in the development cohort 541

APPENDIX E 542
E1. How this research help me: a personal note 543

Abbreviations: FBC=full blood count.

Acknowledgements

Personal:

Firstly, I would like to thank my parents and siblings for encouraging me to pursue this DPhil, putting up with me in times of stress, supporting me throughout my career, and for always believing in me.

My friends, particularly Simran Ghataura and Vesna Zivanic, for listening to my rants, letting me bore you with all the technical “geek-emoji stuff”, always highlighting the importance of my research, keeping me motivated, and letting me practice my presentations on you.

My supervisors, Jacqueline Birks, Tim Holt, Julietta Patnick, and Peter Watkinson, for always being there when I needed support and guidance. Your encouragement, support, and enthusiasm for this research has always inspired me. Without you, this work would not have been possible. I have learned a lot from you over the last few years so know that you will be at the heart of my future successes. Your wealth of knowledge has shaped my work for the better.

Sharon Love, Vicky Strauss, and Joanna Moschandreas for introducing me to the world of research and encouraging me to do this DPhil. My time with you in CSM has given me the tools and skills I needed to do this DPhil.

Finally, to everyone else at CSM for supporting throughout my DPhil journey. Particularly, Leena Elhussein, Ioana Marian, Stephen Gerry, Bethan Copsey, and Jennifer de Beyer for all our chats and keeping me excited about this research. Also, others at NDORMS, particularly Afsie Sabokbar and Samuel Burnell, for supporting me throughout my DPhil.

Technical:

This work was funded by the National Institute for Health Research Doctoral Fellowship programme (DRF-2018-11-ST2-057). This report presents independent research and the views expressed are those of the research team and not necessarily those of the funders, NHS, or Department of Health and Social Care.

The Research Design Service based in Oxford for providing input on my NIHR grant application that would fund the research in this DPhil. Also to Stephen Gerry, Ioana Marian, Paula Dhiman, Leena Elhussein, and Jen de Beyer for the multiple mock interviews that better equipped me for my NIHR interview.

Thesis Oversight Committee members, Vicky Strauss, Lei Clifton, and Sophie Hyndman, for independent input into the design, conduct, and reporting of this research.

Patient and public involvement representatives, Pete Wheatstone, Margaret Ogden, and Julian Ashton, for applying a patient perspective to this work through regular meetings and for reviewing study outputs.

Marialena Trivella for providing independent input to the analysis strategy for the systematic review in Chapter 2 and to Stephen Gerry for letting me work on your systematic review, giving me relevant experience that helped me develop my own.

Shona Kirtley, who helped prepare the search criteria for the systematic review, and Ioana Marian, Anita Mansouri, and Leena Elhussein, who performed eligibility screening and data extraction of papers in the systematic review.

Collaborators in Chapter 3, Alice Fuller, who helped prepare the CPRD data, and Michael Jacobs, who provided oversight in understanding how blood tests are processed in haematology laboratories and giving me a tour of John Radcliffe haematology laboratory. In addition, to Simran Ghataura, who discussed her personal experience on how blood tests work in haematology laboratories.

Michael Jacobs for providing a biologically plausible range of values that FBC levels can take, used in Chapters 3, 4, and 5.

Vesna Zivanic for recommending available high-processing computers, providing advice on appropriate computing specifications, and discussing computing power needed to build the model in Chapter 5. Additionally, the HAVEN team, who helped me gain access to the advanced computational facility at the Kadoorie Centre in the John Radcliffe Hospital, which was used to build and test the prediction model (Chapters 5). Particularly, Robert Hatch for technical assistance.

Abstract

Introduction

Bowel cancer is the fourth most common type of cancer in the UK. Approximately 55% of patients are diagnosed with a late-stage tumour (Stage 3 and 4), where likelihood of survival is reduced: five-year survival is 93% at Stage 1 vs. 10% at stage 4. Early detection is crucial. The full blood count (FBC) blood test is commonly performed in primary care. It may help but currently there is limited evidence on how to utilise this test for early detection.

Methods

I performed four research objectives. I reviewed relevant research studies in the literature to learn about the relevance of the FBC and eliminate some key uncertainties in existing approaches for early detection. I performed a data quality and validation assessment of electronic health record (EHR) patient data. I then performed longitudinal analyses to identify patterns in the FBC over time, which could help identify presence of bowel cancer. I subsequently incorporated my findings into a dynamic prediction model that could facilitate earlier detection and improve survival rates.

Results

My systematic review identified many relevant FBC parameters for bowel cancer detection, such as haemoglobin, mean corpuscular volume, and platelets, and that all existing FBC-based prediction models are flawed. Using primary care data from the Clinical Practice Research Datalink, I identified relevant trajectories in many FBC parameters in patients with bowel cancer, not observed in patients without. Relevant differences in FBC levels occurred for years before diagnosis, suggesting the FBC could facilitate earlier detection. I incorporated the trajectories in haemoglobin, mean corpuscular volume, and platelets into multivariate joint models of

longitudinal and time-to-event data to identify two-year risk of diagnosis. My joint models perform well based on my initial testing.

Discussion

My joint models use routinely available data and would not require additional work for healthcare workers or patients. These models perform well based on my testing, but further testing is required before use in general practice. My models could facilitate earlier detection and improve survival rates.

List of Figures

Main text

Figure 1.1: The Faecal Immunochemical Test for traces of blood in stool samples* 3

Figure 1.2: Colonoscopy – a colonoscope inserted into the anus and directed through the bowel* 4

Figure 2.1: PRISMA diagram* 19

Figure 2.2: Forest plot of mean difference in haemoglobin (g/dL) between patients with and without a diagnosis of bowel cancer up to six months later* 49

Figure 2.3: Forest plot of mean difference in platelet count (10^9/L) between patients with and without a diagnosis of bowel cancer up to six months later* 53

Figure 2.4: Forest plot of mean difference in mean platelet volume (fL) between patients with and without a diagnosis of bowel cancer up to six months later* 54

Figure 2.5: Forest plot of mean difference in white blood cell count (10^9/L) between patients with and without a diagnosis of bowel cancer up to six months later* 56

Figure 2.6: Forest plot of mean difference in lymphocyte count (10^9/L) between patients with and without a diagnosis of bowel cancer up to six months later 58

Figure 3.1: The Faecal Immunochemical Test for traces of blood in stool samples[1] 93

Figure 3.2: The full blood count process 95

Figure 3.3: Example of haematology request paper form – includes the full blood count[1] 97

Figure 3.4: Example of a (dummy) FBC report[1] 98

Figure 3.5: Methodological approach to assess laboratory data quality in CPRD 99

Figure 3.6: Number of different units entered for each FBC parameter (n=769 in total) 114

Figure 4.1: Patient flow diagram (case-control analysis to identify trajectories) 133

Figure 4.2: Histogram of follow-up time from earliest FBC in the 10-year period to index date[1] 136

Figure 4.3: Histogram of time between latest FBC in the 10-year period and index date[1] 136

Figure 4.4: Trajectories (LOWESS and fixed-effects) in haemoglobin levels (g/dL) for males (left) and females (right) with and without a diagnosis of bowel by age at index date[1, 2] 138

Figure 4.5: Trajectories (LOWESS and fixed-effects) in platelet count (10^9/L) for males (left) and females (right) with and without a diagnosis of bowel by age at index date[1, 2] 139

Figure 4.6: Microcytic anaemia prevalence over 6-month time bands before index date between patients with and without bowel cancer – males (top) and females (bottom) 140

Figure 4.7: Trajectories (LOWESS) in haemoglobin (left) and mean corpuscular volume (right) by microcytic anaemia threshold for males, by bowel cancer status and age at index[1, 2, 3] 143

Figure 4.8: Trajectories (LOWESS) in haemoglobin (left) and mean corpuscular volume by microcytic anaemia threshold for females, by bowel cancer status and age at index[1, 2, 3] 144

Figure 5.1: Flow of (dummy) longitudinal data for two-year risk of diagnosis 162

Figure 5.2: Replicating trajectories (LOWESS) in the prior case-control analysis (left) in this joint model development cohort study (right)[1, 2, 3] 164

Figure 5.3: Patient flow diagram (cohort study to develop joint model) 178

Figure 5.4: Distribution of dynamic risk predictions for males (left) and females (right) in the derivation cohort 186

Figure 5.5: Dynamic prediction plots for a female diagnosed (top) and not diagnosed (bottom)[1] 187

Figure 5.6: Calibration plot for the joint models for males (left) and females (right) in the development cohort[1] 189

Figure 5.7: Distribution of dynamic two-year risk predictions from the joint models for males (left) and females (right) in the split sample internal validation cohort 193

Figure 5.8: Distribution of ColonFlag two-year score for males (left) and females (right) in the split sample internal validation cohort 194

Figure 5.9: ROC curves for two-year risk from the joint model and ColonFlag for males (left) and females (right) in the split sample internal validation cohort 196

Figure 5.10: Calibration plot for the joint models in males (left) and females (right) in the split sample internal validation cohort 197

Figure 5.11: C-statistic for the joint model and ColonFlag algorithm by age (at baseline) group in males (split sample internal validation cohort) 200

Figure 5.12: C-statistic for the joint model and ColonFlag algorithm by age (at baseline) group in females (split sample internal validation cohort) 201

Figure 5.13: Calibration plot for the joint models by age (at baseline) group in males (left) and females (right) in the split sample internal validation cohort 201

Figure 5.14: C-statistic for the joint model and ColonFlag algorithm by number of FBCs in males (split sample internal validation cohort) 205

Figure 5.15: C-statistic for the joint model and ColonFlag algorithm by number of FBCs in females (split sample internal validation cohort) 206

Figure 5.16: C-statistic for the joint model and ColonFlag algorithm by number of FBCs in females aged 70-89 years at baseline FBC (split sample internal validation cohort) 206

Figure 5.17: Calibration plot for the joint models by number of FBCs in males (left) and females (right) in the split sample internal validation cohort 207

Abbreviations: FBC=full blood count; ROC=receiver operating characteristic.

Appendix C

Figure C1: Trajectories (LOWESS) in red blood cell count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 394

Figure C2: Trajectories (LOWESS) in haemoglobin for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 395

Figure C3: Trajectories (LOWESS) in haematocrit for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2,3] 396

Figure C4: Trajectories (LOWESS) in mean corpuscular volume for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 397

Figure C5: Trajectories (LOWESS) in mean corpuscular haemoglobin for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 398

Figure C6: Trajectories (LOWESS) in mean corpuscular haemoglobin concentration for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 399

Figure C7: Trajectories (LOWESS) in platelets for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 400

Figure C8: Trajectories (LOWESS) in mean platelet volume for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2, 3] 401

Figure C9: Trajectories (LOWESS) in white blood cell count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 402

Figure C10: Trajectories (LOWESS) in basophil count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2, 3] 403

Figure C11: Trajectories (LOWESS) in eosinophil count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 404

Figure C12: Trajectories (LOWESS) in lymphocyte count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 405

Figure C13: Trajectories (LOWESS) in monocyte count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 406

Figure C14: Trajectories (LOWESS) in neutrophil count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2] 407

Figure C15: Trajectories (LOWESS) in red blood cell count (top) and haemoglobin (bottom) for male controls aged 70 years at index date, influenced by the number of FBCs[1, 2] 408

Figure C16: Trajectories (LOWESS) in red blood cell count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 409

Figure C17: Trajectories (LOWESS) in haemoglobin for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 410

Figure C18: Trajectories (LOWESS) in haematocrit for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 411

Figure C19: Trajectories (LOWESS) in mean corpuscular volume for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 412

Figure C20: Trajectories (LOWESS) in mean corpuscular haemoglobin for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 413

Figure C21: Trajectories (LOWESS) in mean corpuscular haemoglobin concentration for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 414

Figure C22: Trajectories (LOWESS) in platelet count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 415

Figure C23: Trajectories (LOWESS) in mean platelet volume for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 416

Figure C24: Trajectories (LOWESS) in white blood cell count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 417

Figure C25: Trajectories (LOWESS) in basophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 418

Figure C26: Trajectories (LOWESS) in eosinophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 419

Figure C27: Trajectories (LOWESS) in lymphocyte count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 420

Figure C28: Trajectories (LOWESS) in monocyte count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 421

Figure C29: Trajectories (LOWESS) in neutrophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 422

Figure C30: Trajectories (LOWESS) in red blood cell count for male controls by age at index date[1, 2] 430

Figure C31: Red blood cell count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 431

Figure C32: Haemoglobin by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 432

Figure C33: Haematocrit by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 433

Figure C34: Mean corpuscular volume by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 434

Figure C35: Mean corpuscular haemoglobin by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 435

Figure C36: Mean corpuscular haemoglobin concentration by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 436

Figure C37: Platelet count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 437

Figure C38: Mean platelet count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 438

Figure C39: White blood cell count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 439

Figure C40: Basophil count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 440

Figure C41: Eosinophil count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 441

Figure C42: Lymphocyte count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 442

Figure C43: Monocyte count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 443

Figure C44: Neutrophil count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2] 444

Figure C45: Trajectories (LOWESS and fixed-effects) in red blood cell count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 464

Figure C46: Trajectories (LOWESS and fixed-effects) in haemoglobin for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 465

Figure C47: Trajectories (LOWESS and fixed-effects) in haematocrit for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 466

Figure C48: Trajectories (LOWESS and fixed-effects) in mean corpuscular volume for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 467

Figure C49: Trajectories (LOWESS and fixed-effects) in mean corpuscular haemoglobin for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 468

Figure C50: Trajectories (LOWESS and fixed-effects) in mean corpuscular haemoglobin concentration for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 469

Figure C51: Trajectories (LOWESS and fixed-effects) in platelets for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 470

Figure C52: Trajectories (LOWESS and fixed-effects) in mean platelet volume for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 471

Figure C53: Trajectories (LOWESS and fixed-effects) in white blood cell count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 472

Figure C54: Trajectories (LOWESS and fixed-effects) in basophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 473

Figure C55: Trajectories (LOWESS and fixed-effects) in eosinophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 474

Figure C56: Trajectories (LOWESS and fixed-effects) in lymphocyte count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 475

Figure C57: Trajectories (LOWESS and fixed-effects) in monocyte count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 476

Figure C58: Trajectories (LOWESS and fixed-effects) in neutrophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2] 477

Figure C59: Individualised haemoglobin (left) and platelet (right) predicted levels for four cases at different ages at index date[1] 478

Figure C60: Individualised haemoglobin (left) and platelet (right) predicted levels for four controls at different ages at index date[1] 479

Figure C61: Trajectories (LOWESS and fixed-effects) in red blood cell count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 480

Figure C62: Trajectories (LOWESS and fixed-effects) in haemoglobin for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 481

Figure C63: Trajectories (LOWESS and fixed-effects) in haematocrit for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 482

Figure C64: Trajectories (LOWESS and fixed-effects) in mean corpuscular volume for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 483

Figure C65: Trajectories (LOWESS and fixed-effects) in mean corpuscular haemoglobin for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 484

Figure C66: Trajectories (LOWESS and fixed-effects) in mean corpuscular haemoglobin concentration for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 485

Figure C67: Trajectories (LOWESS and fixed-effects) in platelets for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 486

Figure C68: Trajectories (LOWESS and fixed-effects) in mean platelet volume for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 487

Figure C69: Trajectories (LOWESS and fixed-effects) in white blood cell count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 488

Figure C70: Trajectories (LOWESS and fixed-effects) in basophil count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 489

Figure C71: Trajectories (LOWESS and fixed-effects) in eosinophil count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 490

Figure C72: Trajectories (LOWESS and fixed-effects) in lymphocyte count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 491

Figure C73: Trajectories (LOWESS and fixed-effects) in monocyte count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 492

Figure C74: Trajectories (LOWESS and fixed-effects) in neutrophil count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3] 493

Figure C75: Trajectories (LOWESS) in haemoglobin for males with and without bowel cancer – cohort (left) vs. case-control (right) design[1, 2, 3] 509

Figure C76: Trajectories (LOWESS) in platelets for males with and without bowel cancer – cohort (left) vs. case-control (right) design[1, 2, 3] 510

Figure C77: Trajectories (LOWESS) in red blood cell count for males by index date definition – random (left) vs. study exit (middle) with and without deaths (right)[1, 2] 512

Figure C78: Trajectories (LOWESS) in platelets for males with and without bowel cancer – random (left) vs. study exit (right) index date in controls[1, 2] 513

Figure C79: Haemoglobin trends by age at index[1, 2]: unmatched (left) and matched (right) ... 515

Figure C80: Platelet trends by age at index[1, 2]: unmatched (left) and matched (right) 516

Abbreviations: FBC=full blood count.

Appendix D

Figure D1: Relationships between haemoglobin and MCV at diagnosis/censor date (left), 5 (middle), and 10 (right) years prior for cases (top) and controls (bottom)[1] 522

Figure D2: Relationships between haemoglobin and platelets at diagnosis/censor date (left), 5 (middle), and 10 (right) years prior for cases (top) and controls (bottom)[1] 522

Figure D3: Relationships between MCV and platelets at diagnosis/censor date (left), 5 (middle), and 10 (right) years prior for cases (top) and controls (bottom)[1] 523

Figure D4: Trends in the Spearman correlation coefficient and 95% CI (shaded area) between haemoglobin and mean corpuscular volume towards index date[1] 525

Figure D5: Trends in the Spearman correlation coefficient and 95% CI (shaded area) between haemoglobin and platelets towards index date[1] 525

Figure D6: Trends in the Spearman correlation coefficient and 95% CI (shaded area) between mean corpuscular volume and platelets towards index date[1] 526

Figure D7: LOWESS curve of bowel cancer presence by age at index FBC to assess functional form for males (left) and females (right) 530

Figure D8: Schoenfeld residuals to assess the proportional hazards assumption for age (fractional polynomial powers: 2, 2) for males (left) and females (right)..... 532

Figure D9: Calibration plot for the ColonFlag algorithm in males (left) and females (right) in the split sample internal validation cohort 540

Figure D10: Calibration plot for the joint models by number of FBCs for males in the development cohort 541

Abbreviations: MCV=mean corpuscular volume.

List of Tables

Main text

Table 1.1: Description of FBC parameters* 5

Table 2.1: Description of each study (n=53)* 21

Table 2.2: Full blood count parameters analysed in each study included in Aim 1 (n=47)* 43

Table 2.3: Risk of bias (assessed using the QUIPS tool) in each study included in Aim 1 (n=47)* 46

Table 2.4: Details of FBC-based prediction models developed for bowel cancer detection (n=13)* 63

Table 2.5: Risk of bias (assessed using the PROBAST tool) in each study included in Aim 2 (n=16)* 66

Table 2.6: Performance statistics from internal (n=9) and external (n=11) validation of 13 FBC-based models (Aim 2)* 73

Table 2.7: Considerations of existing FBC-based prediction models developed for bowel cancer detection (n=13)* 78

Table 3.1: Entity and medical codes for the FBC blood test in CPRD (Test dataset)..... 104

Table 3.2: Number of FBC parameters at different stages of the quality check..... 110

Table 3.3: Description of mismatched pairs..... 111

Table 3.4: Descriptive statistics of each FBC parameter in the final derived dataset..... 115

Table 4.1: ICD-10 codes for bowel cancer to identify cases in the NCRAS 126

Table 4.2: Characteristics of the patient sample 133

Table 4.3: Summary of missing data for full blood count parameters 134

Table 4.4: Age-adjusted odds ratios for diagnosis corresponding to presence of microcytic anaemia at 6-monthly (+/- 3 months) time intervals prior to index date[1] 141

Table 4.5: Number of cases per Dukes' tumour stage at diagnosis 145

Table 5.1: Summary of patient characteristics (joint model development) 179

Table 5.2: Summary of FBC data and follow-up (joint model development) 179

Table 5.3: Mixed-effects sub-models from the joint models (fixed effects)..... 181

Table 5.4: Mixed-effects sub-models from the joint-models (random effects: variance-covariance matrix[1]) 183

Table 5.5: Cox sub-models from the joint models 184

Table 5.6: Performance measures of the joint model (joint model development) 188

Table 5.7: Characteristics of patients with and without a ColonFlag score (split sample internal validation) 190

Table 5.8: Summary of patient characteristics (split sample internal validation) 191

Table 5.9: Summary of FBC data and follow-up (split sample internal validation)..... 191

Table 5.10: Performance measures of the joint models and ColonFlag (split sample internal validation) 195

Table 5.11: Diagnostic test accuracy measures of the joint models for risk percentiles of 75% and above (split sample internal validation)..... 196

Table 5.12: Summary of FBC data and follow-up by age (at baseline) group (split sample internal validation) 198

Table 5.13: Summary of patient characteristics, FBC data, and follow-up by number of FBCs (split sample internal validation) 202

Abbreviations: FBC=full blood count; CPRD=Clinical Practice Research Datalink.

Appendix A

Table A1: Red blood cell count for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 273

Table A2: Haemoglobin levels for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 277

Table A3: Haematocrit for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 295

Table A4: Mean corpuscular volume for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 296

Table A5: Mean corpuscular haemoglobin for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 303

Table A6: Mean corpuscular haemoglobin concentration for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 304

Table A7: Red blood cell distribution width for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 306

Table A8: Platelet count for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 311

Table A9: Mean platelet volume for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 316

Table A10: White blood cell count for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 319

Table A11: Basophil count for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 324

Table A12: Basophil proportion for bowel cancer 326

Table A13: Eosinophil count for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 327

Table A14: Eosinophil proportion for bowel cancer 329

Table A15: Lymphocyte count for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 330

Table A16: Lymphocyte proportion for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 333

Table A17: Monocyte count for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 334

Table A18: Monocyte proportion for bowel cancer 336

Table A19: Neutrophil count for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 337

Table A20: Neutrophil proportion for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 340

Table A21: Parameters combined for bowel cancer (analyses sorted by outcome time window, country, and then strata)* 341

Table A22: Development details of FBC-based prediction models..... 347

Abbreviations: FBC=full blood count.

Appendix B

Table B1: Number of FBC parameters assigned medical code=0 (i.e. missing) 355

Table B2: Number of rectified and non-rectified mismatches in strata 1[1] 356

Table B3: Number of rectified and non-rectified mismatches in strata 2[1] 356

Table B4: Number of rectified and non-rectified mismatches in strata 3[1] 357

Table B5: Number of red blood cell count records per unit of measure 359

Table B6: Number of white blood cell count records per unit of measure entered..... 360

Table B7: Number of haemoglobin records per unit of measure entered 362

Table B8: Number of haematocrit records per unit of measure entered..... 364

Table B9: Number of mean corpuscular volume records per unit of measure entered..... 365

Table B10: Number of mean corpuscular haemoglobin records per unit of measure entered 366

Table B11: Number of mean corpuscular haemoglobin concentration records per unit of measure entered..... 367

Table B12: Number of red blood cell distribution width records per unit of measure entered 368

Table B13: Number of platelet records per unit of measure entered 369

Table B14: Number of mean platelet volume records per unit of measure entered 370

Table B15: Number of basophil count records per unit of measure entered..... 370

Table B16: Number of basophil proportion records per unit of measure entered 372

Table B17: Number of eosinophil count records per unit of measure entered..... 372

Table B18: Number of eosinophil proportion records per unit of measure entered 373

Table B19: Number of lymphocyte count records per unit of measure entered 373

Table B20: Number of lymphocyte proportion records per unit of measure entered 375

Table B21: Number of monocyte count records per unit of measure entered 375

Table B22: Number of monocyte proportion records per unit of measure entered..... 376

Table B23: Number of neutrophil count records per unit of measure entered 376

Table B24: Number of neutrophil proportion records per unit of measure entered 378

Table B25: Summary statistics of red blood cell count (10 percentiles) entered as 10^{12}/L..... 379

Table B26: Summary statistics of white blood cell count (10 percentiles) entered as 10^{9}/L ... 379

Table B27: Summary statistics of haemoglobin (10 percentiles) entered as g/dL..... 380

Table B28: Summary statistics of haematocrit (10 percentiles) entered as L/L 381

Table B29: Summary statistics of mean corpuscular volume (10 percentiles) entered as fL ... 381

Table B30: Summary statistics of mean corpuscular haemoglobin (10 percentiles) entered as fL 382

Table B31: Summary statistics of mean corpuscular haemoglobin concentration (10 percentiles) entered as g/dL 383

Table B32: Summary statistics of red blood cell distribution width (10 percentiles) entered as proportions..... 383

Table B33: Summary statistics of platelet (10 percentiles) entered as 10^{9}/L 384

Table B34: Summary statistics of mean platelet volume (10 percentiles) entered as fL..... 385

Table B35: Summary statistics of basophil count (4 percentiles) entered as 10^9/L..... 385

Table B36: Summary statistics of basophil proportion (4 percentiles) entered as proportions 386

Table B37: Summary statistics of eosinophil count (4 percentiles) entered as 10^9/L 386

Table B38: Summary statistics of eosinophil proportion (4 percentiles) entered as proportions 387

Table B39: Summary statistics of lymphocyte count (4 percentiles) entered as 10^9/L 387

Table B40: Summary statistics of lymphocyte proportion (4 percentiles) entered as proportions 388

Table B41: Summary statistics of monocyte count (4 percentiles) entered as 10^9/L..... 388

Table B42: Summary statistics of monocyte proportion (4 percentiles) entered as proportions 388

Table B43: Summary statistics of neutrophil count (4 percentiles) entered as 10^9/L 389

Table B44: Summary statistics of neutrophil proportion (4 percentiles) entered as proportions 389

Abbreviations: FBC=full blood count.

Appendix C

Table C1: Biologically plausible ranges for each FBC parameter 392

Table C2: Reference range for patients aged 13+ years for each FBC parameter[1]..... 393

Table C3: Akaike and Bayesian information criteria to compare methods for non-linearity of time to index date[1, 2, 3] 423

Table C4: Akaike and Bayesian information criteria to compare methods for non-linearity of age at index date[1, 2, 3] 445

Table C5: Mixed-effect model coefficients (95% CI) for red blood cell count parameters..... 450

Table C6: Mixed-effect model coefficients (95% CI) for platelet parameters 457

Table C7: Mixed-effect model coefficients (95% CI) for white blood cell parameters 459

Table C8: Mixed-effect model coefficients (95% CI) to assess red blood cell parameters by Dukes' Stage A vs. D 494

Table C9: Mixed-effect model coefficients (95% CI) to assess platelet parameters by Dukes' Stage A vs. D 500

Table C10: Mixed-effect model coefficients (95% CI) to assess white blood cell parameters by Dukes' Stage A vs. D 503

Abbreviations: FBC=full blood count.

Appendix D

Table D1: Summary of number of patients per correlation pair..... 523

Table D2: Summary of Spearman correlation coefficients per correlation pair 524

Table D3: AIC and BIC of models using various functional forms for age at baseline FBC in the Cox sub-model of a joint model[1]..... 531

Table D4: Diagnostic test accuracy measures of the joint models for males (split sample internal validation) 533

Table D5: Diagnostic accuracy measures of the joint models for females (split sample internal validation) 536

Abbreviations: AIC=Akaike information criteria; BIC=Bayesian information criteria; FBC=full blood count.

Abbreviations

AUC: area under the curve

BLUP: best linear unbiased predictor

CPRD: Clinical Practice Research Datalink

CSM: Centre for Statistics in Medicine

FBC: full blood count

GP: General Practitioner

GPRD: General Practice Research Datalink

GPSoC: GP Systems of Choice

HR: hazard ratio

ICD-10: International Classification of Diseases-10

MAR: missing at random

MCAR: missing completely at random

MNAR: missing not at random

MPVPR: mean platelet volume-platelet ratio

NCRAS: National Cancer Registration and Analysis Service

NHS: National Health Service

NICE: National Institute for Health and Care Excellence

NLR: neutrophil-lymphocyte

NPV: negative predictive value

OR: odds ratio

PLR: platelet-lymphocyte ratio

PPI: patient and public involvement

PPV: positive predictive value

PRISMA: Preferred Reporting Items for Systematic review and Meta-Analysis

PROBAST: Prediction model Risk Of Bias ASsessment Tool

QUIPS: Quality In Prognosis Studies

RDW: red blood cell distribution width

RDWLR: red blood cell distribution width-lymphocyte ratio

RECORD: REporting of studies Conducted using Observational Routinely-collected health Data

ROC: receiver operating characteristic

SD: standard deviation

SE: standard error

STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

TRIPOD: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

WHO: World Health Organisation

Publications & dissemination

DPhil-related dissemination is described here.

Publications:

Chapter 2 (systematic review):

Pradeep S. Virdee, Shona Kirtley, Leena Elhussein, Peter Watkinson, Tim Holt, Jacqueline Birks. Components of the full blood count as risk factors for colorectal cancer detection: a systematic review protocol. *BMJ Open, 9:e032759*. (2019). DOI: 10.1136/bmjopen-2019-032759.

Pradeep S. Virdee, Ioana Marian, Anita Mansouri, Leena Elhussein, Shona Kirtley, Tim Holt, Jacqueline Birks. The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. (2020). DOI: 10.3390/cancers12092348.

Chapter 3 (methodology to asses EHR data quality):

Pradeep S. Virdee, Alice Fuller, Michael Jacobs, Tim Holt, Jacqueline Birks. Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data, 7, 95:1-18*. (2020). DOI: 10.1186/s40537-020-00375-w.

Chapter 4 (trajectories in the FBC):

Pradeep S. Virdee, Julietta Patnick, Peter Watkinson, Jacqueline Birks, Tim Holt. Trends in the full blood count blood test and colorectal cancer detection: a longitudinal, case-control study of UK primary care patient data. *****Recently submitted to the British Journal of Cancer (pending review)*****

Conference materials:

Chapter 2 (systematic review):

Pradeep S. Virdee, Ioana Marian, Anita Mansouri, Leena Elhussein, Shona Kirtley, Tim Holt, Jacqueline Birks. The full blood count blood test for colorectal cancer detection: a systematic review, meta-analysis, and critical appraisal. ***Poster presentation*** at the National Cancer Research Institute (NCRI) 2021 conference. Available online at https://abstracts.ncri.org.uk/abstract/the-full-blood-count-blood-test-for-colorectal-cancer-

detection-a-systematic-review-meta-analysis-and-critical-appraisal/. Last accessed 25th November 2021.

Chapter 3 (methodology to assess EHR data quality):

Pradeep S. Virdee, Alice Fuller, Michael Jacobs, Tim Holt, Jacqueline Birks. A methodological approach to assess data quality from the Clinical Practice Research Datalink. ***Poster presentation*** at the International Society for Clinical Biostatistics (ISCB) 2021 conference.

Chapter 4 (trends over time):

Pradeep S. Virdee, Tim Holt, Julietta Patnick, Jacqueline Birks. Patterns in haemoglobin levels over 10 years to predict diagnosis of colorectal cancer by Dukes' staging: preliminary findings using UK primary care routine blood test data. ***Poster presentation*** at the National Cancer Research Institute (NCRI) 2018 conference. Available at <https://abstracts.ncri.org.uk/abstract/patterns-in-haemoglobin-levels-over-10-years-to-predict-diagnosis-of-colorectal-cancer-by-dukes-staging-preliminary-findings-using-uk-primary-care-routine-blood-test-data/>. Last accessed 6th December 2020.

Pradeep S. Virdee, Tim Holt, Julietta Patnick, Jacqueline Birks. Patterns in haemoglobin levels over 10 years to predict diagnosis of colorectal cancer by Dukes' staging: preliminary findings using UK primary care routine blood test data. ***Poster presentation*** at the Young Statistician's Meeting (YSM) 2018 conference.

Pradeep S. Virdee, Tim Holt, Jacqueline Birks. Trends over time in the full blood count blood test for colorectal cancer detection. ***Poster presentation*** at the Early Detection conference (EdX) 2018 conference.

Chapter 5 (joint model development):

Pradeep S. Virdee, Jacqueline Birks, Tim Holt (2021). A dynamic prediction model for early detection of colorectal cancer using routine blood test results from primary care. ***Oral presentation*** at the Society for Academic Primary Care (SAPC) 2021 conference. Available online at https://sapc.ac.uk/conference/2021/abstract/dynamic-prediction-model-early-detection-of-colorectal-cancer-using-routine. DOI: 10.37361/asm.2021.1.1

Pradeep S. Virdee, Jacqueline Birks, Tim Holt (2021). A dynamic prediction model for early detection of colorectal cancer using routine blood test results from primary care. ***Poster***

presentation at the National Cancer Research Institute (NCRI) 2021 conference. Available online at https://abstracts.ncri.org.uk/abstract/a-dynamic-prediction-model-for-early-detection-of-colorectal-cancer-using-routine-blood-test-results-from-primary-care/. Last accessed 25th November 2021.

Other:

Pradeep S. Virdee. Incorporating patient and public involvement into medical research and grant applications: approach, perspective, and lessons learnt from an early-career statistician. ***Oral presentation*** at the Young Statistician's Meeting (YSM) 2018 conference.

Blogs & interviews:

Chapter 2 (systematic review):

Clarivate Web of Science. Young statistician aims to predict the risk of deadly cancer with a blood test. (2020). Available at <https://clarivate.com/webofsciencegroup/article/young-statistician-aims-to-predict-the-risk-of-deadly-cancer-with-a-blood-test/>. Last accessed 6th December 2020.

NDORMS. Full blood count blood test could help detect early stage bowel cancer. (2020). Available at <https://www.ndorms.ox.ac.uk/news/blog/full-blood-count-blood-test-could-help-detect-early-stage-bowel-cancer>. Last accessed 15th December 2020.

Structure of thesis

Chapter 1: this chapter introduces the clinical problem of bowel cancer, the current UK screening programme, the ability of the full blood count blood test to detect bowel cancer, and describes the aims and objectives of the research.

Chapter 2: this chapter reports a systematic review performed to assess the prognostic ability of the full blood count blood test assessed by individual studies. Existing prediction models that use this blood test for bowel cancer risk are critically appraised and individual studies critiqued.

Chapter 3: this chapter presents my approach to preparing and assessing the quality of full blood count test results stored in a nationwide electronic health record database. This includes the process of entering and downloading data from practice into the database.

Chapter 4: this chapter reports findings of a case-control study assessing trajectories in each full blood count parameter over a 10-year period for those with and without a diagnosis separately.

Chapter 5: this chapter presents a new prediction model that incorporates trajectories among repeated full blood counts to identify two-year risk of diagnosis of bowel cancer. The model was built using relatively new methodology, joint modelling of longitudinal and time-to-event data, combined with findings from all previous chapters. Model performance was compared to the ColonFlag algorithm, an existing model using similar inputs, and assessed in relevant subgroups.

Chapter 6: this chapter provides an overall discussion of the research, including summarising the key findings and novelties of this work, future work planned, and an overall conclusion.

1

Introduction

1.1 The problem addressed

1.1.1 Health problem

Bowel cancer (also referred to as colon, rectal, or colorectal cancer) is the fourth most common type of cancer in the UK. Around 42,300 new cases were diagnosed each year in 2015-2017, accounting for 11% of new cancer cases per year[1]. Bowel cancer is the second most common cause of death by cancer in the UK. Around 16,300 patients died each year in 2015-2017 due to bowel cancer[2]. Survival is associated with tumour stage at diagnosis[3-5]. Five-year survival is approximately 93% if the tumour is diagnosed at stage 1 (earliest stage), where it is confined to the bowel lining and easiest to treat, but reduces to 10% if diagnosed at stage 4 (latest stage), where it has spread to other organs[6, 7]. Approximately 55% of bowel cancers in the UK are diagnosed late-stage, where the likelihood of survival is low[7]. Early detection of bowel cancer can save many lives.

Most symptoms for bowel cancer are non-specific and include abdominal pain, change in bowel habit, and rectal bleeding[8-12]. Symptoms often appear in patients when the disease has developed to a late-stage and are commonly first reported to a general practitioner (GP) within six months prior to diagnosis[13]. Identifying cancer at earlier stages, before overt symptoms appear, would be of considerable benefit.

1.1.2 Impact on patients and health services

The National Health Service (NHS) offer a screening programme to detect bowel cancer, consisting of faecal blood tests and bowel scope assessments[14, 15]. The Faecal Immunochemical Test (FIT) examines stool samples for traces of blood[16] (Figure 1.1). Sensitivity (proportion of cases tested positive) of FIT is around 85% overall, but is lower at 68% for stage 1 bowel tumours[17, 18]. The impact of FIT was limited until the last 12 months because it was only offered routinely to adults aged 60-74 years in England every two years, with incidence rates in younger adults aged 20-50 years increasing over time (55 per 100,000 diagnosed in 2005-2007 vs. 57 per 100,000 diagnosed in 2015-2017[19-22]. Additionally, participation rate is around 55-65% in those offered the test[23, 24], which is considered low by the bowel cancer community[25]. However, over the last 12 months, FIT testing has proved a useful test outside the screening programme for use in patients with low-risk symptoms. Negative findings on FIT avoids the need for colonoscopy, based on a 98% negative predictive value in a recent primary care study[26]. However, these patients are still symptomatic, which, as described earlier, is problematic for early detection.

Figure 1.1: The Faecal Immunochemical Test for traces of blood in stool samples*

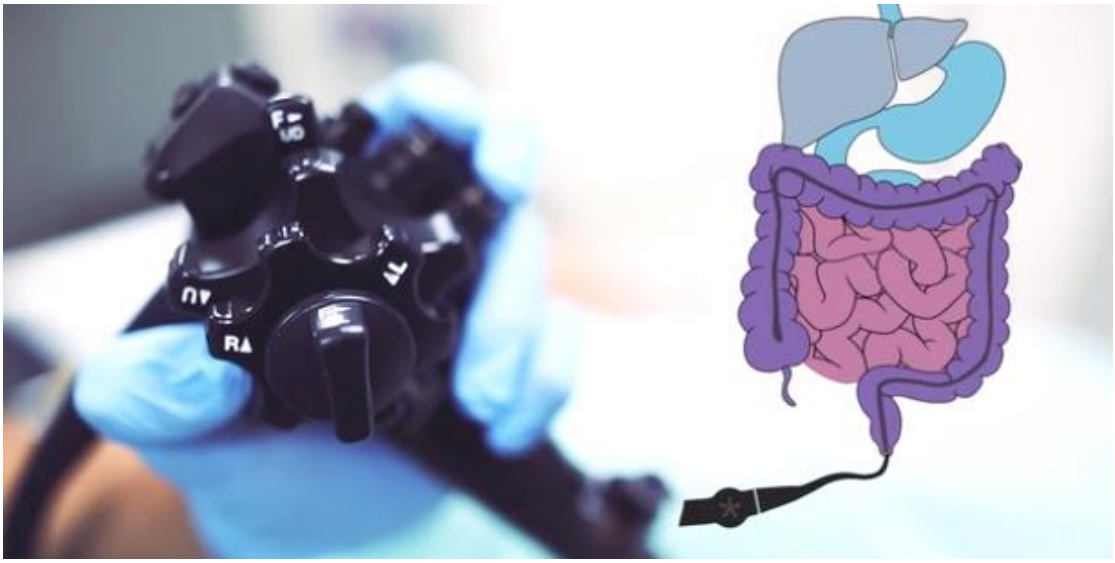


*This figure is taken from: GOV.UK. NHS bowel cancer screening: helping you decide (FIT kit). 2021 [Accessed 1 August 2021; Available from: https://www.gov.uk/government/publications/bowel-cancer-screening-benefits-and-risks/nhs-bowel-cancer-screening-helping-you-decide[27]. Published under an Open Government Licence V3.0 – permissions not required.

Bowel scope screening (or flexible sigmoidoscopy) involves inserting a thin tube with a camera through the anus[28, 29]. The camera is used to detect polyps, which are benign tumours that may become malignant over time, and could be removed during the assessment. This assessment is limited in that it assesses only the lower part of the large bowel (rectum and sigmoid colon) so polyps can be missed[30]. Additionally, it is only offered as a one-off screening assessment to adults aged 55 years[28, 29]. Furthermore, the participation rate is low at 43% in those offered the test[31].

Positive FIT or bowel scope results are followed by a colonoscopy, the gold standard for diagnosing bowel cancer (Figure 1.2)[32]. The participation rate is high at 89% in patients with a positive screening test result[33]. Colonoscopy carries a low risk of serious harm for the patient but is considered invasive[34].

Figure 1.2: Colonoscopy – a colonoscope inserted into the anus and directed through the bowel*



*This figure is taken from: NHS. What is a colonoscopy? 2019 [Accessed 1 August 2021]; Available from: https://www.nhs.uk/conditions/colonoscopy/[35]. Published under an Open Government Licence V3.0 – permissions not required.

Bowel cancer screening facilitates earlier detection and could improve mortality rates by 25%[36-38], but population incidence rates have only slightly improved over the last decade (73 in 2007 vs. 68 in 2017 per 100,000)[20]. The impact of screening is limited due to the reasons described above. Additionally, many barriers to participation have been identified, such as fear, perceived low-risk of bowel cancer, and concerns of bowel preparation and cleanliness[39-43]. A systematic review highlighted many ethical issues in the bowel cancer screening process, such as causing anxiety and poorer quality of life short-term[44]. Studies have shown that additional interventions, such as telephone advice and GP-endorsement, could improve uptake but would increase the burden on healthcare services[45-48].

1.2 How this research can help

1.2.1 The full blood count for bowel cancer

A full blood count (FBC) (also referred to as complete blood count) is a type of blood test commonly performed in primary care. A single FBC consists of 20 individual parameters measured from a patient's blood sample, such as haemoglobin, white blood cells, and platelets. A description of each FBC parameter is provided in Table 1.1[49, 50].

Table 1.1: Description of FBC parameters*

FBC parameter	Description/purpose
Red blood cell count (10^{12}/L)[1]	Number of red blood cells in the blood. Each red blood cell contains levels of haemoglobin.
Haemoglobin (g/dL)[1]	Carries oxygen around the body and is found in red blood cells.
Haematocrit/packed cell volume (%)[2]	Proportion of blood that is occupied by the red cells.
Mean corpuscular volume (fL)[2]	Average size of red blood cells. Calculated as haematocrit (l/l) divided by the number of red blood cells (10^{12}/l).
Mean corpuscular haemoglobin (pg)[2]	Average of amount of haemoglobin that is in each red blood cell. Calculated as haemoglobin (g/dl) divided by the red blood cell count (10^{12}/l).
Mean corpuscular haemoglobin concentration (g/dL)[2]	Average of amount of haemoglobin in individual cells based on the volume of red blood cells. Calculated as haemoglobin (g/dl) divided by the haematocrit (l/l).
Red cell distribution width (%)[2]	Amount of variation in the size of red cells.
Platelet count (10^{9}/L)[1]	Not a cell, but instead fragments of cytoplasm. They bind to sites of damaged blood vessels, e.g. cuts, and clump to form a blood clot to help prevent bleeding.
Mean platelet volume (fL)[2]	Average size of platelets.

White blood cell count (10^9/L)[1]	Number of white blood cells in the blood. They are of the immune system and involved in protecting the body against disease.
Basophil count (10^9/L)[1]	A type of white blood cell. Controls hypersensitivity reactions, allergic and inflammatory responses and fights parasitic infections.
Basophil (%)[2]	Basophil count divided by the number of white blood cells.
Eosinophil count (10^9/L)[1]	A type of white blood cell. Fights infection, including parasitic; has a role in allergic responses.
Eosinophil (%)[2]	Eosinophil count divided by the number of white blood cells.
Lymphocyte count (10^9/L)[1]	A type of white blood cell. Mediates immune responses.
Lymphocyte (%)[2]	Lymphocyte count divided by the number of white blood cells.
Monocyte count (10^9/L)[1]	A type of white blood cell. Kills microorganisms including mycobacteria and fungi; presents antigen to cells of the immune system.
Monocyte (%)[2]	Monocyte count divided by the number of white blood cells.
Neutrophil count (10^9/L)[1]	A type of white blood cell. Fights infection.
Neutrophil (%)[2]	Neutrophil count divided by the number of white blood cells.

[1]This parameter is measured directly from the blood sample.
[2]This parameter is derived using mathematical formulae programmed into the analyser (in haematology) and describes at least one measured parameter.
Abbreviations: FBC=full blood count.
* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

Anaemia (or low haemoglobin) determined from a FBC blood test, often in the presence of iron-deficiency, is a known risk factor for bowel cancer[52] and is recommended by the National Institute for Health and Care Excellence (NICE) and World Health Organisation (WHO) as a guideline for referral for further testing in general practice[53, 54]. Studies have shown that levels in some of the other FBC parameters are also associated with bowel cancer diagnosis, such as

high platelet count, low haematocrit, and high red blood cell distribution width[55-59]. Some studies assessed FBC parameters for long-term (all > 4 years) risk of diagnosis and reported a statistically significant association, indicating the potential for earlier detection[60-62].

1.2.2 A recent model for bowel cancer detection

Medial EarlySign, a private company in Israel, combine clinical expertise with machine-learning methods to improve patient health[63]. In 2016, they reported a machine-learning algorithm called the MeScore (subsequently called the ColonFlag) that assesses changes over time in FBC levels for bowel cancer diagnosis[64]. The commercially developed algorithm examines up to all 20 FBC parameters taken at multiple time-points for each patient in an Israeli population and derives a monotone score between 0 (lowest risk) and 100 (highest risk). This score was estimated by exploring patterns in FBC levels, with no corresponding functional equation relating to bowel cancer derived, hence the ColonFlag does not provide an absolute risk (i.e. probability) of diagnosis. Once derived, the authors assessed the association between the score and diagnosis and found a higher score was associated with an increased likelihood of diagnosis.

My supervisors externally validated the ColonFlag algorithm using Clinical Practice Research Datalink (CPRD) primary care data in the UK in 2017[65]. Discriminative performance, a measure of how well the model separates those who do and do not have the outcome, was measured using the c-statistic (also referred to as the area under the curve (AUC)), which ranges 0-1, where a 0.5 means the model is no better than chance and values higher than 0.7 (rule-of-thumb) represent good discriminative ability. The c-statistic was investigated at multiple time-points up to two years before diagnosis. The c-statistic at 3-6 months before diagnosis (primary outcome window) was similar between the UK (0.84) and Israeli (0.82) datasets. The UK data demonstrated it could discriminate higher-risk from lower-risk patients at 18-24 months prior to diagnosis, with a c-statistic of 0.78.

The ColonFlag is commercially developed so many details of the model remain unknown. This includes details of what changes over time are modelled and how, such as whether absolute difference or rate of change is modelled. It also provides a monotone score, which does not reflect a patient's risk/probability of being diagnosed. NICE have reported key uncertainties about the ColonFlag, which include the lack of published evidence on resource consequences and patient outcomes resulting from using the ColonFlag model[66]. NICE have estimated the ColonFlag would cost the NHS £50,000 per installation (could be at a primary care practice or a central hub) and £40,000 annually for technical support per installation, in addition to variable costs per person tested, but no cost-effective analysis has been performed. To my knowledge, the ColonFlag is currently not being used in any UK primary care practice. A recent study by my supervisors assessed whether the ColonFlag could identify cases before symptoms are first reported to a GP[13].

1.2.3 The evidence gap

To my knowledge, no alternative model to the ColonFlag has previously been built for the UK. I intended to take this research further by exploring how to optimise the use of the FBC test in a new prediction model that would be freely available to the NHS.

The particular FBC parameters that change due to bowel cancer and how they change was unknown before I began this research. For example, two studies have shown that mean platelet volume is associated with the presence of bowel cancer, with one reporting on average 0.29 fL lower and the other 0.59 fL higher levels in the run up to diagnosis compared to those without[67, 68]. The methodology used in these studies had not previously been critiqued. Additionally, the WHO guidelines define anaemia as haemoglobin <13 g/dL in men and <12 g/dL in females, which is a criterion for referral[54], but relevant levels in other FBC parameters remain unknown.

QCancer Colorectal is a prediction model developed using data from the QResearch primary care EHR database and is widely used across UK primary care practices[69]. The model relies on the onset of symptoms to identify a patient's risk of diagnosis but as previously described, symptoms often appear when the disease has developed to a late-stage. There is a need to identify cancer at earlier stages, before overt symptoms appear. The FBC test may play a role in early detection but currently there is limited evidence.

1.3 Aims & objectives

The aim of this research was to create and test a dynamic prediction model that utilises FBC test results in primary care electronic records to identify two-year risk of diagnosis of bowel cancer. The model is dynamic because it incorporates changes over time in repeated FBC tests for each patient, updating the risk when a new FBC becomes available.

The objectives were to:

1. Conduct a systematic review to identify which FBC parameters are prognostic factors and critically appraise existing FBC-based prediction models for bowel cancer detection.
2. Perform a data quality assessment of FBC data from CPRD to derive a reliable dataset for analysis and build a methodological approach for future researchers to perform such quality assessments.
3. Explore how the FBC measurements change over time in the run up to diagnosis to identify relevant trajectories that may be useful to identify cases.
4. Develop and test a dynamic prediction model for identifying two-year risk of bowel cancer diagnosis, incorporating findings from the systematic review and trajectories in the FBC and based on the derived FBC data.

This research was planned to tackle a clinical problem that, despite evolving efforts to improve, still largely exists. From a patient and public involvement (PPI) perspective, a blood test, commonly performed for a variety of reasons, provides a valuable opportunity to refine the group of patients referred for colonoscopy, reduce the clinical problem, and potentially help save many lives.

2

Systematic review of the FBC blood test for bowel cancer detection

This chapter has previously been published[51]:

Virdee P.S. *et al*. The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37* (2020). DOI: 10.3390/cancers12092348. Published under a CC-BY licence – permissions not required.

2.1 Introduction

Anaemia, or low haemoglobin level, often in the presence of other symptoms, is a criterion for referral for further investigation for bowel cancer[53, 54]. The FBC test includes up to 19 other parameters but it is unknown how best to utilise these results to identify bowel cancer cases. For example, can any of the 20 parameters be considered prognostic factors for diagnosis? If so, what blood levels and changes over time can indicate presence of bowel cancer? It is also unknown how many repeat FBC measurements would be needed to identify the presence of bowel cancer and whether the FBC test can identify early-stage tumours.

This review has two main aims:

1. Identify individual FBC parameters as prognostic factors for bowel cancer diagnosis (Aim 1).

2. Critically appraise existing FBC-based prediction models for bowel cancer diagnosis (Aim 2).

Studies of FBC data for bowel cancer detection are becoming popular, increasingly found in the literature over the last decade. To my knowledge, this is the first systematic review of the FBC blood test for bowel cancer detection. This review will inform the development and validation of future FBC-based prediction models for early detection of bowel cancer, namely the model developed in Chapter 5.

2.2 Methods

This review follows the Preferred Reporting Items for Systematic review and Meta-Analysis (PRISMA) reporting guidelines[70]. There was no direct patient investigation and only published articles were systematically reviewed, hence ethical approval was not required. The protocol for this review is registered with the International Prospective Register of Systematic Reviews database (PROSPERO: CRD42019134400)[71]. The protocol has previously been published[72]. Stata/SE V15.1 was used for all analyses.

2.2.1 Search criteria

Published journal articles are indexed in electronic databases. One study identified that the combination of MEDLINE, EMBASE, Google Scholar, and Web of Science databases performed best at retrieving articles, providing >98% recall of references from 58 published reviews included in their study[73]. Another database, CINAHL, specialises in nursing and allied healthcare publications and was considered relevant for this review because blood tests are commonly performed by nurses[74]. I therefore used CINAHL instead of the Google Scholar database, as the latter is considered a supplementary database, as identified in existing studies[75-77], and to make

management of the large number of references identified practical and achievable. The final databases used to identify articles were MEDLINE, EMBASE, CINAHL, and Web of Science.

Search terms included MesH, EMTREE, and Web of Science headings. Additionally, free-text terms included a variety of terminology relating to bowel cancer, the FBC test, and prediction or prognosis terms. There were no language or date restrictions applied to the search. The search strategy for each database is provided in Appendix A1, developed in collaboration with an experienced librarian, Shona Kirtley. Once the sample of eligible studies was determined, I searched through the reference list of each article to identify eligible studies that were not identified by the search strategy.

2.2.2 Study selection criteria

Articles that satisfied the following eligibility criteria were included in the study:

Inclusion:

- *Participants*: patients aged 18 years or above with at least one FBC parameter measured.
- *Outcome*: diagnosis of bowel cancer (yes/no).
- *Type of study (Aim 1 only)*: any in-human research article reporting on the association between any FBC parameter(s) and risk of diagnosis of bowel cancer.
- *Type of study (Aim 2 only)*: any in-human research article describing the development or validation of a FBC-based prediction model for bowel cancer detection, where the model was derived using at least two risk factors, with at least one being a parameter of the FBC test.

Exclusion:

- *Participants*: patients whose FBCs were collected and studied post-diagnosis of bowel cancer, such as for survival or treatment response outcomes, unless pre-treatment FBCs were compared with FBCs from patients who were not diagnosed with bowel cancer.
- *Type of study*: abstracts, conference proceedings, correspondence, and case studies, as they produced incomplete data for a thorough review. Existing systematic reviews were excluded.
- *Type of study (Aim 1 only)*: clinical trials were excluded as the interest here was in FBC data, not intervention.

2.2.3 Screening of articles

Initially, Endnote X9[78] was used to identify and remove duplicate articles in the entire reference set. Screening of articles was subsequently performed in two stages: 1) title and abstract screening and 2) full-text screening. Title and abstract screening was performed in Rayyan, a web and mobile tool designed for conducting systematic reviews[79], where the de-duplicated references were cross-checked with the selection criteria described in Section 2.2.2. Full-text screening was subsequently performed on the remaining references to determine the eligible study sample for data extraction and analysis. At both stages of screening, each article was assessed for eligibility independently by two reviewers: I screened 100% of articles and Ioana Marian, Leena Elhussein, and Anita Mansouri screened one-third each. Disagreements between the two reviewers were discussed until agreement was reached.

2.2.4 Data extraction

I developed two data extraction forms in Microsoft Excel: one for Aim 1 and one for Aim 2. These forms were piloted on a random sample of eight eligible articles. I extracted from all eight and

four independent extractors (Ioana Marian, Anita Mansouri, Leena Elhussein, and Jacqueline Birks) each extracted from two articles during data extraction pilot. The final data extraction forms are provided in Appendix A2. Two reviewers independently extracted data from each eligible full-text article: I extracted from 100% of studies and Ioana Marian, Anita Mansouri, and Leena Elhussein extracted one-third each. Disagreements between the two reviewers were resolved until agreement was reached.

2.2.5 Data analysis & synthesis

2.2.5.1 Analysis methods

Continuous data were summarised using means with standard deviations (SD) and categorical data using counts with proportions. I performed two sets of analyses: one for each aim.

For Aim 1, I stratified analyses by four time intervals between the FBC test and diagnosis of bowel cancer: 0<time≤6 months, 6<time≤12 months, 12<time≤36 months, and >36 months. The 0-6 month interval represented a time when patients may be in the process of clinical investigation for bowel cancer. The 6-12 month interval could be seen to bring diagnosis forward clinically significantly in symptomatic patients, where survival is more likely. The 12-36 month interval reflects pre-symptomatic patients, although it is still within the timescale of when we expect symptoms to become overt. The interval of >36 months was considered to reflect pre-symptomatic patients.

Analyses were performed for each FBC parameter separately. *A priori*, I planned to perform an (inverse-variance) random-effects meta-analysis of effect estimates, such as odds ratios (ORs), if at least three studies analysed the same FBC parameter and had available data per strata. The choice of three studies was advised by Marialena Trivella, an expert in meta-analysis based in

the Centre for Statistics in Medicine (CSM), Oxford. There were less than three individual studies, either with available data reported or for me to calculate effect estimates myself, so I did not analyse these effect estimates. Alternatively, a random-effects meta-analysis of the mean difference (with standard error (SE)) in individual FBC parameters between patients with and without a diagnosis was performed, as this analysis was performed by at least three individual studies and had available data. Heterogeneity between studies was measured using the I^2 statistic, which describes the proportion of variability in estimates that is due to heterogeneity (rather than sampling error/chance)[80]. Cohort and case-control studies were planned to be meta-analysed separately, as the two study designs can give different results, such as incidence rates and odds ratios, which have different interpretations between study designs[81]. However, data required for meta-analysis was not reported for cohort studies and could not be derived from other data reported in the study articles so only meta-analyses of case-control studies were performed. For FBC parameters that could not be meta-analysed, the mean (SD) of the mean difference in levels between groups is reported. A two-sided 5% significance level was used for all analyses.

For Aim 2, I critically appraised FBC-based clinical prediction models. Narrative discussions included model-building strategies, predictors in the final models, and model performance for identifying cases.

2.2.5.2 Missing data

If the data required to perform meta-analysis (mean difference and SE for patients with and without a diagnosis separately) was not directly reported in the study article, I followed a four-step approach to account for the missing data:

1. If relevant data (number of patients, mean FBC level and SD per outcome group) was available in the article, I estimated the mean difference and associated SE in FBC levels

between patients with and without a bowel cancer diagnosis using the formulae provided in the Cochrane Handbook for Systematic Reviews of Interventions[82]. Where the SD per outcome group was not reported but the p-value of the statistical t-test was, SEs were calculated using the p-value. Where non-exact p-values were reported, the nearest value was used in the estimation. For example, if an article reported $p<0.001$, then the nearest exact value of $p=0.0009$ was used.

2. If a study did not provide sufficient data for a mean difference and SE to be derived, I contacted the study authors and requested the additional data or asked for clarifications.
3. If no additional information was obtained from the authors, I approximated the data by measuring the means from graphs in articles, where reported.
4. If none of the above were possible, the study was not included in the meta-analysis (but any available data was still used in other analyses, such as for other FBC parameters).

2.2.6 Risk of bias

Risk of bias was assessed for Aim 1 and Aim 2 separately. I assessed risk of bias for all studies included. A second reviewer (Ioana Marian, Leena Elhussein, or Anita Mansouri) assessed risk of bias independently in their allocated studies. Disagreements were discussed until agreement was reached.

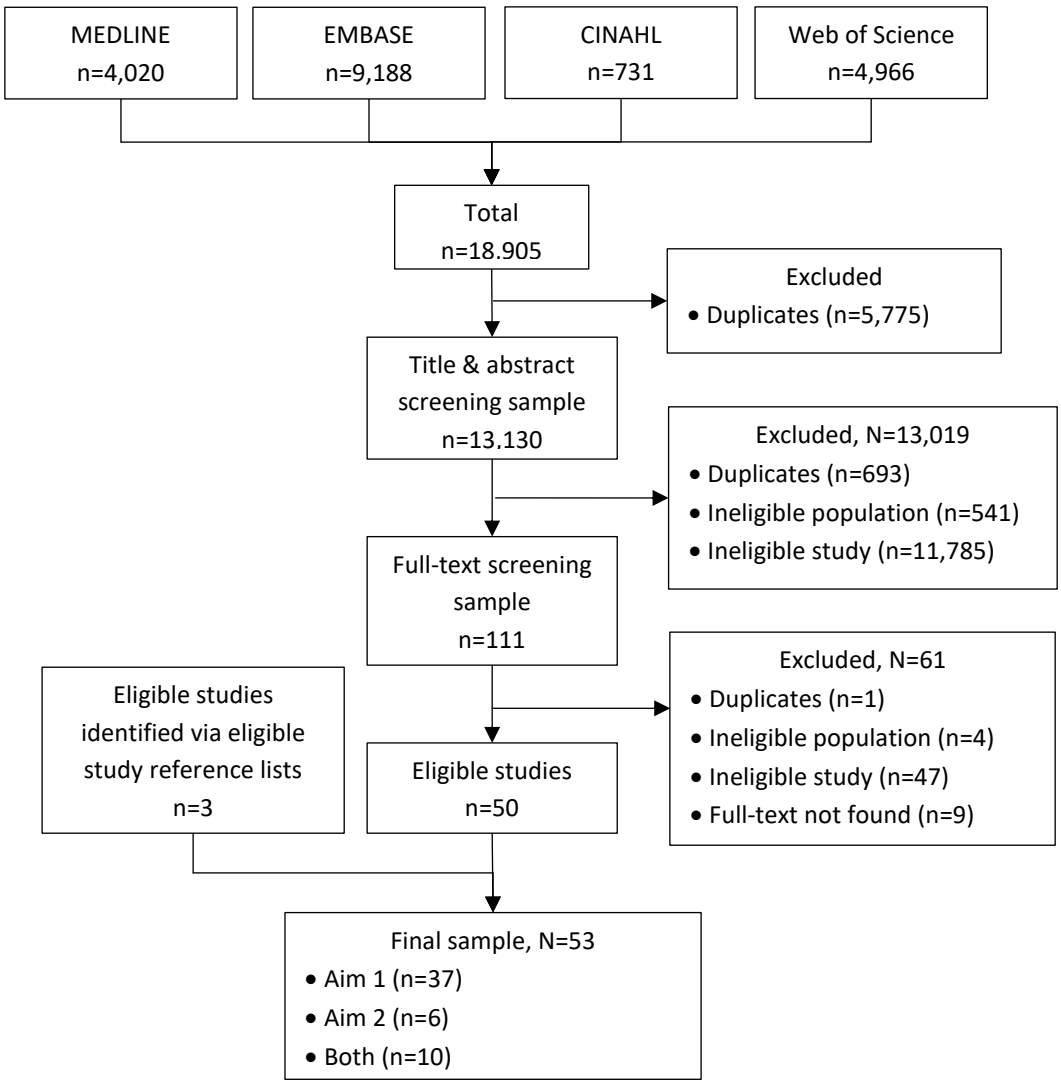
For studies of associations between FBC and bowel cancer diagnosis (Aim 1), risk of bias in each study was assessed using the Quality In Prognosis Studies (QUIPS) tool[83]. The QUIPS tool assesses potential biases in six domains: 1) participation, 2) attrition, 3) prognostic factor, 4) outcome, 5) confounders, and 6) analysis & reporting. Each domain is scored as either low, moderate, or high risk of bias.

For studies developing or validating clinical prediction models (Aim 2), the Cochrane Prediction model Risk Of Bias ASsessment Tool (PROBAST) was used[84]. The PROBAST tool assesses potential biases in four domains: 1) participants, 2) predictors, 3) outcome, and 4) analysis. Each domain is scored as either low, high, or unclear risk of bias.

2.3 Results

Studies published up until 3rd September 2019 (search date) were extracted from each database by an experienced librarian, Shona Kirtley. There were 18,905 references identified and 13,130 remained after I de-duplicated in EndNote X9[78]. Of these, 512 studied FBC data, of which 50 were for bowel cancer and met the eligibility criteria (described in Section 2.2.2) and were included in the review. An additional three studies were identified from reference lists in eligible articles. Figure 2.1 shows the full PRISMA diagram.

Figure 2.1: PRISMA diagram*



* This figure is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

2.3.1 Description of studies

The 53 eligible articles were published between 1995 and 2019, with 55% (n=29) published in the five years before the search date (2015-2019). A detailed description of each study is in Table 2.1.

Forty-seven studies assessed the relationship between at least one FBC parameter and diagnosis of bowel cancer and were included in the analysis of Aim 1. This includes 10 studies that

developed prediction models that were also included in Aim 2. There were 285 analyses in total among the 47 studies, resulting from some studies performing more than one analysis, such as using multiple outcomes, time windows, or FBC parameters.

Sixteen studies described the development or validation of a FBC-based prediction model and were included in the analysis of Aim 2. This includes the 10 studies included in Aim 1, with the remaining six externally validating existing FBC-based prediction models.

Table 2.1: Description of each study (n=53)*

Article	Included in Aim	Study design	Country	Patient setting	Patient type	Patient population	Average age	% male
Archer 2003[85]	Aim 1	Cohort	UK	Unclear	Anyone	**Inclusion**: men aged >50 years and women >55 years with histologically proven bowel cancer in 1996–1999. **Exclusion**: patients with recurrent bowel cancer.		
Ankus 2018[55]	Aim 1	Cohort	UK	Primary care	Anyone	**Inclusion**: random 10,000 from CPRD with first platelet count from 2000-2013 of 150-399 10^9/L, aged ≥40 years at the time of the platelet count with no prior cancer diagnosis. **Exclusion**: diagnosed with non-melanoma skin cancer after index date.		22.6
Ay 2015[86]	Aim 1	Case-control	Turkey	Unclear	Anyone	**Inclusion**: FBC within one week of diagnosis. **Exclusion**: patients with anaemia, haematological disorders, active infection, blood transfusion made <3 months, venous thrombosis <6 months, receiving iron deficiency treatment, hypertension, cardiac failure, inflammatory intestinal disease and rheumatoid arthritis.	60.3	

Ayling 2019[87]	Aim 2	Cohort	UK	Secondary care	Symptomatic	**Inclusion**: patients in the Gastroenterology Clinic in Derriford Hospital, Plymouth, between March 2014 and March 2017, referred with a low haemoglobin on a 2-week wait cancer pathway. Additionally, a cohort of consecutive patients who attended the Royal London Hospital for colonoscopy during 2017.		48.1
Bafandeh 2008[88]	Aim 1	Cohort	Iran	Unclear	Symptomatic	**Inclusion**: 480 consecutive patients with unexplained lower gastrointestinal tract symptoms for >3 months who underwent total colonoscopy between May 2005-April 2007. **Exclusion**: failure to reach the caecum or referred for polypectomy.	42.7	56
Bailey 2017[56]	Aim 1	Cohort	UK	Primary care	Anyone	**Inclusion**: patients who had had a primary care FBC taken.		30.7
Birks 2017[65]	Aim 2	Cohort	UK	Primary care	Anyone	**Inclusion**: all patients with ≥1 FBC present in their record. **Exclusion**: <12 months registered with the general practice, <2 years of follow-up data following the index	54.2	44.1

						date, history of bowel cancer before the index date, bowel cancer precursors, haemoglobin gene defects.		
Boursi 2016[57]	Both	Case-control	UK	Primary care	Anyone	**Inclusion**: all those in 1995-2013 from THIN. **Exclusion**: those with a diagnosis of bowel cancer syndromes, familial history of bowel cancer, IBD, or unacceptable medical records.	69.7	47.4
Cakmak 2017[89]	Aim 1	Case-control	Turkey	Unclear	Anyone	**Inclusion**: patients who underwent colonoscopy screening and diagnosed with colon adenocarcinoma from biopsy. **Exclusion**: patients with co-existing infections, hematologic diseases, renal diseases, vascular diseases, or other cancer types.	65.4	53.4
Collins 2012[90]	Aim 2	Cohort	UK	Primary care	Anyone	The same entry criteria as the original model development study were used (Hippisley-Cox 2012).	48	49.6
Cross 2019[91]	Aim 1	Cohort	UK	Secondary care	Symptomatic	**Inclusion**: patients from the SIGGAR trials, who were ≥55 years and judged to be in need of and fit enough for a whole colon investigation with full bowel preparation. **Exclusion**: if they were in follow-up for bowel cancer, had	69	41

						undergone whole colon investigation <6 months, familial adenomatous polyposis or Lynch syndrome, previously diagnosed with irritable bowel disease.		
Cubiella 2016[92]	Both	Cohort	Spain	Other	Symptomatic	**Inclusion**: the derivation cohort consisted of consecutive patients with gastrointestinal symptoms referred for colonoscopy from primary and secondary health care to Complexo Hospitalario Universitario de Ourense, Spain. The validation cohort included a prospective cohort of patients with gastrointestinal symptoms referred for colonoscopy in 11 hospitals in Spain. **Exclusion**: age <18 years, pregnant, asymptomatic individuals undergoing colonoscopy for bowel cancer screening, previous history of colonic disease who underwent a surveillance colonoscopy, requiring hospital admission, symptoms ceased <3 months before evaluation, and declined to participate after reading the informed consent form.	66	50.1

Fijten 1995[93]	Aim 1	Cohort	Netherlands	Primary care	Symptomatic	**Inclusion**: overt rectal bleeding was the reason for encounter or <3 months visible rectal blood loss. **Exclusion**: age <18 or >75 years, pregnant, urgent admission to a hospital, and no follow-up data available.	42	44
Firat 2016[94]	Both	Case-control	Turkey	Unclear	Anyone	**Inclusion**: bowel cancer cases and controls between 1 January 2010 and 1 March 2014 from Inonu University Turgut Ozal Center of Medicine, Department of Oncology.	58.6	56.3
Goldshtein 2010[95]	Aim 1	Case-control	Israel	Primary care	Anyone	**Inclusion**: MHS members aged 45–75 years diagnosed with bowel cancer between 1/1/2004 and 14/1/2009. **Exclusion**: haemoglobin values below 11.7 g/dl for women and 12.6 g/dl for men at any point during the first year of follow-up. Controls had no documented history of cancer.		
Goshen 2017[67]	Both	Case-control	Israel	Primary care	Anyone	**Inclusion**: MHS enrolees with and without a bowel cancer diagnosis between 40 and 75 years of age in 2002-2011, ≥1 blood test recorded before diagnosis. **Exclusion**: individuals with any form of cancer before 2002.		52.1

Hamilton 2005[96]	Aim 1	Case-control	UK	Primary care	Anyone	**Inclusion**: patients aged ≥40 years with a primary bowel cancer diagnosed in 1998-2002 at the Royal Devon and Exeter Hospital. Cases without positive histology were included if the records contained a specialist diagnosis of cancer based on strong clinical evidence. Controls were alive at the time of diagnosis of their case. **Exclusion**: unobtainable records, no consultations <2 years before diagnosis, previous bowel cancer, or residence outside Exeter at the time of diagnosis.		50.7
Hamilton 2008[52]	Aim 1	Case-control	UK	Primary care	Anyone	**Inclusion**: patients with bowel cancer aged ≥30 years and diagnosed between January 2000 and July 2006. Controls were free from bowel cancer. All participants had ≥2 years of electronic records prior to the date of diagnosis of the case.		53.1
Hamilton 2009[97]	Aim 1	Case-control	UK	Primary care	Anyone	**Inclusion**: patients aged ≥30 years between January 2001 and July 2006 and ≥2 years of full electronic records before		53.1

						diagnosis. Cases had bowel cancer diagnosis and controls did not.		
Hilsden 2018[98]	Aim 2	Cohort	Canada	Secondary care	Symptomatic	**Inclusion**: individuals aged 50-75 years who underwent a successful colonoscopy between January 2013 and June 2015 with bowel preparation rated by the endoscopist as adequate to detect polyps >5 mm, at average risk for bowel cancer, with a personal or family history of polyps or bowel cancer. **Exclusion**: positive guaiac or immunochemical fecal occult blood test, history of bowel cancer, known or suspected genetic predisposition to cancer or no FBC result <1 year prior to their colonoscopy.		45.3
Hippisley-Cox 2012[69]	Both	Cohort	UK	Primary care	Symptomatic	**Inclusion**: patients aged 30–84 years registered from practices between 1 January 2000 and 30 September 2010. **Exclusion**: no postcode-related Townsend score, history of bowel cancer at baseline, and recorded red flag symptom ≤12 months to the study entry date that might indicate bowel cancer.		50.1

Hippisley-Cox 2013[99]	Both	Cohort	UK	Primary care	Symptomatic	**Inclusion**: males aged 25–89 years from practices between 1 January 2000 and 1 April 2012. **Exclusion**: no postcode-related Townsend score or recorded red flag symptom ≤12 months before the study entry date were excluded.	48	100
Hippisley-Cox 2013[100]	Both	Cohort	UK	Primary care	Symptomatic	**Inclusion**: females aged 25–89 years from practices between 1 January 2000 and 1 April 2012. **Exclusion**: no postcode-related Townsend score or recorded red flag symptom ≤12 months before the study entry date were excluded.	50.2	0
Hornbrook 2017[101]	Aim 2	Case-control	USA	Unclear	Anyone	**Inclusion**: bowel cancer from the Kaiser Permanente Tumor Registry diagnosed with bowel cancer, had multiple FBCs ≤6 months of bowel cancer diagnosis, and ≥180 days of continuous enrolment prior to bowel cancer diagnosis. Controls received at least one outpatient FBC between 2000 and 2013, were aged 40-89 years at time of at least one FBC, had no history of cancer diagnoses in the database, were continuously enrolled from 180 days prior	58	44.2

						to FBC date through 24 months after the FBC date. **Exclusion**: bowel cancer patients with any cancer diagnosis prior to the bowel cancer diagnosis date.		
Huang 2019[102]	Aim 1	Case-control	China	Unclear	Anyone	**Inclusion**: patients newly diagnosed with bowel cancer at the first affiliated Hospital of Guangxi Medical University (Nanning, China) from June 2017 to October 2018. Controls had benign colorectal polyps or were healthy. **Exclusion**: haematological disorders, kidney disease, acute/chronic infections, coronary artery disease, hypertension, diabetes mellitus, medical treatment with anticoagulant, undergone transfusions ≤3 months, received neoadjuvant therapy, or had other cancers.	53.4	62
Hung 2015[103]	Aim 1	Cohort	Taiwan	Unclear	Symptomatic	**Inclusion**: patients newly diagnosed with iron deficiency anaemia between January 1, 2000 and December 31, 2010, aged ≥20 years at the time of IDA, and with no prior malignancies.		24

Joosten 2008[104]	Aim 1	Case-control	Belgium	Secondary care	Symptomatic	**Inclusion**: patients admitted to the acute geriatric ward and the geriatric day care centre of the University Hospitals Leuven, referred for colonoscopy during January 2002 to June 2007. **Exclusion**: patients with a history of bowel cancer, incomplete colonoscopy, polyp surveillance, previous colon surgery, red cell transfusion, or iron therapy ≤2 months.	82.3	61.6
Kilincalp 2015[68]	Aim 1	Case-control	Turkey	Unclear	Anyone	**Inclusion**: bowel cancer cases diagnosed by colonoscopy with colorectal resection thereafter and those with histological confirmation of adenocarcinomas. **Exclusion**: coexistent haematological disorders, renal disease, chronic infection, coronary artery or cerebrovascular disease, other types of cancers, received preoperative chemoradiotherapy and postoperative infections including wound infections.	60.7	67.9
Kinar 2016[64]	Both	Cohort	Israel	Primary care	Anyone	**Inclusion**: all insured individuals above age 40 years.	57.8	46.8

Kinar 2017[105]	Aim 2	Cohort	Israel	Primary care	Anyone	**Inclusion**: the development set was aged 50-75 on January 1, 2008 with ≥1 FBC recorded in the MHS during the six-month testing period. **Exclusion**: a diagnosis of any cancer recorded in the National Cancer Registry prior to January 1, 2008, or no blood test taken during the testing period.	60.9	44
Lawrenson 2006[106]	Aim 1	Cohort	UK	Primary care	Symptomatic	**Inclusion**: patients aged 40–89, registered in practices from England and Wales contributing to the GPRD at any time between 1 January 1992 to 31 December 1999, and with at least 1 year of data.		
Lee 2006[60]	Aim 1	Cohort	Korea	Unclear	Anyone	**Inclusion**: government employees, teachers, and their dependents insured by the Korean Medical Insurance Corporation in 1993 and 1995. **Exclusion**: no white blood cell record in their examinations, history of any cancer at enrolment, cancer-related death before the start of follow-up and missing data on any covariate.	56.7	25.7
Margolis 2007[61]	Aim 1	Cohort	USA	Unclear	Anyone	**Inclusion**: postmenopausal women aged 50-79 years recruited at 40 clinical centres throughout the United	63	0

						States between September 1, 1993 and December 31, 1998 from a hormone trial, dietary modification trial, and calcium/vitamin D supplementation trial. The observational study included women screened but ineligible for the trials or recruited through a direct invitation for screening into the observational study. **Exclusion**: history of cancer except non-melanoma skin cancer at baseline, missing baseline white blood cell count, missing data regarding cancer history at baseline, white blood cell count >15.0 10^9/L or <2.5 10^9/L.		
Marshall 2011[107]	Both	Case-control	UK	Primary care	Anyone	**Inclusion**: The development set had patients aged ≥30 years with or without a diagnosis of bowel cancer between January 2001 and July 2006 and ≥2 years of records before diagnosis. The validation was a case-control study in a single primary care trust in Exeter UK, aged >40 years between 1998 and 2002.		53

Mashlab 2018[108]	Aim 1	Cohort	UK	Secondary care	Symptomatic	**Inclusion**: patients referred under the 2-week wait pathway for suspected bowel cancer from the referral database created by specialist nurses at the colorectal service. **Exclusion**: duplicate and rejected referrals, cases with no FBC on referral, no investigations or an unknown outcome.		45.4
Naef 1999[109]	Aim 1	Cohort	Switzerland	Unclear	Anyone	**Inclusion**: primary and secondary small-bowel tumours treated in the department between January 1984 and December 1993, as well as associated syndromes. **Exclusion**: ileocecal valve and peri-ampullary duodenal tumours.	61.4	55.6
Nakama 2000[110]	Aim 1	Cohort	Japan	Unclear	Asymptomatic	**Inclusion**: asymptomatic persons aged 40–60 years who participated in a medical check-up for bowel cancer as recommended by the companies with which they were employed.		
Panagiotopoulou 2014[111]	Aim 1	Unclear	UK	Unclear	Symptomatic	**Inclusion**: consecutive referrals for suspected bowel cancer received at Centre A between November 2008 and		46.2

						June 2009 and Centre B between April 2010 and March 2011 using the cancer services prospectively maintained database. **Exclusion**: no blood tests available, previous history of bowel cancer/panproctocolectomy and diagnosis of bowel cancer in another hospital.		
Panzuto 2003[112]	Aim 1	Cohort	Italy	Primary care	Symptomatic	**Inclusion**: outpatients with symptoms considered suspicious for the presence of a colon disease to rule out bowel cancer. **Exclusion**: previous diagnoses of colorectal disorders or a recent large bowel examination.	61	42.9
Pilling 2018[62]	Aim 1	Cohort	UK	Other	Anyone	**Inclusion**: volunteers aged 40-70 years recruited by postal invitation from the UK Biobank study, living ≤25 miles of assessment centres in Great Britain, seen between 2006 and 2010. **Exclusion**: those with anaemia, coronary artery disease, cancer, type-2 diabetes, stroke, chronic obstructive pulmonary disease, or hypertension.	55	51.8
Prizment 2011[113]	Aim 1	Cohort	USA	Other	Anyone	**Inclusion**: patients aged 45-69 years from the ARIC study of atherosclerosis in 1987–1989, from suburban	53.9	46

						Minneapolis, Forsyth County, Jackson, and Washington County. **Exclusion**: prevalent cancer at the start of follow-up, did not consent to participate, or had missing biomarker information.		
Raje 2007[114]	Aim 1	Cohort	UK	Secondary care	Asymptomatic	**Inclusion**: females >50 and males >40 years with iron deficiency anaemia referred to one district general hospital during 2003. **Exclusion**: patients with haemoglobinopathy.		40.1
Schneider 2018[115]	Aim 1	Case-control	UK	Primary care	Anyone	**Inclusion**: patients in CPRD aged 18-89 years with a read-coded bowel cancer diagnosis and matched control. **Exclusion**: history of any cancer before the index date except non-melanoma skin cancer.		55.5
Shi 2019[58]	Aim 1	Case-control	China	Unclear	Anyone	**Inclusion**: patients with bowel cancer from historical biopsy undergoing radical surgery at the People's Hospital of Liuzhou or those with colon polyps, with blood test data from 2 weeks before surgery. **Exclusion**: previous	61.7	54.5

						neoadjuvant therapy, presence of infection, and age of >85 years.		
Song 2018[116]	Aim 1	Case-control	China	Unclear	Anyone	**Inclusion**: patients with bowel cancer diagnosed at Fujian Medical University Union Hospital (China) from June 2015 to November 2017, or controls with colorectal adenomas patients or healthy participants. **Exclusion**: patients with anaemia, hematologic disorders, blood transfusion made ≤3 months, receiving iron deficiency treatment and with active inflammation.		59.7
Spell 2004[59]	Aim 1	Case-control	USA	Other	Anyone	**Inclusion**: bowel cancer cases aged ≥18 years with ≥1 FBC recorded prior to surgery, performed at University of Texas Medical Branch. Controls were without bowel cancer during the same time with a routine flexible sigmoidoscopy and ≥1 FBC <6 months of the procedure. **Exclusion**: other colon malignancies (cases only), no FBC data prior iron therapy or red blood cell transfusion, chemotherapy or radiation therapy <1 year of diagnosis,		47.5

						documented vitamin B12/folate deficiencies, or rectal cancer.		
Stapley 2006[117]	Aim 1	Cohort	UK	Primary care	Symptomatic	**Inclusion**: primary bowel cancer in patients aged ≥40 years from Exeter Primary Care Trust, diagnosed between 1998 and 2002.	73	51
Thompson 2017[118]	Both	Cohort	UK	Secondary care	Symptomatic	**Inclusion**: newly referred to a colorectal surgical clinic undergoing sigmoidoscopy and/or whole colon investigation. **Exclusion**: previous bowel cancer diagnosis and subsequent referral to the colorectal clinic, or no sigmoidoscopy/WCI performed.	60.1	43.8
van Boxtel-Wilms 2016[119]	Aim 1	Case-control	Netherlands	Primary care	Symptomatic	**Inclusion**: patients with or without bowel cancer between 1 January 1992 and 31 December 2011 with ≥2 years of record before index date, and controls with a GP encounter <1 month of the index date.		56.5
Wu 2019[120]	Aim 1	Case-control	China	Unclear	Anyone	**Inclusion**: patients who underwent surgical resection after bowel cancer diagnosis but did not receive pharmacological treatment. **Exclusion**: pregnancy or	53.6	59.9

						lactation, other malignancies, thyroid disease, diabetes, cardiovascular disease, autoimmune diseases, kidney disease, haematological disease, or blood transfusion <3 months before admission.		
Yang 2018[121]	Aim 1	Case-control	China	Unclear	Anyone	**Inclusion**: newly diagnosed and pathologically proven patients with bowel cancer or benign colon polyps admitted to Shanghai Tongji Hospital between July 2014 and June 2017. **Exclusion**: patients with cardiovascular, kidney, blood, or other malignant diseases, or blood transfusion <3 months of admission.		
Zhou 2017[122]	Aim 1	Case-control	China	Unclear	Anyone	**Inclusion**: patients with bowel cancer or adenomatous polyp histologically confirmed whose families provided written informed consent. Healthy people had no symptoms and cancer history. **Exclusion**: acute infective disease and haematological disorders.		
Zhu 2018[123]	Aim 1	Case-control	China	Unclear	Anyone	**Inclusion**: bowel cancer at Fujian Medical University Union Hospital (China) from June 2015 to October 2017 with no	60	60.1

						prior treatment, or controls with colorectal adenomas or healthy volunteers. **Exclusion**: haematological disorders, coronary artery disease, hypertension, diabetes mellitus, medical treatment with anticoagulant, and acetylic salicylic acid.		

Abbreviations: CPRD=Clinical Practice Research Datalink, FBC=full blood count, THIN=The Health Improvement Network, MHS=Maccabi Health Services, IDA=iron deficiency anaemia, GPRD=General Practice Research Database.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

2.3.1.1 Study design & participants

A case-control design was used by 43% (n=23) and cohort design by 55% (n=29) of the 53 studies. The study design was unclear in one study, Panagiotopoulou 2014[111]. The median (min, max) number of patients analysed was 2,398.5 (54, 2,793,468) over all the studies. The period of recruitment ranged from 1984 to 2018. Most studies were conducted in UK (43%, n=23) and Chinese participants (13%, n=7) and spanned a region of 15 countries. There were 36% (n=19) of studies in symptomatic patients, 2% (n=1) in asymptomatic patients, and 62% (n=33) in either. The mean (SD) age at study entry among 29 studies that reported age was 58.8 years (SD=8.6). On average, 47.7% (SD=15.8) of participants were male among 46 studies that described sex.

2.3.1.2 Data considerations

Of the 53 studies, 6% (n=3) performed some data validation or quality assessment to ensure reliability and accuracy of the data: Boursi 2016[57], Firat 2016[94], and Prizment 2011[113]. Seventeen percent (n=9) performed a complete-case analysis, two used imputation methods, and one kept the data as-is. Some parameters are mathematically related to others and can be derived using existing formulae (described in detail in Chapter 3), performed by Birks 2017 alone[65]. In 9% (n=5) of studies, missing data was analysed as a separate category or combined with other categories. Methods for accounting for missing data were not discussed in 66% (n=35) of studies.

2.3.1.3 Outcome details (Aim 1)

In Aim 1, studies reported multiple analyses and I therefore describe the different bowel cancer outcomes at an analysis-level. All outcomes were categorical among the 285 analyses, with 83.2% (n=160) using ‘yes/no’, 12.3% (n=35) using ‘yes/benign polyp’, and remaining using other outcomes.

The outcome time window, such as one-year or two-year risk of diagnosis following the time of the FBC measurement, was provided for 50.5% (n=146) of the 285 analyses. Among the 47 studies included in Aim 1, 87% (n=41) used a single outcome time window, one study used two outcome windows, one study used three outcome windows, and four did not clearly report the outcome window used. The outcome time window ranged from the time of diagnosis to 19 years earlier. The most common outcome window was two years, used in seven studies. Four analyses from three studies justified their choice of outcome time window, which was two-year risk of diagnosis, stating that this represents the time in which existing cancers are likely to manifest clinically.

2.3.1.4 Outcome details (Aim 2)

For Aim 2, the 13 prediction models from ten studies all used 'yes/no' as the outcome. In two of these studies, patients without a diagnosis were considered not to have any other type of cancer (Hippisley-Cox 2013[99] and Hippisley-Cox 2013[100]). The outcome time window ranged from the time of diagnosis to three years earlier. The most common outcome window was two-year risk of diagnosis, used for five models built among four studies. Four analyses from three studies justified their choice of outcome time window, which was 2-year risk of diagnosis, stating that this represents the time in which existing cancers are likely to manifest clinically.

Among all internal validation studies (n=7), one study assessed the performance of a prediction model using four outcome windows (Kinar 2016[64]) and the remaining used one outcome window. Among all external validation studies (n=9), one study assessed the performance of a prediction model using 15 outcome windows (Kinar 2016[64]), one study used five (Birks 2017[65]), and the remaining used one window. The most commonly used outcome window among validation studies was 2 years, used in the internal validation of five models and external

validation of five models. In one external validation study, Hilsden 2018[98], the outcome window was unclear.

All external validation studies used the same outcome and time window as the development study corresponding to the model undergoing validation (with or without additional outcome windows). Internal validation studies that used more than one outcome time window did not justify their choice of any of these additional outcome time windows. One external validation study, Birks 2017[65], that used more than one outcome time window provided justification for their choice of windows, stating that the opportunity to modify bowel cancer prognosis requires an adequate time window for diagnosis and intervention.

2.3.2 FBC for bowel cancer detection (Aim 1)

Table 2.2 shows the number and proportion of individual FBC parameters analysed in the 47 studies. Haemoglobin was most commonly analysed (81%, n=38), followed by mean corpuscular volume (34%, n=16), platelet count (26%, n=12), and white blood cell count (23%, n=11). Kinar 2016[64] was the only study that analysed all 20 FBC parameters and developed a machine-learning algorithm called the ColonFlag (previously the MeScore). Due to the nature of machine-learning, no functional equation was derived so no individual effect estimates were reported in their article. No other study analysed basophil percentage, eosinophil percentage, or monocyte percentage.

Table 2.2: Full blood count parameters analysed in each study included in Aim 1 (n=47)*

Article	RBC	Hb	Hc	MCV	MCH	MCHC	RDW	Plat	MPV	WBC	BasC	BasP	EosC	EosP	LymC	LymP	MonC	MonP	NeuC	NeuP
Acher 2003		✓		✓		✓														
Ankus 2018								✓												
Ay 2015		✓		✓			✓	✓												
Bafandeh 2008		✓																		
Bailey 2017								✓												
Boursi 2016	✓	✓	✓	✓	✓			✓		✓	✓		✓		✓		✓		✓	
Cakmak 2017		✓					✓	✓												
Cross 2019		✓																		
Cubiella 2016		✓		✓																
Fijten 1995		✓								✓										
Firat 2016		✓						✓		✓										
Goldshtein 2010		✓																		
Goshen 2017	✓	✓		✓			✓	✓	✓	✓	✓		✓		✓		✓		✓	
Hamilton 2005		✓																		
Hamilton 2008		✓		✓																
Hamilton 2009		✓		✓																

Hippisley-Cox 2012		✓																		
Hippisley-Cox 2013		✓																		
Hippisley-Cox 2013		✓																		
Huang 2019		✓								✓					✓					
Hung 2015		✓																		
Joosten 2008		✓																		
Kilincalp 2015		✓						✓	✓											
Kinar 2016	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Lawrenson 2006		✓																		
Lee 2006										✓										
Margolis 2007										✓										
Marshall 2011		✓		✓																
Mashlab 2018		✓		✓																
Naef 1999		✓																		
Nakama 2000		✓																		
Panagiotopoulou 2014		✓		✓																
Panzuto 2003		✓		✓																
Pilling 2018		✓		✓			✓													

Prizment 2011										✓										
Raje 2007		✓		✓																
Schneider 2018	✓	✓		✓																
Shi 2019							✓													
Song 2018							✓													
Spell 2004		✓		✓			✓													
Stapley 2006		✓																		
Thompson 2017		✓																		
Wu 2019	✓	✓						✓	✓	✓					✓		✓		✓	
Yang 2018		✓					✓	✓							✓				✓	
Zhou 2017										✓						✓				✓
Zhu 2018								✓	✓											
van Boxtel-Wilms 2016		✓																		
Total	**5**	**38**	**2**	**16**	**2**	**2**	**9**	**12**	**5**	**11**	**3**	**1**	**3**	**1**	**6**	**2**	**4**	**1**	**5**	**2**
% of 47 studies	**11%**	**81%**	**4%**	**34%**	**4%**	**4%**	**19%**	**26%**	**11%**	**23%**	**6%**	**2%**	**6%**	**2%**	**13%**	**4%**	**9%**	**2%**	**11%**	**4%**

Abbreviations: RBC=red blood cells, Hb=haemoglobin, Hc=haematocrit, MCV=mean corpuscular volume, MCH=mean corpuscular haemoglobin, MCHC=mean corpuscular haemoglobin concentration, RDW=red blood cell distribution width, Plat=platelet count, MPV=mean platelet volume, WBC=white blood cells, BasC=basophil count, BasP=basophil %, EosC=eosinophil count, EosP=eosinophil %, LymC=lymphocyte count, LymP=lymphocyte %, MonC=monocyte count, MonP=monocyte %, NeuC=neutrophil count, NeuP=neutrophil %

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

2.3.2.1 Risk of bias

Risk of bias in each study, assessed using the QUIPS tool, is reported in Table 2.3. Risk of bias was considered highest in the confounders domain (41%, n=19), followed by outcome (23%, n=11) and analysis and reporting (19%, n=9).

Table 2.3: Risk of bias (assessed using the QUIPS tool) in each study included in Aim 1 (n=47)*

Article	Participation	Attrition	Prognostic factor	Outcome	Confounders	Analysis & reporting
Acher 2003	High	Low	Moderate	Low	High	High
Ankus 2018	Low	Low	Low	Moderate	High	Low
Ay 2015	Moderate	Low	Low	Moderate	High	High
Bafandeh 2008	Low	High	High	Moderate	High	High
Bailey 2017	Low	Low	Low	Low	High	Low
Boursi 2016	Low	Low	Moderate	Low	Moderate	Moderate
Cakmak 2017	Low	Low	Moderate	Low	High	Moderate
Cross 2019	Moderate	Low	Low	High	High	Low
Cubiella 2016	Low	Moderate	Low	Moderate	Moderate	Low
Fijten 1995	Low	Moderate	Moderate	High	High	Moderate
Firat 2016	Moderate	Low	High	Moderate	Moderate	High
Goldshtein 2010	Moderate	Low	Moderate	High	High	High
Goshen 2017	Moderate	Low	High	Moderate	Moderate	High
Hamilton 2005	Moderate	Low	Low	Low	Low	Moderate
Hamilton 2008	Low	Low	Moderate	Low	Low	Moderate
Hamilton 2009	Low	Low	Low	Low	Low	High
Hippisley-Cox 2012	Low	Low	Moderate	Low	Low	Low
Hippisley-Cox 2013	Low	Low	Moderate	Low	Low	Low
Hippisley-Cox 2013	Low	Low	Moderate	Low	Low	Low
Huang 2019	Low	Low	Low	Low	Moderate	Moderate

Hung 2015	Low	Low	High	Moderate	Moderate	Low
Joosten 2008	Low	Low	Low	Low	Moderate	Moderate
Kilincalp 2015	Low	Low	Low	Low	Moderate	Low
Kinar 2016	Low	Low	Moderate	Low	Moderate	High
Lawrenson 2006	Moderate	Low	High	Moderate	Moderate	Moderate
Lee 2006	Low	Low	Moderate	Moderate	Moderate	Low
Margolis 2007	Low	Low	Moderate	Low	Moderate	Moderate
Marshall 2011	Low	Low	Low	Low	Low	Low
Mashlab 2018	Low	Low	Low	Moderate	High	Low
Naef 1999	High	Low	Low	High	High	Moderate
Nakama 2000	High	Low	Low	Moderate	High	Low
Panagiotopoulou 2014	Moderate	Low	Low	High	Moderate	Moderate
Panzuto 2003	Moderate	Low	Low	Low	Moderate	Moderate
Pilling 2018	Low	Low	Moderate	High	Low	Moderate
Prizment 2011	Low	Low	Low	Moderate	Low	Moderate
Raje 2007	Moderate	Low	Low	Low	High	Moderate
Schneider 2018	Low	Low	Low	Low	Low	Moderate
Shi 2019	Low	Low	Low	Moderate	High	Low
Song 2018	Low	Low	Low	High	High	Moderate
Spell 2004	Moderate	Low	Low	Low	Moderate	Moderate
Stapley 2006	Low	Low	Low	Low	Moderate	High
Thompson 2017	Moderate	Low	Low	Low	Moderate	Low
Wu 2019	Low	Low	Low	High	High	Low
Yang 2018	Moderate	Low	Low	High	High	Low
Zhou 2017	Low	Low	Low	High	High	Low
Zhu 2018	Low	Low	Low	High	High	Low
van Boxtel-Wilms 2016	Low	Low	Moderate	Low	Moderate	Low
Total low (%)	**31 (66%)**	**44 (94%)**	**31 (66%)**	**23 (49%)**	**10 (21%)**	**20 (43%)**

Total moderate (%)	**13 (28%)**	**2 (4%)**	**13 (28%)**	**13 (28%)**	**18 (38%)**	**18 (38%)**
Total high (%)	**3 (6%)**	**1 (2%)**	**5 (11%)**	**11 (23%)**	**19 (41%)**	**9 (19%)**

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

Risk of bias was highest in Bafandeh 2008[88], having scored high in four of the six domains. The authors reported a prospective cohort study of bowel cancer in Iranian patients with gastrointestinal symptoms. Patients were excluded if the outcome was not measured, with the number of patients excluded unspecified. The response rate is therefore unknown. Key characteristics were not described for patients who were excluded or compared to those from patients who did have the outcome measured. The authors used multivariable logistic regression for bowel cancer and included symptoms as prognostic factors in the model. The prognostic factors were not clearly defined. For example, it is unclear what haemoglobin threshold was used to define anaemia in the study and how the presence of each symptom was defined. Important confounders were not recorded and therefore were not adjusted for in the study design or analysis. Effect estimates from the multivariable logistic regression model were not reported – only p-values for prognostic factors with a statistically significant association with bowel cancer were reported.

2.3.2.2 Red blood cell count

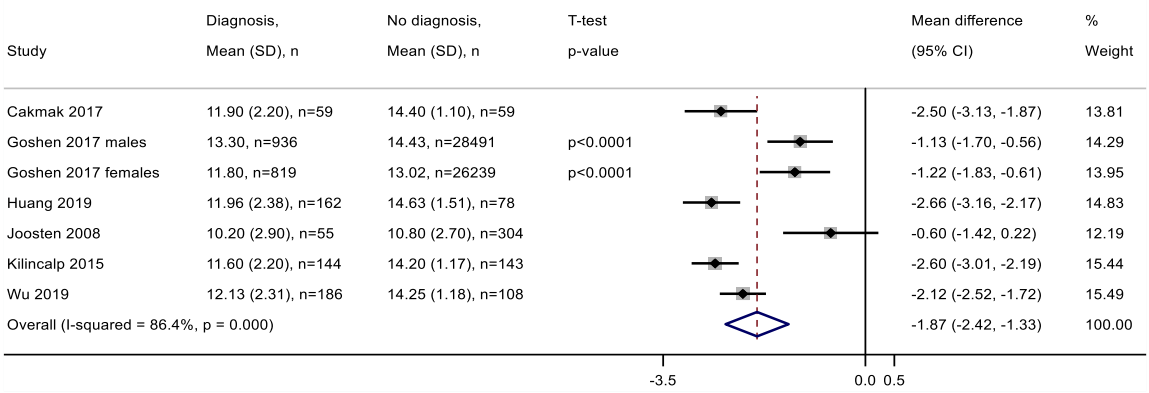
Findings from each study that analysed red blood cell count are in Table A1 (Appendix A3). All analyses indicated that patients with bowel cancer have lower red blood cell count within one year of diagnosis compared to patients without bowel cancer. No study analysed red blood cell counts measured in other time strata. There were only two studies with available data for meta-analysis, which was therefore not performed. Among the two studies, patients with bowel cancer had on average 0.17 10^9/L (SD=0.12) lower red blood cell count than patients without a diagnosis, where red blood cell count was measured up to 6 months before bowel assessment.

Wu 2019[120] also assessed benign polyps and reported 0.36 10^9/L lower red blood cell count in patients with a diagnosis compared to patients with benign polyps.

2.3.2.3 Haemoglobin

Findings from each study that analysed haemoglobin are in Table A2 (Appendix A4). In each strata, all studies reported lower haemoglobin levels in patients who were diagnosed compared to patients who were not. On meta-analysis using six studies with available data, the pooled mean difference indicated that patients with a diagnosis up to six months from the haemoglobin measurement have 1.87 (95% CI=1.33-2.42) g/dL lower haemoglobin (g/dL) than patients without a diagnosis (Figure 2.2). The I^2 statistic was 86.4%, indicating high heterogeneity between the six studies. Data were not available for meta-analysis in other time strata.

Figure 2.2: Forest plot of mean difference in haemoglobin (g/dL) between patients with and without a diagnosis of bowel cancer up to six months later*



Abbreviations: SD=standard deviation, CI=confidence interval.
* This figure is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

Joosten 2008[104] reported no statistically significant difference between groups and scored low-moderate risk of bias. The lack of statistical significance may be due to the patient sample being hospitalised and older on average (82.3 years) compared to the other studies included in the

meta-analysis (53.4 – 65.4 years). Each study was iteratively removed and the meta-analysis repeated with the remaining studies. The I^2 statistic remained high (minimum 81.5%), including when Joosten 2008 was removed (84.1%), with a corresponding pooled mean difference of -2.06 (95% CI=-2.57–-1.54).

Four studies compared haemoglobin levels between patients diagnosed and patients with benign polyps and reported statistically significant differences. From three studies with available data, haemoglobin was on average 0.98 (SD=0.53) g/dL lower in patients diagnosed than patients with benign polyps. Haemoglobin levels for both groups were not assessed for any other time strata.

2.3.2.4 Haematocrit

Findings from each study that analysed haematocrit are in Table A3 (Appendix A5). No study reported the mean difference between patients with and without a bowel cancer diagnosis so no meta-analysis was performed. One study, Boursi 2016[57], reported an odds ratio that indicated higher levels of haematocrit (analysed as a continuous variable) are associated with decreased odds of diagnosis (OR=0.97, 95% CI=0.95-0.98).

2.3.2.5 Mean corpuscular volume

Findings from each study that analysed mean corpuscular volume are in Table A4 (Appendix A6). All studies reported a statistically significant relationship between mean corpuscular volume and bowel cancer diagnosis, where the blood test was measured up to 36 months before diagnosis. In all studies, patients with a diagnosis had lower mean corpuscular volume than patients without. Panagiotopoulou 2014[111] reported no statistically significant difference in odds of 3-month diagnosis between patients with mean corpuscular volume <80 fL and ≥80 fL when unadjusted (OR=1.73, 95% CI=0.96-3.1), but this achieved significance when adjusted for other

factors (OR=2.2, 95% CI=1.2-4.1). Goshen 2017 was the only study that reported the mean difference between groups, with patients diagnosed having 3.67 fL lower mean corpuscular volume on average up to six months before bowel assessment compared to patients without a diagnosis. No meta-analysis was therefore performed. In the ≥36 month strata, one study, Pilling 2018[62], suggested no association between mean corpuscular volume and bowel cancer.

Ay 2015[86] was the only study that compared mean corpuscular volume between patients diagnosed and patients with benign polyps. Patients diagnosed had on average 1.5 fL lower mean corpuscular volume compared to patients with benign polyps, but the statistical test failed to achieve significance (t-test p≥0.05).

2.3.2.6 Mean corpuscular haemoglobin

This parameter was analysed by two studies: Kinar 2016[64] and Boursi 2016[57] (see Table A5, Appendix A7). Only the latter reported on the relationship with diagnosis, which was an odds ratio indicating that higher levels of mean corpuscular haemoglobin (analysed as a continuous variable) are associated with decreased odds of diagnosis (OR=0.75, 95% CI=0.74-0.76).

2.3.2.7 Mean corpuscular haemoglobin concentration

This parameter was analysed by two studies: Kinar 2016[64] and Acher 2003[85] (see Table A6, Appendix A8). Neither study reported sufficient results for me to assess the diagnostic ability of this FBC parameter.

2.3.2.8 Red blood cell distribution width

Findings from each study that analysed red blood cell distribution width are in Table A7 (Appendix A9). No study was in the 6-12 month or 12-36 month strata. In the 0-6 month strata, all analyses indicated that patients with a diagnosis have higher red blood cell distribution width

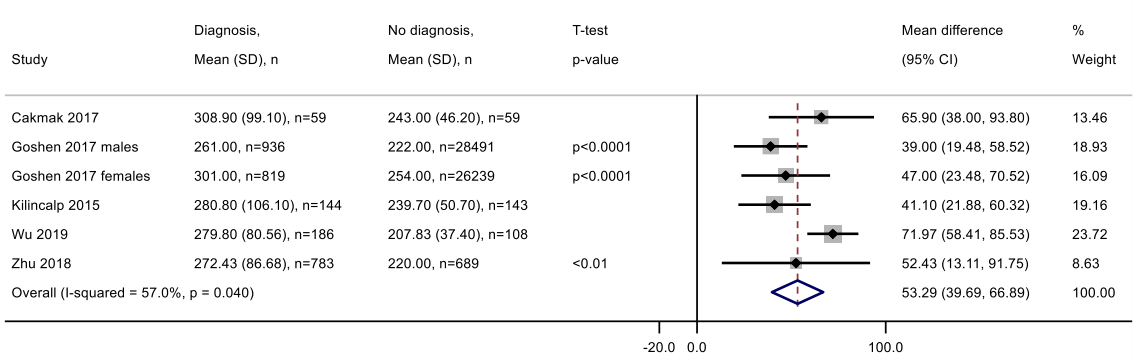
than patients without a diagnosis. Song 2018[116] performed a ROC analysis and derived a cut-off of 13.95%, with higher levels indicating diagnosis. The cut-off had 41% sensitivity, 94% specificity, and the c-statistic was 0.64 (95% CI=0.61-0.67). From two studies with available data, patients with bowel cancer had on average 1.4% (SD=1.0) higher red blood cell distribution width up to six months earlier than patients without. One study, Pilling 2018[62], analysed red blood cell distribution width earlier than 36 months prior to bowel assessment. The authors reported that higher levels were associated with increased rate of diagnosis up to 4.5 years after the blood test but not at later time points.

Four studies compared red blood cell distribution width between patients diagnosed and patients with benign polyps. All four studies performed ROC analyses and reported cut-offs for diagnosis ranging from 13.2% to 17.5%, where the blood test was performed up to six months prior to bowel assessment. From two studies with available data, patients diagnosed with bowel cancer had 1.9% (SD=0.4) higher red blood cell distribution width than patients with benign polyps.

2.3.2.9 Platelets

Findings from each study that analysed platelets are in Table A8 (Appendix A10). Within one year prior to diagnosis, platelet count was reported to be higher in patients with a diagnosis compared to patients without, which reached statistical significance in all studies. Platelet levels were not analysed by studies in other strata. I meta-analysed the mean difference between patients with and without a diagnosis of bowel cancer using five studies with available data. The pooled estimate indicated a statistically significant difference in platelet count, with patients diagnosed having on average 53.29 (95% CI=39.69-66.89) 10^9/L higher platelet count than patients without a diagnosis up to six months earlier (Figure 2.3). Heterogeneity among the studies was moderate (I^2=57.0%). Data was not available for meta-analysis in other time strata.

Figure 2.3: Forest plot of mean difference in platelet count (10^9/L) between patients with and without a diagnosis of bowel cancer up to six months later*



Abbreviations: SD=standard deviation, CI=confidence interval.
* This figure is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

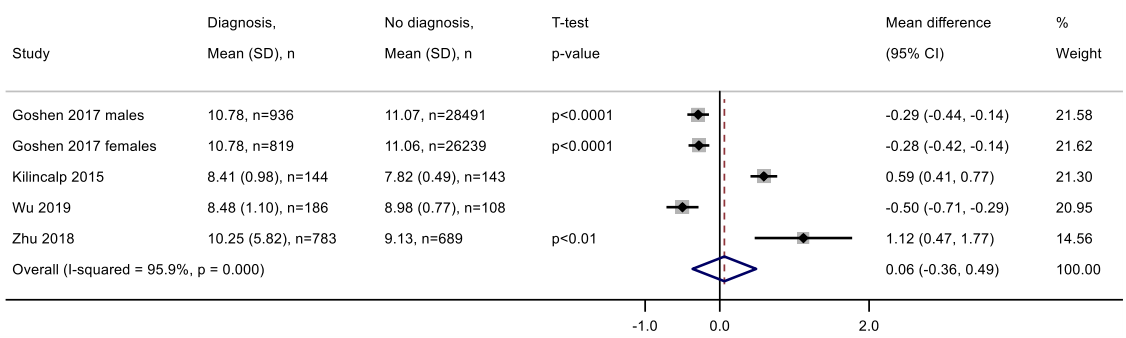
All studies that compared platelet levels between patients with benign polyps and patients diagnosed were in the 0-6 month strata. Ay 2015[86] reported no statistically significant difference in mean platelet levels between patients with benign polyps and patients diagnosed (t-test p≥0.05). Zhu 2018[123] performed a receiver operating characteristic (ROC) analysis and derived an optimal cut-off of 242.5 10^9/L to distinguish patients with a diagnosis to patients with benign polyps in a Chinese population. Sensitivity was 62% and specificity was 72% (c-statistic=0.71, 95% CI=0.68-0.74). Among three studies with available data, on average, patients with a bowel cancer diagnosis had 40.16 (SD=27.16) 10^9/L higher platelet count than patients with benign polyps up to six months earlier. Platelet count for both groups was not assessed for any other time strata.

2.3.2.10 Mean platelet volume

Findings from each study that analysed mean platelet volume are in Table A9 (Appendix A11). All studies were in the 0-6 month time strata. All analyses indicated a statistically significant association with diagnosis. Two studies also assessed diagnostic performance of mean platelet

volume and derived cut-offs to distinguish between patients with and without a diagnosis. Kilincalp 2015[68] reported a cut-off of 8.25 fL, with sensitivity 54%, specificity 76%, and c-statistic 0.72. Wu 2019[120] did not report the cut-off but reported corresponding sensitivity 93%, specificity 45%, and c-statistic 0.66 (95% CI=0.60-0.71). On meta-analysis of four studies with available data, the pooled estimate indicated no difference in mean platelet volume between patients with and without a diagnosis of bowel cancer (0.06 (95% CI=-0.36-0.49)) fL (Figure 2.4). Although there was a statistically significant difference in each individual study, the pooled estimate indicates no difference due to some studies reporting higher levels in patients diagnosed and others reporting lower levels. Heterogeneity among the five studies was high (I^2=95.9%).

Figure 2.4: Forest plot of mean difference in mean platelet volume (fL) between patients with and without a diagnosis of bowel cancer up to six months later*



Abbreviations: SD=standard deviation, CI=confidence interval.
* This figure is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

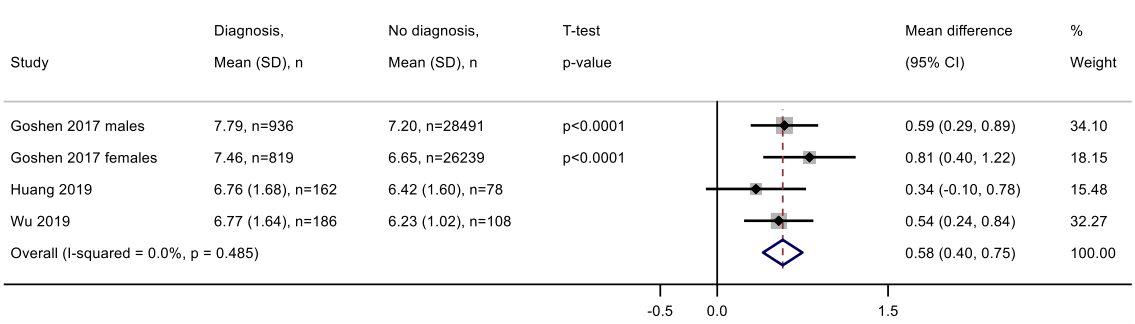
Two studies compared mean platelet volume between patients with benign polyps and patients diagnosed: Wu 2019[120] and Zhu 2018[123]. Among them, mean platelet volume was on average 0.39 (SD=1.04) fL higher in patients diagnosed compared to patients with benign polyps. Zhu 2018[123] also performed a ROC analysis and derived an optimal cut-off of 9.25 fL to distinguish

patients with a diagnosis and patients with benign polyps in a Chinese population, with sensitivity 69%, specificity 59%, and c-statistic 0.66 (95% CI=0.66-0.69).

2.3.2.11 White blood cell count

Findings from each study that analysed white blood cell count are in Table A10 (Appendix A12). No study was in the 12-36 month strata. In all other strata, all analyses showed patients with a diagnosis have higher white blood cell count than patients without. Up to 36 months before diagnosis, all studies reported a statistically significant difference between patients with and without a diagnosis, except two. Huang 2019[102] reported no statistically significant difference in mean white blood cell count at the time of diagnosis compared to patients without a diagnosis in a Chinese population (t-test $p \geq 0.05$). Firat 2016[94] categorised white blood cell count and reported no statistically significant association with bowel cancer in a Turkish population (chi-squared p=0.463). The authors did not report the white blood cell categories, number of individuals per category, or number with a diagnosis per category, making it difficult to assess the quality of the analysis. Three studies with available data were meta-analysed (Figure 2.5). The pooled estimate indicated that patients with a diagnosis up to six months from the white blood count measurement have on average 0.58 (95% CI=0.40-0.75) 10^9/L higher white blood count than patients without a diagnosis. There was little-to-no heterogeneity among the studies (I^2=0%).

Figure 2.5: Forest plot of mean difference in white blood cell count (10^9/L) between patients with and without a diagnosis of bowel cancer up to six months later*



Abbreviations: SD=standard deviation, CI=confidence interval.
* This figure is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

In measurements taken at least 36 months before diagnosis, studies reported mixed associations with bowel cancer (Table A10, Appendix A12). Data were not available for meta-analysis in other time strata.

Three studies compared white blood cell count between patients with benign polyps and patient with a diagnosis and each reported a statistical significant difference in mean levels up to six months prior to bowel cancer assessment ($p<0.05$). From two studies with available data, white blood cell count was on average 0.48 (SD=0.04) 10^9/L higher in patients diagnosed than patients with benign polyps. White blood cell count for both groups was not assessed for any other time strata.

2.3.2.12 Basophil count

Findings from each study that analysed basophil count are in Table A11 (Appendix A13). From one study with available data, the mean difference in basophil count between patients with and without a diagnosis up to six months later was <0.01 10^9/L. Goshen 2017[67] reported that higher basophil count statistically significantly increased odds of diagnosis in up to six months. Similarly,

Boursi 2016[57] reported higher basophil count increased odds of diagnosis in 6-12 months but this analysis failed to reach statistical significance (OR=1.34 (95% CI=0.93-1.95)). No study was in the 12-36 month and ≥36 month strata.

2.3.2.13 Basophil proportion

One study analysed this parameter (Table A12, Appendix A14). Kinar 2016[64] derived a machine-learning algorithm so did not report effect estimates or differences in FBC levels between outcome groups.

2.3.2.14 Eosinophil count

Findings from each study that analysed eosinophil count are in Table A13 (Appendix A15). From one study with available data, the eosinophil count was on average 0.03 10^9/L higher in patients diagnosed than patients not diagnosed. Goshen 2017[67] reported that higher eosinophil count statistically significantly increased odds of diagnosis in up to six months. Similarly, Boursi 2016[57] reported higher eosinophil count increased odds of diagnosis in 6-12 months but this analysis failed to reach statistical significance (OR=1.09 (95% CI=0.98-1.20)). No study was in the 12-36 month and ≥36 month strata.

2.3.2.15 Eosinophil proportion

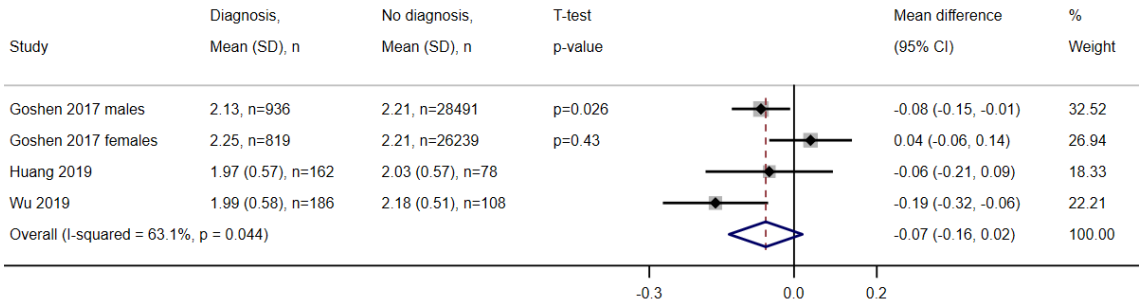
One study analysed this parameter (Table A14, Appendix A16). Kinar 2016[64] derived a machine-learning algorithm and did not report effect estimates or differences in FBC levels between outcome groups.

2.3.2.16 Lymphocyte count

Findings from each study that analysed lymphocyte count are in Table A15 (Appendix A17). All analyses indicated that patients diagnosed with bowel cancer have lower lymphocyte count up

to six months earlier compared to patients with no diagnosis, except Goshen 2017[67] for females. I meta-analysed the mean-difference from three studies with available data (Figure 2.6). The pooled estimate indicated that patients diagnosed have on average 0.07 10^9/L (95% CI=-0.16-0.02) lower lymphocyte count up to six months earlier compared to patients without a diagnosis, but this did not reach statistical significance. Heterogeneity between the studies was moderate (I^2=63.1%). Data was not available for meta-analysis in any other time strata.

Figure 2.6: Forest plot of mean difference in lymphocyte count (10^9/L) between patients with and without a diagnosis of bowel cancer up to six months later



Abbreviations: SD=standard deviation, CI=confidence interval.

Three studies compared lymphocyte count between patients with a diagnosis and patients with benign polyps and were all in the 0-6 month strata. Among two with available data, on average, patients diagnosed had 0.005 (SD=0.007) 10^9/L lower lymphocyte count than patients not diagnosed.

2.3.2.17 Lymphocyte proportion

Two studies analysed this parameter (Table A16, Appendix A18): Kinar 2016[64] and Zhou 2017[122]. The latter reported that patients diagnosed up to six months after the blood test was performed had statistically significantly lower lymphocyte proportion than patients without a diagnosis (median difference=11.2%) and patients with benign polyps (median difference=7.55%).

2.3.2.18 Monocyte count

Findings from each study that analysed monocyte count are in Table A17 (Appendix A19). All studies reported statistically significantly higher levels on average in patients diagnosed in up to six months from the blood test than patients not diagnosed. Two studies had data required for meta-analysis, which was therefore not performed. From the two studies, patients diagnosed had 0.07 (SD=0.02) 10^9/L monocyte count higher up to six months earlier compared to patients without a diagnosis, on average. Boursi 2016[57] reported that higher monocyte count statistically significantly increased odds of diagnosis in 6-12 months (OR=1.08 (95% CI=1.03-1.14)). No study was in the 13-36 month or ≥36 month strata.

One study, Wu 2019[120], assessed benign polyps. The authors reported that patients diagnosed had 0.03 10^9/L higher monocyte count up to six months earlier than patients with benign polyps.

2.3.2.19 Monocyte proportion

One study analysed this parameter (Table A18, Appendix A20). Kinar 2016[64] derived a machine-learning algorithm and did not report effect estimates or differences in FBC levels between outcome groups.

2.3.2.20 Neutrophil count

Findings from each study that analysed neutrophil count are in Table A19 (Appendix A21). All analyses in the 0-6 month strata indicated that patients diagnosed had statistically significantly higher neutrophil count than patients not diagnosed. Two studies had data required for meta-analysis, which was therefore not performed. Among them, patients diagnosed had 0.57 (SD=0.06) 10^9/L neutrophil count higher than patients not diagnosed. Boursi 2016[57] reported

that higher neutrophil count statistically significantly increased odds of diagnosis in 6-12 months (OR=1.24 (95% CI=1.21-1.27)). No study was in the 12-36 month or ≥36 month strata.

Two studies compared neutrophil count between patients diagnosed and patients with benign polyps and were both in the 0-6 month strata. Wu 2019[120] reported a mean difference of 0.35 10^9/L and Yang 2018[121] reported a median difference of 0.4 10^9/L, with higher levels on average in patients diagnosed.

2.3.2.21 Neutrophil proportion

Two studies analysed this parameter (Table A20, Appendix A22): Kinar 2016[64] and Zhou 2017[122]. The latter reported that patients diagnosed up to six months after the blood test was performed had statistically significantly higher lymphocyte proportion than patients without a diagnosis (median difference=9.75%) and patients with benign polyps (median difference=8.35%).

2.3.2.22 Ratio of parameters

I identified 32 analyses that assessed the association between a ratio of two FBC parameters and bowel cancer diagnosis (Table A21, Appendix A23). This excludes basophil, eosinophil, lymphocyte, monocyte, and neutrophil proportions (of white blood cell count), as these are considered their own FBC parameters and are described earlier. The most commonly analysed ratio was neutrophil-lymphocyte ratio (NLR) (n=16 analyses), followed by platelet-lymphocyte ratio (PLR) (n=9 analyses), mean platelet volume-platelet ratio (MPVPR) (n=4 analyses), and red blood cell distribution width-lymphocyte ratio (RDWLR) (n=3 analyses).

All analyses that reported the mean between patients with and without a diagnosis indicated statistically significantly higher levels of NLR, PLR, and MPVPR in patients diagnosed up to six

months after the blood test, on average, except Wu 2019[120], who reported statistically significantly lower MPVPR in patients diagnosed (mean=0.0330) than patients not diagnosed (mean=0.0447). Boursi 2016[57] reported that the odds of one-year diagnosis increase as NLR increases (OR=1.21, 95% CI=1.18-1.24).

Kilincalp 2015[68] also compared NLR and PLR in non-anaemic (haemoglobin ≥13 g/dL in men and ≥12 g/dL in women) cases to controls. Using FBC measurements taken at the time of diagnosis, NLR (mean difference=3.06, p<0.001) and PLR (mean difference=51.1, p<0.001) were both higher in non-anaemic cases than controls.

Compared to patients with benign polyps, NLR, PLR, and MPVPR were on average higher in patients diagnosed in all studies except in Wu 2019[120], who reported statistically significantly lower MPVPR in patients diagnosed. Mean differences between patients diagnosed and patients with benign polyps reached statistical significance in all studies except Huang 2019[102] (RDWLR p≥0.05) and Yang 2018[121] (NLR p=0.091, PLR p=0.059).

2.3.2.23 Microcytic anaemia

Panagiotopoulou 2014[111] compared the proportion of patients diagnosed up to three months after the FBC test between patients with microcytic anaemia (mean corpuscular volume <80 fL and haemoglobin < lower limit of normal for each centre) and patients without (Table A21, Appendix A23). Patients with microcytic anaemia had greater odds of diagnosis compared to patients with normocytic anaemia in both recruiting centres (OR=1.3, 95% CI=0.5-3.9 and OR=1.6, 95% CI=0.8-3.3) but both failed to achieve statistical significance. Patients with microcytic anaemia had greater odds of diagnosis compared to patients without microcytosis and without anaemia, which achieved statistical significance in one centre (OR=2.0, 95% CI=1.1–3.7) but not in the other (OR=1.7, 95% CI=0.6–4.5).

2.3.3 Appraisal of FBC-based prediction models (Aim 2)

Of the 53 studies included in the review, 16 developed or externally validated a FBC-based prediction model for bowel cancer diagnosis. Ten studies built 13 models among them (Table 2.4). The majority of models were built using UK patient data (62%, n=8).

Table 2.4: Details of FBC-based prediction models developed for bowel cancer detection (n=13)*

Article	Model name/description	Country	Outcome window	No. diagnosed/not diagnosed	Model building approach	Predictors
Boursi 2016	Laboratory model	UK	1 year	4929/11311	Logistic regression	Haematocrit, mean corpuscular volume, lymphocyte count, neutrophil-lymphocyte ratio
Boursi 2016	Combined model	UK	1 year	3375/8560	Logistic regression	Haemoglobin, mean corpuscular volume, white blood cell count, neutrophil-lymphocyte ratio, platelets, sex, previous metformin prescriptions, previous prescriptions for oral hypoglycemic drugs other than metformin
Cubiella 2016	COLONPREDICT	Spain	1 week	214/1358	Logistic regression	Change in bowel habit, rectal bleeding, benign anorrectal lesion, rectal mass, serum CEA, haemoglobin, faecal haemoglobin, previous colonoscopy, aspirin use, sex, age
Firat 2016		Turkey	At diagnosis		Machine-learning	Platelets, haemoglobin, sodium, total bilirubin, creatinine, calcium

Goshen 2017	Model for males	Israel	1-6 months	936/28491	Logistic regression	Haemoglobin, mean corpuscular volume, monocyte count, platelets, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, iron, ferritin
Goshen 2017	Model for females	Israel	1-6 months	819/26239	Logistic regression	Haemoglobin, mean corpuscular volume, neutrophil count, platelets, red blood cell distribution width, alanine aminotransferase, protein, iron, ferritin
Hippisley-Cox 2012	QCancer Colorectal males	UK	2 years		Cox regression	Alcohol status, family history of gastrointestinal cancer, haemoglobin, current/recent: rectal bleeding, abdominal pain, appetite loss, weight loss, or change in bowel habit in previous year
Hippisley-Cox 2012	QCancer Colorectal females	UK	2 years		Cox regression	Family history of gastrointestinal cancer, haemoglobin, rectal bleeding, abdominal pain, appetite loss, weight loss
Hippisley-Cox 2013	QCancer females	UK	2 years	3250/1240550	Logistic regression	Haemoglobin, family history of gastrointestinal cancer, alcohol status, abdominal distension,

						abdominal pain, appetite loss, rectal bleeding, weight loss, change in bowel habit, constipation
Hippisley-Cox 2013	QCancer males	UK	2 years	2607/1217648	Logistic regression	Haemoglobin, family history of gastrointestinal cancer, alcohol status, abdominal distension, abdominal pain, appetite loss, rectal bleeding, venous thromboembolism, weight loss, change in bowel habit, constipation
Kinar 2016	ColonFlag	Israel	3-6 months	2437	Machine-learning	Age, sex, all 20 FBC parameters
Marshall 2011	Bristol-Birmingham	UK	2 years	5477/38314	Logistic regression	Constipation, diarrhoea, change in bowel habit, flatulence, irritable bowel syndrome, abdominal pain/antispasmodic, rectal bleeding, haemoglobin, mean corpuscular volume, weight loss, deep venous thrombosis/pulmonary embolism, diabetes, obesity
Thompson 2017		UK	3 years	990/16413	Logistic regression	Age, sex, symptom combinations, physical signs, iron-deficiency anaemia, rectal bleeding, change in bowel habit, other colorectal cancer characteristics

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

2.3.3.1 Risk of bias

Risk of bias, assessed using the PROBAST tool, is reported in Table 2.5. The ten studies that developed prediction models (n=13 models) all scored high risk of bias in the analysis domain. Often, this was due to removing patients with missing data (whereas more accepted methods for handling missing data could have been used, such as multiple imputation) from all their analyses and not assessing or adjusting the developed model for overfitting or optimism, with no justification provided for these decisions. All nine studies externally validating models (n=11 models) were considered low risk of bias in the participants domain and all but one in the predictors domain, which scored unclear risk of bias. Six of the nine studies were considered to have high risk of bias in the analysis domain. This was commonly due to removing patients with missing data from analyses, with no justification provided for these decisions.

Table 2.5: Risk of bias (assessed using the PROBAST tool) in each study included in Aim 2 (n=16)*

Article	Model(s) name/description	Participants	Predictors	Outcome	Analysis
Development:					
Boursi 2016	1) Laboratory model 2) Combined model	Low	Low	Low	High
Cubiella 2016	COLONPREDICT	Low	Low	Unclear	High
Firat 2016		High	Unclear	Unclear	High
Goshen 2017	1) Model for males[1] 2) Model for females[1]	High	High	High	High
Hippisley-Cox 2012	1) QCancer Colorectal males 2) QCancer Colorectal females	Low	Low	Low	High
Hippisley-Cox 2013	QCancer males	Low	Low	Low	High
Hippisley-Cox 2013	QCancer females	Low	Low	Low	High
Kinar 2016	ColonFlag	Low	Low	Low	High

Marshall 2011	Bristol-Birmingham	Low	Low	Low	High
Thompson 2017		Low	Low	Unclear	High
Total low ***Total high*** ***Total unclear***		***7*** ***2*** ***0***	***7*** ***1*** ***1***	***5*** ***1*** ***3***	***0*** ***9*** ***0***
External validation:					
Ayling 2019	ColonFlag	Low	Unclear	Unclear	High
Birks 2017	ColonFlag	Low	Low	Low	Low
Collins 2012	1) QCancer Colorectal males 2) QCancer Colorectal females	Low	Low	Low	Low
Cubiella 2016	COLONPREDICT	Low	Low	Unclear	High
Hilsden 2018	ColonFlag	Low	Low	Unclear	High
Hornbrook 2017	ColonFlag	Low	Low	Low	High
Kinar 2016	ColonFlag	Low	Low	Low	High
Kinar 2017	ColonFlag	Low	Low	Low	High
Marshall 2011	1) Bristol-Birmingham 2) CAPER[2]	Low	Low	Low	Low
Total low ***Total high*** ***Total unclear***		***9*** ***0*** ***0***	***8*** ***1*** ***0***	***6*** ***0*** ***3***	***3*** ***6*** ***0***

[1]As recommended by the PROBAST study group, these models scored a high risk of bias in all domains because they were not internally or externally validated.
[2]The CAPER model, which includes haemoglobin level as a predictor, was not included in this review because it was never published as full-text. Only a published conference abstract is available[124].
* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

Goshen 2017[67] developed separate prediction models for males and females. As recommended by the Cochrane PROBAST study group[84], this study scored high risk of bias in all four domains because neither model underwent internal or external validation.

2.3.3.2 Model building approaches

Of the 13 models developed, 85% (n=11) were derived using statistical methods: logistic regression (n=9) and Cox regression (n=2). Two models, by Firat 2016[94] and Kinar 2016[64], were derived using machine-learning methodology. The former used multilayer perceptron artificial neural networks and the latter used ensembles of decision trees. A line listing of development details is provided in Table A22 (Appendix A24).

In seven models, variable selection was based on univariable analyses, with statistically significant variables subsequently included in a multivariable model. Of these, four models were further refined using backward selection, a process where variables in a model are iteratively removed based on pre-specified exclusion criteria. Four models used backward selection alone to build the model and one model, by Thompson 2017[118], used both forward and backward selection. One model, the ColonFlag by Kinar 2016[64], included all candidate variables.

A list of predictors in each prediction model is in Table 2.4. All models included haemoglobin as a predictor, except the Laboratory model by Boursi 2016[57], which included other FBC parameters.

2.3.3.3 Modelling FBC parameters

Of the 13 models, nine (from seven studies) included FBC parameters as categorised groups but only three studies stated the reason for the choice of categories. Cubiella 2016[92] defined each haemoglobin category as per established clinical guidelines. Goshen 2016[67] based their decision on a data-driven approach for both the model for males and females, categorising many FBC parameters into quintiles, with cut-offs not reported. Thompson 2017[118] categorised haemoglobin to model the presence or absence of anaemia but the levels used to define anaemia were not reported, nor how they were chosen.

In four models that kept the FBC data as continuous, two models, both developed by Boursi 2016[57], included them as fractional polynomials (non-linear variables) and two models, by Kinar 2016[64] and Firat 2016[94], kept them as-is (linear variables).

2.3.3.4 Correlation between FBC parameters

FBC parameters may be considered correlated because they are taken from the same patient at the same time. Additionally, some are mathematically dependent on others (further details are in Chapter 3). It is important to assess correlation in prediction model building, as including correlated predictors in a model can under/over-estimate the effect of the predictors. Of the 13 models developed, five from three studies reported assessing correlation between FBC parameters during model development.

Boursi 2016[57] developed two models: the Combined model and Laboratory model. In the Combined model, correlation between red blood cell count and lymphocyte count with other predictors was assessed. The authors did not report what the other predictors were, the methods used to assess correlation, or why red blood cell count and lymphocyte count were removed from the model instead of the correlated predictors. In the Laboratory model, two sets of mathematically correlated FBC parameters were included, despite the authors assessing correlation (using unknown methods). Firstly, haematocrit and mean corpuscular volume, where the latter is a ratio of the former and red blood cells. Secondly, lymphocyte count and neutrophil-lymphocyte ratio, where the latter is the neutrophil count divided by the former. Consequently, their combined model is likely to under/over-estimate risk of diagnosis.

Cubiella 2016[92] assessed correlation between predictors but did not state which predictors, the methods used to assess correlation, or present the results.

Goshen 2017[67] built a model for males and females separately. For each model, the authors assessed correlation between FBC parameters using logistic regression. All parameters statistically significant in univariable analysis were included in the final model. This approach is unorthodox and incorrect because the significant parameters may still be correlated with non-significant parameters, regardless of their significant association with the outcome.

2.3.3.5 Importance of FBC parameters

Kinar 2016[64] included all 20 FBC parameters in their machine-learning algorithm. The authors assessed which FBC parameters most increased the accuracy of predictions of diagnosis in 3-6 months from the latest FBC. For each FBC parameter, the difference in the AUC between models with and without the parameter was derived. Their analysis suggested that red blood cell-related parameters were the most important for identifying cases, of which haemoglobin was the important parameter, increasing the AUC by approximately 0.1. Platelet-related parameters were the next to contribute the most and white blood cell-related parameters contributed the least.

Firat 2016[94] defined (relative) importance of each covariate in the final model as the ratio of each log-transformed p-value from a Chi-squared test of association with bowel cancer presence divided by the log-transformed p-value corresponding to the maximum p-value from the entire set of covariates in the model. Their analysis identified platelets as the most important predictor, followed by haemoglobin.

2.3.3.6 Model reporting

Due to the nature of machine-learning methodology, prediction models derived by Kinar 2016[64] and Firat 2016[94] have no functional form (i.e. equation) so were not reported in the study

articles or referenced elsewhere. Using these models for independent external validation becomes impossible.

Where statistical methods were used, a full risk equation is needed to make predictions about an individual's risk of diagnosis. For example, this includes the intercept from logistic regression and baseline survival/hazard estimate from Cox regression, together with the effect estimates for all covariates in the final model. Of the 11 models developed using statistical methods, Cubiella 2016[92] provided the full risk equation in their article, reporting both the intercept and effect estimates for all covariates in a logistic regression model (named COLONPREDICT). No other study reported the full risk equation. Access to the QCancer model (which identifies many cancers, including colorectal) for males[99] and females[100] by Hippisley-Cox 2013 was provided in the articles as a reference to a website but the baseline survival/estimate needed for identifying risk of diagnosis remains unknown.

2.3.3.7 Internal validation

Internal validation is a process that addresses problems in the specification and estimation of a model, primarily overfitting, where a model appears to perform too optimistically in new patients[125]. Overfitting can be adjusted for using penalisation or shrinkage techniques, such as reducing the coefficients in a model to less extreme values based on performance measures in a separate dataset. Among the nine studies, there were on average 372,245.4 patients in the internal validation sample, ranging from 4,946 to 679,174.

Nine of the 13 models underwent internal validation (Table 2.6). All nine models used random data splitting to obtain the internal validation sample, where observations are randomly assigned to a development sample or validation sample. In addition to data splitting, Kinar 2016[64] used cross-validation for internal validation, which is an extension to data splitting

involving several split samples and iteratively building the model on all but one sample, with the one remaining sample used to testing. The authors estimated performance measures of the ColonFlag at 15 outcome time windows before diagnosis. I report performance measures from the split sample approach for the primary 3-6 month risk outcome but also from the cross-validation approach for the two-year outcome for comparability with other models (Table 2.6).

All nine models that underwent internal validation assessed discrimination, which measures how well the model distinguishes between patients with and without the outcome. All models used the c-statistic (also referred to as the c-index or AUC) to assess discrimination. QCancer Colorectal for males by Hippisley-Cox 2012[69] had the highest c-statistic (0.91, 95% CI=0.90-0.91), corresponding to two-year risk of diagnosis. The ColonFlag model by Kinar 2016[64] had the lowest c-statistic (0.72, 95% CI=0.69-0.75), corresponding to 22-24-month risk of diagnosis, with higher c-statistics reported at outcome windows closer to diagnosis, including the primary 3-6 window (0.82, 95% CI=0.79-0.84).

Seven of the nine models were assessed for calibration, which identifies how closely model predictions match the observed data. Calibration plots (of observed vs. predicted risk) were used to assess calibration for each model, and showed that predictions matched the observed probabilities adequately. Additionally, the Hosmer-Lemeshow goodness-of-fit test was performed by Kinar 2016[64] (p=0.47) and Thompson 2017[118] (p=0.430) to assess calibration.

No study reported assessing or correcting model performance statistics for optimism or adjusted their derived model for overfitting.

Table 2.6: Performance statistics from internal (n=9) and external (n=11) validation of 13 FBC-based models (Aim 2)*

Article	Model name/description	Primary outcome window	No. cases	No. non-cases	Discrimination: C-statistic (or AUC) (95% CI)	Calibration
Internal validation:						
Boursi 2016	Laboratory model	1 year			0.77 (0.75-0.78)	
Boursi 2016	Combined model	1 year	1702	3324	0.73 (0.71-0.74)	Calibration plot
Firat 2016		At diagnosis			0.81	
Hippisley-Cox 2012	QCancer Colorectal males	2 years			0.91 (0.90-0.91)	Calibration plot
Hippisley-Cox 2012	QCancer Colorectal females	2 years			0.89 (0.88-0.90)	Calibration plot
Hippisley-Cox 2013	QCancer males	2 years	125	667261	0.90 (0.90-0.91)	Calibration plot
Hippisley-Cox 2013	QCancer females	2 years	1356	655311	0.89 (0.88-0.90)	Calibration plot
Kinar 2016	ColonFlag	3-6 months 22-24 months	698		0.82 (0.79-0.84) 0.72 (0.69-0.75)	Calibration plot; Hosmer-Lemeshow: p=0.47
Thompson 2017		3 years	636	10966	0.86 (0.84-0.87)	Calibration plot; Hosmer-Lemeshow: p=0.430
External validation:						
Ayling 2019	ColonFlag		21	571		
Birks 2017	ColonFlag	3-6 months 6-12 months 12-24 months 18-24 months	5935 6821 5744 5141	2478764 2429503 2328636 2220108	0.84 (0.84-0.85) 0.81 (0.81-0.82) 0.79 (0.79-0.80) 0.78 (0.78-0.78)	

		24-36 months	7360	2102947	0.75 (0.75-0.76)	
Collins 2012	QCancer Colorectal males	2 years	2036	1057729	0.92 (0.91-0.92)	Calibration plot
Collins 2012	QCancer Colorectal females	2 years	1676	1074099	0.91 (0.90-0.92)	Calibration plot
Cubiella 2016	COLONPREDICT	1 week	136	1345	0.92 (0.90-0.94)	
Hilsden 2018	ColonFlag	1 year	60	8704		
Hornbrook 2017	ColonFlag	6 months	900	16195	0.80 (0.79-0.82)	
Kinar 2016	ColonFlag	3-6 months	5061	20552	0.81 (0.80-0.83)	Calibration plot; Hosmer-Lemeshow: p<0.001
Kinar 2017	ColonFlag	18 months	133	112451		
Marshall 2011	Bristol-Birmingham	2 years	349	1744	0.92 (0.91-0.94)	
Marshall 2011	CAPER[1]	2 years	5477	38314	0.79 (0.77-0.80)	

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

[1]The CAPER model, which includes haemoglobin level as a predictor, was not included in this review because it was never published as full-text. Only a published conference abstract is available[124].

2.3.3.8 External validation

External validation assesses the performance of the final model (after amendments made during internal validation) in an external dataset. Eleven models underwent external validation (Table 2.6). The ColonFlag model by Kinar 2016[64] was commonly externally validated (n=6 studies). The remaining models were externally validated once. There were on average 455,610.1 patients included in the external validation sample, ranging from to 592 to 2,225,249 among the 11 models.

Eight models were assessed for discrimination, each using the c-statistic. Birks 2017[65] reported the lowest c-statistic, for diagnoses in 24-36 months (0.75, 95% CI=0.75-0.76) using the ColonFlag by Kinar 2016[64] in a UK population. Cubiella 2016[92] and Marshall 2011[107] reported the highest c-statistic, for one-week risk of diagnosis by COLONPREDICT (0.92, 95% CI=0.90-0.94) and two-year risk by Bristol-Birmingham (0.92, 95% CI=0.91-0.94), respectively.

Three models were assessed for calibration, each using calibration plots. Kinar 2016[64] used unorthodox methods, comparing predicted risk in one cohort to observed risk in different cohorts, which does not reflect the accuracy (and therefore calibration) of the ColonFlag model. Collins 2012[90] show the predicted risk of diagnosis from and QCancer Colorectal models were similar to the observed risk for lower- and middle-risk UK patients but over-predicted risk slightly in the highest risk group. Kinar 2016[64] also performed a Hosmer-Lemeshow test that indicated the ColonFlag does not fit the UK data adequately ($p<0.001$).

Three studies externally validated the ColonFlag and did not assess discrimination or calibration, providing only sensitivity and specificity for various risk groups. This may be because the ColonFlag is a monotone score and not an absolute risk score so cannot be calibrated using conventional methods.

2.3.3.9 Reliability of existing models

Larger sample sizes are generally known to improve reliability of results[126, 127]. Where the sample size was known, models were derived using large sample sizes, although QCancer for males by Hippisley-Cox 2013[99] included relatively low number of events (n=125). Similarly, all external validations included large sample sizes, although the external validation of the ColonFlag by Ayling 2019[87] and Hilsden 2018[98] had relatively low number of events.

None of the 13 models assessed or corrected the model for overfitting and no article reported model performance statistics adjusted for optimistic performance so may appear better than they do in reality. Table 2.7 gives full details.

The four FBC-based prediction models with the highest c-statistic from internal and external validation (all c-statistics >0.85) were the QCancer Colorectal model for males and females by Hippisley-Cox 2012[69], Bristol-Birmingham model by Marshall 2011[107], and the model by Thompson 2017[118]. Each of these models relies on reported symptoms to identify risk of diagnosis for outcomes over a long period: two-year risk (both QCancer Colorectal models and Bristol-Birmingham) or three-year risk (model by Thompson 2017[118]). Recent evidence indicates that the majority of diagnoses in symptomatic patients are made within six months before first reporting symptoms to a GP[13]. For the four models, it is unclear how many early diagnoses there were within the two or three year outcome period (following reporting of symptoms) and if early events drive the good performance reported for each model. A more reliable approach to assess the relevance of symptoms and how well they perform for risk of diagnosis in two or three years should only include diagnoses around the two- or three-years after reporting them. This was not assessed in any internal or external validation for any of these

models. We therefore likely cannot consider the high performance reported to correspond to identification of cases in two or three years.

The Hippisley-Cox 2012[69] QCancer Colorectal models include "current" and "recent" symptoms as covariates. These terms are not adequately described or defined in their article, such as symptoms experienced in the last day, week, month, or 3 months, and these criteria are likely subjective decisions made by GPs. It is unclear how the timing of the symptoms entered into the model affects predicted risk.

The red blood cell distribution width FBC parameter has historically been unavailable in UK primary care so is often unavailable in corresponding electronic records[128]. The ColonFlag model, which uses this parameter to identify a score for each individual, was externally validated in UK data by Kinar 2016[64], Birks 2017[65], and Ayling 2019[87]. Kinar 2016[64] removed this parameter from the model and validated the resulting model in UK data. At minimum, the derived model should undergo external validation so the Kinar 2016[64] external validation is incomplete. Birks 2017[65] reported 98% of FBCs had missing red blood cell distribution width but most tests had a ColonFlag score available. It is therefore likely that a score was derived (by Medial EarlySign) using a modified ColonFlag algorithm, with red blood cell distribution width removed. Ayling 2019[87] reported that the ColonFlag score was not calculated for patients with missing FBC parameters but did not report which parameters were missing.

Table 2.7: Considerations of existing FBC-based prediction models developed for bowel cancer detection (n=13)*

Article	Model name/description	Outcome window	Consideration
Boursi 2016	Laboratory model	1 year	• Not assessed/corrected for overfitting or optimistic performance
Boursi 2016	Combined model	1 year	• Not assessed/corrected for overfitting or optimistic performance • Includes mathematically correlated covariates
Cubiella 2016	COLONPREDICT	1 week	• Not assessed/corrected for overfitting or optimistic performance
Firat 2016		At diagnosis	• Not assessed/corrected for overfitting or optimistic performance • Machine-learning methods not transparent or easy to incorporate into practice • Not designed to predict risk of future diagnosis
Goshen 2017	Model for males	1-6 months	• Not assessed/corrected for overfitting or optimistic performance • Not validated
Goshen 2017	Model for females	1-6 months	• Not assessed/corrected for overfitting or optimistic performance • Not validated
Hippisley-Cox 2012	QCancer Colorectal males	2 years	• Not assessed/corrected for overfitting or optimistic performance • Symptom-covariates not adequately defined (i.e. “currently” or “recently”) • Timing of covariate measured potentially not suitable for outcome window
Hippisley-Cox 2012	QCancer Colorectal females	2 years	• Not assessed/corrected for overfitting or optimistic performance • Symptom-covariates not adequately defined (i.e. “currently” or “recently”) • Timing of covariate measured potentially not suitable for outcome window
Hippisley-Cox 2013	QCancer females	2 years	• Not assessed/corrected for overfitting or optimistic performance • Timing of covariate measured potentially not suitable for outcome window
Hippisley-Cox 2013	QCancer males	2 years	• Not assessed/corrected for overfitting or optimistic performance • Timing of covariate measured potentially not suitable for outcome window • Relatively low number of events

Kinar 2016	ColonFlag	3-6 months	• Not assessed/corrected for overfitting or optimistic performance • Machine-learning methods not transparent or easy to incorporate into practice • Derives a monotone score, not a risk score • Commercially developed and costly to the NHS • UK external validations are often of a modified ColonFlag (red cell distribution width removed) • Relatively low number of events in some external validation studies
Marshall 2011	Bristol-Birmingham	2 years	• Not assessed/corrected for overfitting or optimistic performance • Timing of covariate measured potentially not suitable for outcome window
Thompson 2017		3 years	• Not assessed/corrected for overfitting or optimistic performance • Timing of covariate measured potentially not suitable for outcome window

2.3.4 Use of repeated measures

Among all 53 studies, 75% (n=40) analysed a single FBC blood test from each patient, 15% (n=8) used multiple per patient, and it was unclear whether a single or multiple were used in 9% (n=5). In the 40 studies that analysed a single test, the first available FBC (from study entry) per patient was used in 10% (n=4) of studies, the last FBC (closest to the time of outcome assessment) in 15% (n=6), and it was unclear which was chosen in 75% (n=30). Among the eight studies that analysed multiple FBC tests per patient, one (Goldshtein 2010[95]) explored associations with bowel cancer diagnosis, one (Boursi 2016[57]) developed a prediction model only, one (Kinar 2016[64]) developed and externally validated a prediction model, and five studies externally validated an existing model (ColonFlag model).

Goldshtein 2010[95] used mixed-effects modelling to identify changes in haemoglobin over a 10-year period between Israeli patients with and without a diagnosis. Patients diagnosed had a greater decline in haemoglobin over (log-transformed) time (slope=0.3, intercept=12.72) than patients not diagnosed (slope=0.04, intercept=13.111). Levels in patients diagnosed diverged from those in patients not diagnosed approximately four years beforehand and the rate of decline increased in the run up to diagnosis.

Boursi 2016[57] assessed the association between changes in FBC levels and diagnosis in a UK population. Changes were calculated as the difference in blood levels between the two most recent FBCs available per patient prior to outcome assessment. The time interval between the two sets of FBCs was not summarised so may vary substantially across patients. The authors reported an association between change in mean corpuscular volume (OR=1.23, 95% CI=1.01-1.51) and mean corpuscular haemoglobin (OR=1.27, 95% CI=1.04-1.55) with diagnosis. Changes in red blood cell count (OR=1.09, 95% CI=0.97-1.23), haematocrit (OR=1.04, 95% CI=0.98-1.11),

and white blood cell count (OR=0.97, 95% CI=0.94-1.00) were not significantly associated with diagnosis.

Kinar 2016[64] incorporated trends over three sets of FBCs into their ColonFlag model: measured at 36 months, 18 months, and 3-6 months before outcome assessment. It is unclear why only three FBCs or these chosen time points were used. The Israeli model uses all 20 FBC parameters to derive a monotone score between 0-100. Due to the nature of machine-learning, the algorithm was not reported in the article. It is therefore unknown how higher variations in FBC levels over time gives higher scores and how these scores reflect future diagnoses of bowel cancer.

2.4 Discussion

This systematic review identified that the association between the FBC test and bowel cancer detection has been studied for over 20 years, with the earliest paper published in 1995. Most studies were published in the five years before the search date (2015-2019), indicating an increasing interest in the use of the FBC for detecting bowel cancer.

2.4.1 FBC for bowel cancer (Aim 1)

2.4.1.1 Summary of main findings

This review identified many studies that suggest the FBC is influenced by the presence of bowel cancer. Patients diagnosed up to six months after the FBC was measured had lower levels of red blood cells, haemoglobin, and mean corpuscular volume and higher levels red blood cell distribution width, white blood cell count, and platelets compared to patients not diagnosed. The findings for these parameters were consistent at earlier outcome windows before diagnosis but the differences between patients with and without a diagnosis remained small and confined within approximately five years prior to diagnosis. This highlights that the FBC may assist early

detection, potentially before overt symptoms appear, but FBC levels in patients with a future diagnosis may reside in the normal reference range of blood levels, which represent little-to-no cause for concern and differ between health systems. Such small changes in the FBC may not be obvious to a clinician, resulting in no for further investigation unless there were other signs or symptoms.

On meta-analysis, mean platelet volume and lymphocyte count did not statistically significantly differ between patients with and without a diagnosis. Other FBC parameters, including haematocrit, mean corpuscular haemoglobin, and differential white blood cell counts were not assessed by enough studies to draw conclusions regarding their ability for detection. These FBC parameters require further research to determine their prognostic ability to detect bowel cancer.

2.4.1.2 Comparison with existing literature

To my knowledge, no previous reviews of the FBC blood test for bowel cancer detection have been reported. However, a review of symptoms for bowel cancer, which included low haemoglobin[129], and red blood cell distribution width for a range of cancers including bowel[130] exist.

This review identified similar results for haemoglobin reported in the review of bowel cancer symptoms by Del Giudice 2014[129]. They reported that low haemoglobin levels are associated with increased likelihood of diagnosis. Furthermore, haemoglobin <13 g/dL for males and <12 g/dL for females is an established criteria currently aiding referral for further investigation in general practice[53, 54, 131]. The focus of their review was on symptoms and not the FBC blood test so no other FBC parameter was reviewed. This review therefore builds on their findings, suggesting that many other FBC parameters could also be used to support referral.

2.4.2 Appraisal of FBC-based prediction models (Aim 2)

2.4.2.1 Summary of main findings

Among the 13 FBC-based models, poor methods were often employed and many were inadequately reported. In result, all 13 models scored a high risk of bias in the analysis domain of the PROBAST tool. For example, some studies excluded patients with missing prognostic factor data from the study entirely, did not adequately define their prognostic variables, or did not assess correlation between variables. Additionally, many studies used data from electronic health records but only two assessed data quality, which is often considered an important step because these datasets were not primarily designed for use in research studies and may require some data validation before analysis[132-135]. Furthermore, only one study, Cubiella 2016[92], reported the full risk-equation in their article, which is required to perform independent external validations and make risk predictions in practice.

All models were either internally or externally validated, except the two models by Goshen 2017[67]. Most models were insufficiently validated, with only one of discrimination and calibration often reported and neither in some studies, although the ColonFlag machine-learning algorithm is a monotone score so calibration cannot be assessed using traditional methods. Furthermore, the FBC-based prediction models may not perform as well as reported because no model assessed or reported adjusting for overfitting or optimistic performance.

Prediction models with long-term outcome windows may not identify cases as well as reported in the study articles. For example, symptomatic patients are likely to be diagnosed soon after reporting symptoms to primary care[13, 112] so the good performance for long-term outcomes is likely driven by early events and may therefore be misleading. Independent external validation

studies provide an opportunity to assess such scenarios, but no such assessment was reported among the external validations included in this review.

2.4.2.2 Comparison with existing literature

To my knowledge, no previous reviews of FBC-based prediction models for bowel cancer detection have been reported. A systematic review of all identified prediction models for bowel cancer reported in 2000-2014 was published in 2016[136], but the focus was not on the FBC blood test and its ability to identify cases. Additionally, none of the models included in this review were included in that review: five published before 2014 that were likely missed from their review and the remaining eight published after 2014.

2.4.3 Summary of repeated measures analysis

Three studies assessed the association between change in FBC levels over repeated measurements and bowel cancer diagnosis. They suggest trends in some FBC parameters may indicate bowel cancer and these become apparent within 3-4 years prior to diagnosis. The particular FBC parameters with trends influenced by bowel cancer presence were contradicted across the studies. Further research is needed in this area.

2.4.4 Recommendations for future research

Here, I provide recommendations for preparation and analysis of FBC data and areas for future research.

2.4.4.1 Use appropriate analytic methods for FBC analysis

Often, the FBC parameters were categorised for analysis, causing loss of information and reduced statistical power[137]. My recommendation is that the FBC be analysed as continuous variables to allow for an efficient analysis, such as assessing non-linear relationships between

trends and diagnosis. If categorised, data-driven approaches should not inform the choice of categories. Instead, this should be based on prior evidence or current clinical guidelines[54, 131].

For prediction models, researchers should perform correlation analyses and only include FBC parameters with no/weak correlation. For example, mathematically related parameters likely have high correlation and should not be included together in a model to prevent biased estimates.

2.4.4.2 Account for missing data

Nine of the 20 FBC parameters are measured from the blood sample and 11 are summary parameters derived using mathematical equations that describe the measured parameters (see Chapter 3 for more details). I recommend researchers derive missing measured parameters by applying mathematical equations to available summary parameters. Imputation methods may subsequently be used for measured parameters, if appropriate. Any missing data among summary parameters can then be derived by applying these mathematical equations to maintain functional relationships between parameters between measured and summary parameters.

2.4.4.3 Research change in FBC levels over time

Studies that analysed repeated FBC measures for bowel cancer detection used sub-optimal or non-transparent methods and showed inconsistent results. I recommend further research be performed using efficient methodology that explores whether changes over time are related to future diagnosis, such as mixed-effects modelling, as these changes may be better indicators of bowel cancer than the blood levels alone.

2.4.4.4 Research FBC thresholds for referral

Currently, haemoglobin <13 g/dL for men and <12 g/dL for women aids referral for further investigation[53, 54, 131]. This review identified that red blood cell distribution width 13.95-14%, platelets >242.5 10^9/L, and mean platelet volume 8.25-9.25 fL may be possible criteria to support referral. However, future studies should first formally assess the performance of these cut-offs and combinations of parameters for bowel cancer detection.

2.4.4.5 Research the use of FBC for early-stage diagnosis

No study reported on the association of the FBC test and early-stage diagnosis. Such analyses would be useful to assess the ability of the FBC to identify early-stage disease, where likelihood of survival is higher: five-year survival is approximately 93% for stage 1 diagnosis and 10% for stage 4 diagnosis[6, 7]. Further research is needed to identify the clinical usefulness of the FBC test for early-stage diagnosis.

2.4.4.6 Choice of outcome time window

This review suggests changes in FBC levels are most relevant within four years prior to diagnosis, as differences in many parameters between patients with and without a diagnosis become apparent in this window. I therefore recommend studies use outcome windows within this four-year period. This time window may also be clinically relevant for early detection and improved survival.

2.4.4.7 Assess overfitting of prediction models

No FBC-based prediction model reported assessing (during internal validation) or adjusting the model for overfitting or optimistic performance, where performance measures appear too large. These adjustments are important to derive reliable risk predictions and performance measures.

I recommend new prediction models undergo internal validation and assess/adjust the model for overfitting using established methods such as shrinkage (of regression coefficients)[138].

For each model, the internal validation sample was obtained using random data splitting. This approach can be inefficient, particularly with small sample sizes, as they further reduce the sample size used for model building. I recommend using bootstrapping or cross-validation, which resample from the existing dataset.

2.4.4.8 Assess model discrimination and calibration

For an efficient validation, I recommend internal and external validation studies assess both discrimination and calibration of models. The c-statistic is often used to assess discrimination but the D-statistic may also be appropriate for survival models[139]. Calibration, which was not always assessed by studies included in this review, should be assessed using a calibration plot of observed versus predicted risk of diagnosis, where possible, and can be supported by other estimates, such as the calibration intercept and slope. Overall performance of a model can also be assessed using statistics such as the R^2 or Brier score[140].

2.4.4.9 Robust external validation

External validation studies should, and often did, employ the same methods used in the study that developed the model undergoing validation. However, I recommend these validations be more robust and critical. For example, they should consider the appropriateness of the patient sample and the likely course of disease in this group. Additionally, performance should be assessed using short- and long-term events separately, especially if the model is designed for earlier detection and use wide outcome time windows. This would give a more reliable overview of model performance for long-term outcomes. A robust assessment of a model's performance will improve its integrity.

2.4.4.10 Reporting models

Many aspects of study design and results were often not included in the studies. For example, age and sex of the patients and the number of patients with and without a diagnosis were often not reported. I recommend studies follow the appropriate reporting guidelines for transparency. These include STROBE[141] for observational studies, RECORD[142] for studies using electronic health records, and TRIPOD[143] for prediction model studies.

Studies developing statistical prediction models should report the full risk-equation. This would allow models to be used by readers, independently validated, and embedded into practice for testing and actual usage.

2.4.5 Strengths & limitations

Strengths:

A strength of this systematic review is that it is the first review of the FBC test for bowel cancer detection. It brings together many individual studies to provide a comprehensive review of the current literature, including the relevance of the FBC and individual parameters, reliability of existing prediction models, and quality of evidence. Another strength is that this review identified many grey areas and future research needed.

Limitations:

Only published, full-length articles were included in this review. Two relevant abstracts were identified but were excluded during screening because they provided limited data for extraction and review and no corresponding full-length article was published (confirmed by the authors)[124, 144]. Additionally, some titles and abstracts did not mention the FBC or bowel cancer but the main text may have. Therefore, some relevant studies may have been missed during screening.

I was unable to obtain full texts for some articles and hence could not assess the study's eligibility at full text screening.

Many studies analysed only two or three of the 20 FBC parameters. This resulted in small number of studies for most of the parameters for which I could not perform a thorough review or draw any conclusions regarding their prognostic ability for bowel cancer detection.

The pre-specified analysis involved meta-analysing effect estimates for diagnosis, such as odds ratios. Most studies did not report the same type of effect estimate for meta-analysis. I derived many odds ratios using data available in the study articles, but these were often not comparable so could not be pooled. Comparability was defined as FBC data being analysed using the same methods across studies, such as all as a continuous variable or all as categorical. Where categorised, different cut-off levels and number of categories were used. Furthermore, the time interval between FBC measurements and diagnosis varied across studies so there were often less than three comparable odds ratios within each time strata.

I derived the mean difference by measuring graphs in Zhu 2018[123]. I also derived the SE of the mean difference using the nearest p-value approach for Goshen 2017[67] and Zhu 2018[123]. This could have slightly under/over-estimated these estimates and influenced the weight of the study on the pooled estimate in the meta-analysis. However, these were the best estimates I could obtain for inclusion in the meta-analysis.

Thresholds for FBC levels to distinguish between patients with and without a diagnosis were often not derived or reported. I was therefore unable to analyse and recommend cut-off values for most parameters to assist referral and diagnosis in general practice.

3

Preparation and quality assessment of FBC data from CPRD

This chapter has previously been published[128]:

Virdee P.S. *et al*. Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96* (2020). DOI: 10.1186/s40537-020-00375-w. Published under a CC-BY licence – permissions not required.

3.1 Introduction

Electronic health records (EHRs) are software implemented in clinical systems. Databases, such as the Clinical Practice Research Datalink (CPRD), extract from these EHR systems to hold anonymised patient data obtained by health services, such as hospitals and GP clinics. EHRs were introduced to improve efficiency in clinical practices by aiding patient care and monitoring clinical services. For example, to manage patient appointments and offer a repeat prescriptions service. Their use for research is a secondary development and has become increasingly popular over the last decade, with patient anonymity ensured. EHRs allow the conduct of large-scale medical research studies with large sample sizes.

Data cleaning is a common form of data preparation before analysis in studies with active patient involvement and data collection, such as clinical trials. However, data quality assessments and

validation are often not performed in studies using EHR data. The systematic review reported in Chapter 2 identified that only 6% of 53 eligible studies reported performing data validation before analysis[57, 94, 113]. A recent systematic review identified many barriers to performing EHR data quality checks, which include handling large amounts of unstructured data, challenges with patient identification, matching and data extraction, and unfamiliarity with performing data quality assessments[145]. Such checks have been reported to be crucial, with all individual studies in one systematic review agreeing that data accuracy and data completeness were key factors to consider when designing EHR studies[146]. A second review highlighted a need for a generalised approach to assess EHR data quality[147].

This data validation had two main aims:

1. To derive a methodological approach to assess EHR data quality, specifically laboratory data from CPRD.
2. To apply this approach to the FBC blood test to derive a more reliable dataset for analyses (performed for subsequent chapters).

As laboratory data are frequently used across medical research (the systematic review in Chapter 2 identified many studies that used FBC data alone (n=512)[51]), I provide recommendations and guidance for researchers who wish to access and analyse EHR data. I have provided the statistical program used to perform the data validation in the corresponding publication for other researchers to use[128].

3.2 The Clinical Practice Research Datalink

3.2.1 Source of data

GP Systems of Choice (GPSoC) was a framework that supplied computation systems and services to clinical practices in England and ended in 2018[148]. Various computation systems were offered

and primary care practices could choose a system that best suited their data management needs. The CPRD is a national database of patient records from EHRs among contributing GP practices in the UK that use the Vision and, more recently, EMIS software systems[149]. Anonymised patient records are extracted by the CPRD through regular downloads. The extraction process does not require active intervention by practices. Over 2,000 primary care practices contribute to CPRD in the UK. The CPRD data includes around 60 million patients, of which around 16 million are currently registered active patients. A full description of CPRD can be found on the CPRD website[150, 151].

3.2.2 Database structure & coding

The CPRD GOLD Data Specification explains that the database consists of 10 individual datasets[152]. Each dataset holds different bits of information from a patient's record. In each dataset, patient records are coded using medical codes (or 'medcodes'), which are the numeric equivalent of Read codes available in patient records in the GP system and are specific to the CPRD database[153, 154].

Read codes are a coded thesaurus of clinical terms, providing a standard vocabulary for clinicians to record patient findings and procedures across clinical practice. Read codes have been used in the NHS since 1985, with version 2 and version 3 (also called Clinical Terms Version 3) actively used until April 2016 [155], when they were replaced with SNOMED CT codes, which provide the most comprehensive and precise clinical terminology and are used in health care internationally[156]. Cross-maps of both Read codes version 2[157] and version 3[158] to SNOMED CT codes are publicly available by NHS Digital. Cross-maps of Read codes version 2 and 3 and SNOMED CT codes to CPRD medical codes are available in the CPRD medical dictionary, available from CPRD staff on request[149]. Other existing coding systems, such as the International

Classification of Diseases-10 (ICD-10) for identifying diseases and malignancies, are available in linked EHRs only[159].

The 'Test' dataset is one of the 10 CPRD GOLD datasets. It holds records of tests and examinations performed in primary care, including laboratory tests. The Test dataset is coded using both medical codes (translated from Read codes, as described above) and entity codes. The latter are unique numbers that identify the measurement from every type of laboratory test available in patient records. For example, red blood cell count and white blood cell count are measured as part of a FBC blood test and assigned a unique entity code of 194 and 207, respectively. The entity code is generated by the CPRD system and is not visible to staff at the primary care practice – it is used internally by the CPRD to provide a convenient way to file data that preserves the original data structure in practice. Both medical and entity codes are therefore specific to the CPRD database.

Data items in the Test dataset include the anonymised patient identification number, laboratory test result, corresponding medical and entity code, date of test, and unit of measurement[152]. A snapshot of a CPRD Test data extract (with patient data removed) is in Figure 3.1.

Figure 3.1: The Faecal Immunochemical Test for traces of blood in stool samples[1]

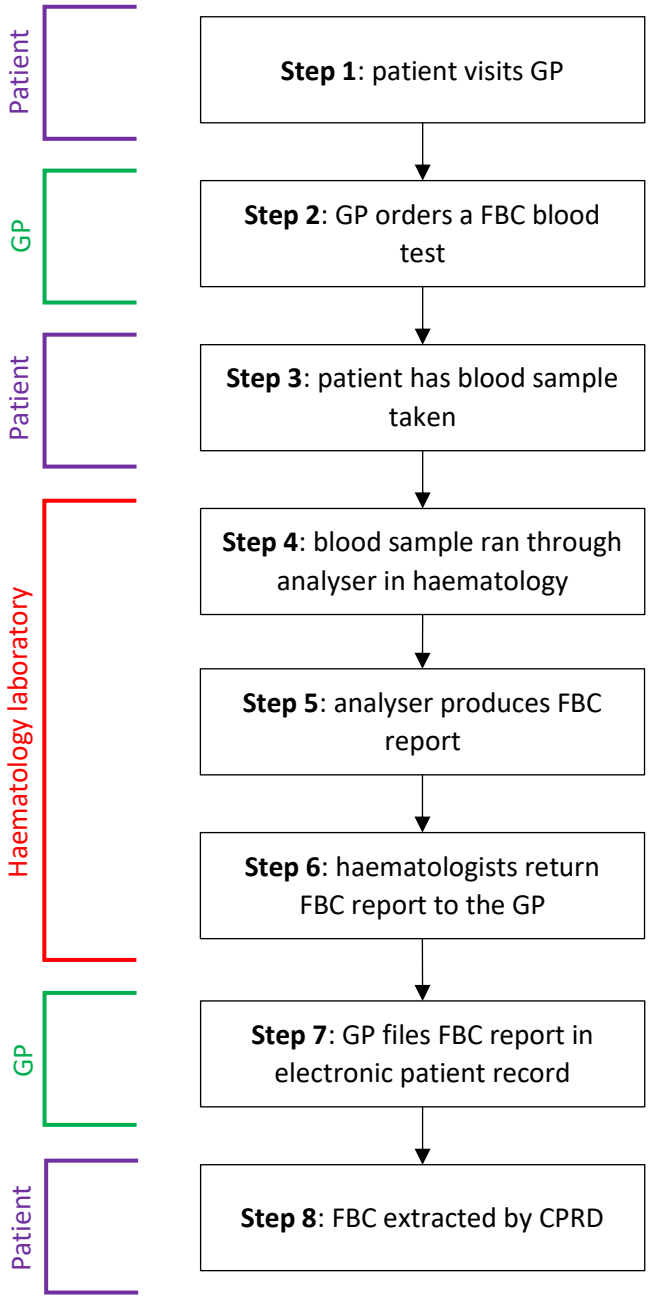
	epatid	eventdate	medcode	enttype	data2	data3
1	.	.	.	.	.	.
2	.	.	.	.	.	.
3	.	.	.	.	.	.
4	.	.	.	.	.	.
5	.	.	.	.	.	.
6	.	.	.	.	.	.
7	.	.	.	.	.	.

[1]Data items: epatid=patient identifier, eventdate=date of test, medcode=medical code, enttype=entity code, data=measurement value, data=unit of measure

3.3 The full blood count process

I was given a tour of a haematology laboratory in the John Radcliffe Hospital by Michael Jacobs, a senior haematologist. I witnessed how the FBC blood test is processed and analysed and results returned to the GP. I combined the lessons learnt from this tour with discussions from various haematologists (described later) and knowledge of the CPRD database provided by Alice Fuller, the CPRD representative involved in this research. I describe these lessons in this section. A flow chart describing the FBC process is in Figure 3.2.

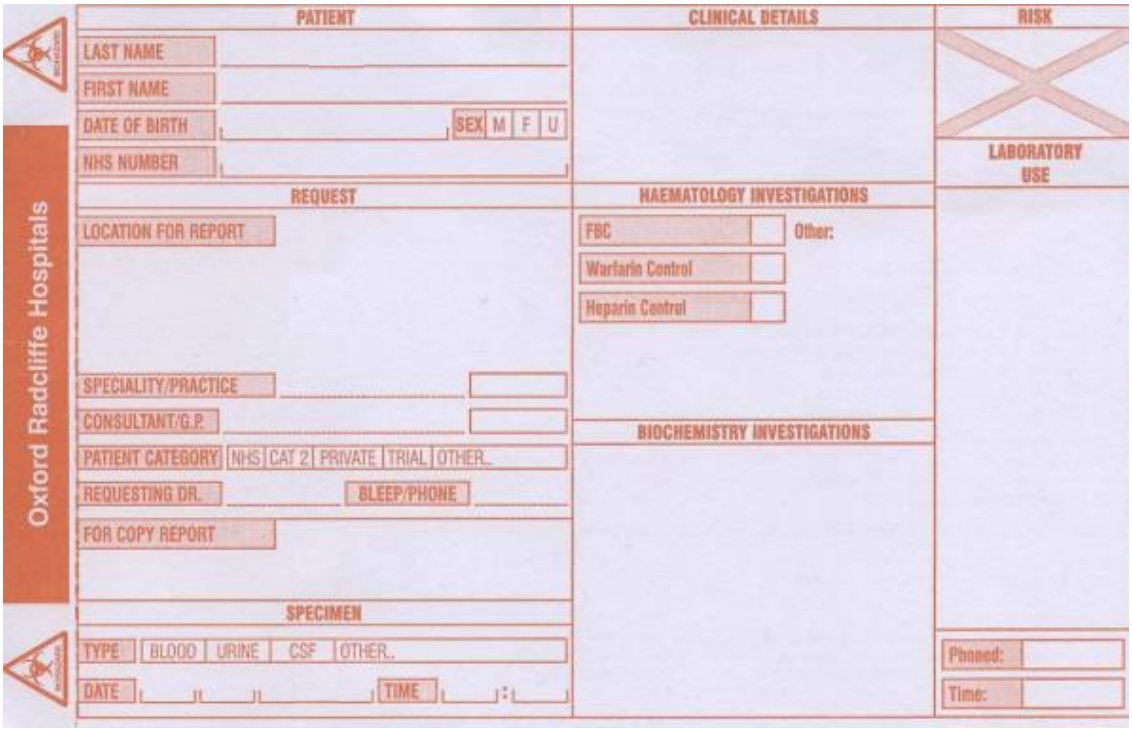
Figure 3.2: The full blood count process



A patient may make an appointment to visit their primary care GP (**step 1**). A GP may then order a FBC blood test, which is ordered electronically or, less commonly, using a paper form (Figure 3.3) (**step 2**). Once ordered, a sample of blood is required and taken from the patient by health care professionals (**step 3**). The blood sample, labelled using the patient's name and NHS number, is delivered to a haematology laboratory and run through an analyser (a machine) that

processes the sample (**step 4**). Historically, not all parameters were always derived by analysers, possibly due to technology catching up with changes in general practice as new parameters become available. In the last decade or so, analysers have derived the values for all 20 FBC parameters. Nine of the parameters are measured directly from the blood sample: red blood cell count, haemoglobin, platelet count, basophil count, eosinophil count, lymphocyte count, monocyte count, and neutrophil count. The remaining 11 parameters describe the nine measured parameters and are derived using mathematical formulae programmed into the analyser: haematocrit (or packed cell volume), mean corpuscular volume, mean corpuscular haemoglobin, mean corpuscular haemoglobin concentration, red blood cell distribution width, mean platelet volume, basophil percentage, eosinophil percentage, lymphocyte percentage, monocyte percentage, and neutrophil percentage. A FBC report is then generated by the analyser and made available to haematology laboratory staff (**step 5**). The report includes the derived FBC blood levels and corresponding unit of measurements and Read code (version 2) for each FBC parameter, assigned by the haematologists.

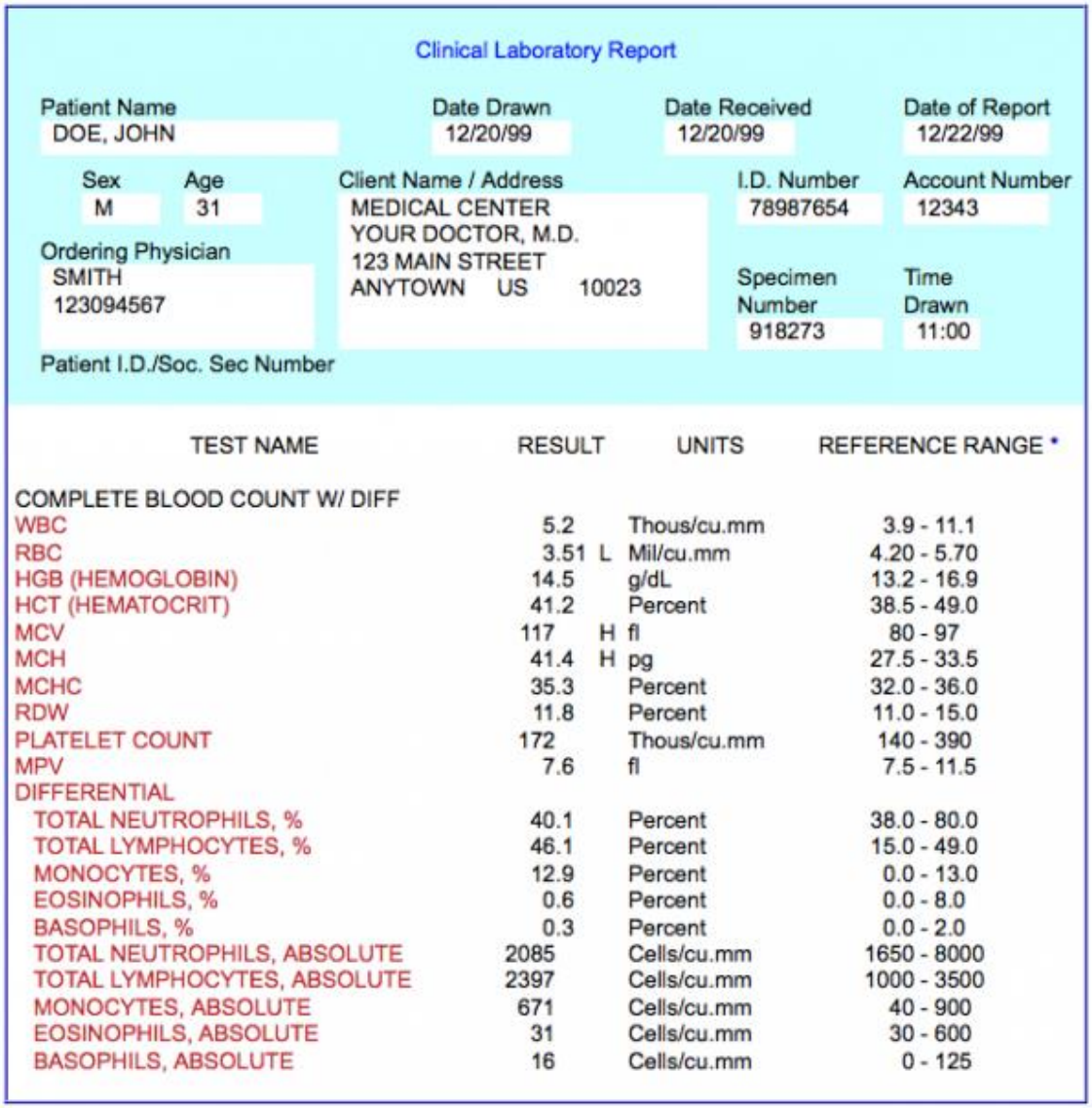
Figure 3.3: Example of haematology request paper form – includes the full blood count[1]



[1]This figure is taken and adapted from: Platt, A. The Department of Laboratory Haematology - OUH Haematology Laboratory Handbook. 2020 [Accessed 2 August 2021]; Available from: https://www.ouh.nhs.uk/services/departments/laboratory-medicine/haematology-laboratories/documents/haematology-handbook.pdf[160]. Figure is copyright-free (confirmed by the author) and permission granted for re-use in this thesis (Appendix B1).

The FBC report is labelled with the patient's name and NHS number and electronically returned to the GP practice that ordered the test (**step 6**). An example of a FBC report (with dummy data) is in Figure 3.4. At the primary care practice, the FBC report is electronically assigned to the patient's electronic record, a process that is largely automated (**step 7**). The report is examined by a clinician, who then decides on the appropriate action to take, e.g. refer the patient for further testing. The FBC report, now filed in the patient's electronic record, contributes to the CPRD database at its next download, with each FBC parameter extracted from the GP's patient record by the EHR system (**step 8**). Specifically, Read codes (assigned by the haematology staff) are extracted and converted to medical codes (or 'medcodes'). During this process, entity codes are assigned to each FBC parameter automatically by the EHR system.

Figure 3.4: Example of a (dummy) FBC report[1]



Clinical Laboratory Report

Patient Name: DOE, JOHN | Date Drawn: 12/20/99 | Date Received: 12/20/99 | Date of Report: 12/22/99

Sex: M | Age: 31 | Client Name / Address: MEDICAL CENTER, YOUR DOCTOR, M.D., 123 MAIN STREET, ANYTOWN US 10023 | I.D. Number: 78987654 | Account Number: 12343

Ordering Physician: SMITH 123094567 | Specimen Number: 918273 | Time Drawn: 11:00

Patient I.D./Soc. Sec Number

TEST NAME	RESULT		UNITS	REFERENCE RANGE *
COMPLETE BLOOD COUNT W/ DIFF				
WBC	5.2		Thous/cu.mm	3.9 - 11.1
RBC	3.51	L	Mil/cu.mm	4.20 - 5.70
HGB (HEMOGLOBIN)	14.5		g/dL	13.2 - 16.9
HCT (HEMATOCRIT)	41.2		Percent	38.5 - 49.0
MCV	117	H	fl	80 - 97
MCH	41.4	H	pg	27.5 - 33.5
MCHC	35.3		Percent	32.0 - 36.0
RDW	11.8		Percent	11.0 - 15.0
PLATELET COUNT	172		Thous/cu.mm	140 - 390
MPV	7.6		fl	7.5 - 11.5
DIFFERENTIAL				
TOTAL NEUTROPHILS, %	40.1		Percent	38.0 - 80.0
TOTAL LYMPHOCYTES, %	46.1		Percent	15.0 - 49.0
MONOCYTES, %	12.9		Percent	0.0 - 13.0
EOSINOPHILS, %	0.6		Percent	0.0 - 8.0
BASOPHILS, %	0.3		Percent	0.0 - 2.0
TOTAL NEUTROPHILS, ABSOLUTE	2085		Cells/cu.mm	1650 - 8000
TOTAL LYMPHOCYTES, ABSOLUTE	2397		Cells/cu.mm	1000 - 3500
MONOCYTES, ABSOLUTE	671		Cells/cu.mm	40 - 900
EOSINOPHILS, ABSOLUTE	31		Cells/cu.mm	30 - 600
BASOPHILS, ABSOLUTE	16		Cells/cu.mm	0 - 125

[1]This figure is taken from: POZ.com. Understanding Your Lab Work (Blood Tests). 2021 [Accessed 12 November 2021]; Available from: https://www.poz.com/basics/hiv-basics/understanding-lab-work-blood-tests[161]. Permission has been granted by the authors for re-use in this thesis (Appendix B1).

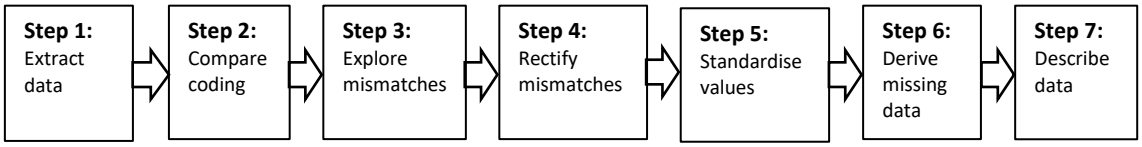
3.4 Methods

Access to the CPRD GOLD data was approved by the CPRD Independent Scientific Advisory Committee, who also provided ethical approval for the study (CPRD protocol: 14_195RMn2A2R). The CPRD data accessed spanned a study period of 1st January 2000 to 28th April 2015 (data cut date). This start date was advised by CPRD staff, with earlier data known to be of poorer quality[162]. Patients aged at least 40 years before the end of the study period with at least one

FBC blood test available in the Test dataset were extracted from CPRD and included in the study. This age group was chosen because the ColonFlag machine-learning algorithm was developed for adults aged at least 40 years[64]. Additionally, this age group allows for some FBC data prior to age 50, when bowel cancer incidence rates begin to rise steeply in the UK[163].

My approach to assess the quality of the laboratory data in the CPRD Test dataset is provided in Figure 3.5.

Figure 3.5: Methodological approach to assess laboratory data quality in CPRD



*This figure is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Stata/SE V15.1 was used for all analyses.

3.4.1 Step 1: extract FBC data

The Test dataset was downloaded from the CPRD database by Alice Fuller (CPRD representative) and provided to me for preparation and analysis. I actively searched through the list of medical and entity codes to identify FBC blood test records, which were reviewed by my supervisors. I validated the code list by comparing it to independent lists from relevant published studies and clinical code repositories[164-166]. I then kept only those records where either the medical code or entity code corresponded to a FBC parameter.

3.4.2 Step 2: compare medical and entity codes

I compared the medical and entity code assigned to each parameter by checking the types of medical codes assigned to each FBC-related entity code and vice-versa. FBC parameters were considered to be consistently coded if the medical and entity code corresponded to the same parameter and considered mismatched if otherwise.

I was provided with the CPRD entity code dictionary by Alice Fuller (CPRD representative), where I identified that six FBC parameters do not have existing entity codes: mean platelet volume, basophil proportion, eosinophil proportion, lymphocyte proportion, monocyte proportion, and neutrophil proportion. They were considered mismatched because they were assigned an entity code for a different FBC parameter by the EHR. Consequently, mismatches were stratified into three strata:

1. Mismatches with existing medical or entity codes for both parameters that could have been assigned.
2. Mismatches where the medical code suggested one of the six parameters without an existing entity code.
3. Mismatches where one coding system suggested a FBC parameter but the other suggested the measurement was not a FBC parameter.

3.4.3 Step 3: explore mismatches further

In the mismatched pairs, the parameter suggested by either coding system was checked to see if it was already available for that FBC test. For example, if the medical code suggested haemoglobin and entity code suggested neutrophil count, I checked whether (consistently coded) haemoglobin and neutrophil count were already available in that FBC.

3.4.4 Step 4: rectify mismatches

The mismatched pairs were classified as one of the two suggested parameters, where possible. The classification was based on the value and corresponding unit that best reflected the parameter and biological plausibility of the value.

3.4.5 Step 5: standardise FBC data to a single unit

The resulting dataset contained consistently coded or rectified FBC parameters. I then planned to standardise FBC levels to a single, conventional unit of measurement. Parameters were excluded if they had unreliable units, such as partially entered or where the value and unit did not appear to match. Subsequently, to aid standardisation, the extent of any extreme or implausible blood levels were identified to assess whether the blood levels or the units were incorrectly entered. This was done by categorising each parameter by percentiles and the mean, median, and range in each percentile calculated. All parameters were divided into 10 percentiles (deciles), except basophils, eosinophils, lymphocytes, monocytes, and neutrophils, which were divided into four percentiles (quartiles) because the range of possible values is very small such that deciles could not be derived.

3.4.6 Step 6: derive missing FBC levels using functional equations

Many FBC parameters are mathematically related. Therefore, missing FBC parameter values can often be derived using known values of other parameters from the same FBC blood test. As a pre-requisite to derivation, I examined the blood test date, which was the only indicator to separate each FBC in the CPRD data if a patient had multiple, to ensure I derived values in a single FBC test using other values within that same FBC.

I subsequently applied 25 known mathematical equations to derive missing FBC levels (see Appendix B2). Equations to derive haematocrit, red blood cell distribution width, and mean

platelet volume exist but rely on data not recorded in the CPRD database so could not be used. Once a missing FBC parameter was derived, it could subsequently be used to derive other missing FBC levels, if used in other equations. Therefore, I re-iteratively applied the 25 equations until no further values were being derived.

3.4.7 Step 7: describe the resulting data

The final dataset was described using mean with SD or median with range for continuous variables and counts and proportions for categorical variables. Descriptions include the number of parameters and FBCs available, extent of extreme or implausible blood values, and amount of missing data.

3.4.8 Discussions with haematologists to understand any issues

Once the seven steps described were executed, I spoke with haematologists and GPs to develop insight and understand any identified issues, particularly the mismatches in coding, reliability of units, and causes for missing data. Seven haematologists and nine GPs across the UK were randomly chosen and approached – the number approached was initially five each but increased due to low response rate. Discussions were carried out by email or telephone if an email address was unknown.

Discussion items with haematologists:

1. Do you electronically deliver FBC reports to the clinics that ordered them or are they also delivered by other means (e.g. post)?
2. Is the FBC report you deliver labelled with the patient's name and NHS number or does it include other identifiers?
3. Does the FBC report contain the units of measurement for each FBC parameter and are they always entered in full (i.e. no missing or partially entered units)?

4. Do you always include clinical codes in the FBC report for each blood level?
5. Are red blood cell distribution width and mean platelet volume always included in the FBC report?

Discussion items with GPs:

1. Do you use either the Vision or EMIS computer software in your practice?
2. Do you electronically receive FBC reports from labs or are they also received by other means (e.g. post)?
3. Is the FBC report labelled with only a name and NHS number or are other identifiers included?
4. How do you assign the FBC report to a patient's record?
5. Is there any manual entry or editing of the results?
6. Are red blood cell distribution width and mean platelet volume blood levels usually present in the FBC report?

Responses to the questions were not analysed in any specific way, but were instead used informally for me to develop a fundamental understanding of what happens in practice.

3.5 Results

The Test dataset provided by a CPRD representative (Alice Fuller) contained 695,139,617 laboratory test results from patients aged at least 40 years before the end of the study period among 658 primary care practices that use the Vision software system. Contribution to the CPRD from GP practices using the EMIS software system was not until a later date, hence practices using EMIS are not included in this CPRD extract.

3.5.1 Step 1: extract FBC data

There were 325 unique entity codes and 10,963 unique medical codes assigned to patient records in the Test dataset. Table 3.1 contains medical and entity codes I identified to be for the FBC. These codes were approved by my supervisors and were consistent with existing lists[164-166]. Keeping records with any of these codes assigned resulted in 228,707,454 FBC parameters among 2,914,589 patients in the data.

Table 3.1: Entity and medical codes for the FBC blood test in CPRD (Test dataset)

FBC parameter	Entity code	Entity term (as in CPRD)	Medical code	Medical term (as in CPRD)
Red blood cell count	194	Red blood cell count	17	Red blood cell (RBC) count
			13788	Nucleated red blood cell count
			26931	RBC count NOS
			26932	RBC count normal
			26933	RBC count low
			44213	RBC count abnormal
			50182	Red cell mass
			57136	RBC count raised
			58853	RBC count borderline low
			70079	RBC count borderline raised
Haemoglobin	173	Haemoglobin	4	Haemoglobin estimation
			739	Anaemia unspecified
			795	Iron deficiency anaemias
			3942	Haemoglobin low
			10404	Hb estimation
			13755	Haemoglobin - sample sent
			26272	Haemoglobin borderline high

			26908	Haemoglobin abnormal
			26909	Haemoglobin normal
			26910	Haemoglobin borderline low
			26912	Haemoglobin estimation NOS
			26913	Haemoglobin high
			33284	Haemoglobin requested
			35749	Haemoglobin very low
			39601	Haemoglobin not estimated
			41531	Haemoglobin very high
Haematocrit (packed cell volume)	312	Packed cell volume	40	Haematocrit
			99	Haematocrit – PCV
			14240	Packed cell volume
			19836	Packed cell volume – PCV
			23476	Haematocrit - PCV – high
			27143	Haematocrit - PCV – NOS
			27144	Haematocrit - PCV – normal
			27145	Haematocrit - PCV – low
			41478	Haematocrit - PCV – abnormal
			55365	Haematocrit - borderline low
			62347	Haematocrit - borderline high
Mean corpuscular volume	182	Mean corpuscular volume	10	Mean corpuscular volume (MCV)
			2480	MCV – raised
			13774	Mean cell volume
			26920	MCV – NOS
			26921	MCV – normal
			26922	MCV - low
			41160	MCV - borderline raised

			52874	MCV - borderline low
Mean corpuscular haemoglobin	180	Mean corpuscular haemoglobin	20	Mean corpusc. haemoglobin(MCH)
			23214	Mean cell haemoglobin
			26917	MCH – normal
			26918	MCH – low
			40170	MCH – NOS
			47174	MCH - borderline raised
			49225	MCH - abnormal
			51616	MCH - borderline low
			61951	MCH - raised
Mean corpuscular haemoglobin concentration	181	MCH Hb Concentration	30	Mean corpusc. Hb. conc. (MCHC)
			26919	MCHC – raised
			39202	MCHC - NOS
			47345	MCHC - normal
			55183	MCHC - low
			64474	MCHC - borderline low
			72488	MCHC - borderline raised
Red blood cell distribution width	361	RBC red blood cell size	64	Red blood cell distribution width
			1191	RBC's – anisocytosis
			1962	RBC's – macrocytic
			9933	RBC's - microcytic
			19837	RBC - red blood cell size
			51484	Red blood cell size normal
			52050	Red blood cell size NOS
Platelets	189	Platelets	7	Platelet count
			3320	Thrombocythaemia
			4006	Thrombocytopenia

			4415	Platelet count abnormal
			26926	Platelet count NOS
			26927	Platelet count normal
			37666	Platelet aggregation test
Mean platelet volume	-	-	14166	Mean platelet volume
White blood cell count	207	Total White Blood cell count	15	Total white cell count
			1955	Leucopenia
			3372	Leucopenia
			4760	Leucocytosis
			4996	White cell count abnormal
			13817	White blood count
			13818	White cell count
			18516	Leucocytosis
			22293	Leucocytosis -high white count
			26325	Leucopenia - low white count
			26946	Total white cell count NOS
			26947	Total white blood count
			26948	White cell count normal
			38198	Leucocyte count
			45115	Diff. white cell count normal
			48015	Total WBC (IMM)
			48341	Polymorphonuclear leukocyte count
			53865	Leucocytosis
			92372	Leukocytosis
Basophil count	313	Basophil count	25	Basophil count
			27146	Basophil count NOS

			27147	Basophil count normal
			27148	Basophil count abnormal
			53404	Basophilia
Basophil proportion	-	-	14096	Percentage basophils
Eosinophil count	168	Eosinophil count	22	Eosinophil count
			13742	Eosinopenia
			18531	Eosinophil count raised
			26905	Eosinophil count NOS
			26906	Eosinophil count normal
Eosinophil proportion	-	-	19760	Percentage eosinophils
Lymphocyte count	208	Lymphocyte count	19	Lymphocyte count
			3189	Lymphocytosis
			11240	Lymphopenia
			23120	Lymphocyte count normal
			23121	Lymphocytosis - absolute
			26949	Lymphocyte count NOS
			26950	Lymphocyte count abnormal
			32932	Reactive lymphocyte count
			34551	Total lymphocyte count (IMM)
			37677	Lymphocytosis - relative
			42346	Abnormal lymphocytes
			74019	Lymphocyte function test
Lymphocyte proportion	-	-	17621	Percentage lymphocytes
Monocyte count	183	Monocyte count	21	Monocyte count

			9248	Monocytosis
			13776	Monocyte count NOS
			26923	Abnormal monocytes
			26924	Monocyte count normal
			26925	Monocyte count raised
			44189	Monocyte count abnormal
			72849	Monocytopenia
Monocyte proportion	-	-	13775	Percentage monocytes
Neutrophil count	184	Neutrophil count	18	Neutrophil count
			4463	Neutropenia
			13777	Neutrophil count NOS
			15725	Neutrophilia
			23112	Neutrophil count normal
			23113	Neutrophil function test
			31382	Neutrophil count abnormal
			105211	Band neutrophil count
Neutrophil proportion	-	-	17622	Percentage neutrophils

Abbreviations: FBC=full blood count, MCH=mean corpuscular haemoglobin, Hb=haemoglobin, RBC=red blood cell.

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

3.5.2 Step 2: compare medical and entity codes

An entity and medical code was assigned to each of the 228,707,454 FBC parameters. Both codes consistently corresponded to the same FBC parameter for 95.2% (n=217,752,448) of parameters and mismatches occurred in 4.8% (n=10,955,006) (Table 3.2). Mismatched pairs were among 94.1% (n=619) of the 658 practices and were present among FBCs performed during the study

period (2000-2015). Mean age at the start of the study period was 42.3 years vs. 42.8 years and 44.6% vs. 43.3% were male for patients with consistently coded parameters and mismatched pairs, respectively.

Table 3.2: Number of FBC parameters at different stages of the quality check

FBC parameter	1. In the CPRD Test dataset			2. After rectifying mismatches	3. After standardising	4. After deriving values
	Consistent[1]	Mismatched[1]				
		Using entity codes	Using medical codes			
RBC	17,496,396	-	2,233	17,496,431	14,483,956	14,989,332
Haemoglobin	18,129,224	425	3,318	18,136,578	13,241,660	15,956,530
Haematocrit	16,229,494	-	90	16,229,502	3,901,766	14,929,672
MCV	16,972,637	3	62	16,972,661	16,190,073	15,508,062
MCH	15,828,539	48	2,443	15,828,610	14,776,728	14,978,026
MCHC	11,600,421	-	99	11,600,434	7,872,690	14,758,965
RDW	7,187,721	-	92,334	7,187,769	319,775	312,052
Platelets	17,321,447	2,267,404	798	17,321,470	16,110,377	15,409,813
MPV	-	-	2,299,660	2,049,996	2,041,360	1,964,340
WBC	17,510,110	4	84,959	17,510,410	16,098,918	13,391,463
Basophil count	15,064,082	1,647,073	625	15,064,119	13,939,287	11,532,656
Basophil %	-	-	1,670,876	1,579,167	1,498,690	11,577,466
Eosinophil count	15,963,917	1,579,993	6,578	15,963,935	14,761,804	11,609,390
Eosinophil %	-	-	1,601,225	1,516,946	1,505,596	11,619,039
Lymphocyte	16,164,725	1,634,084	6,825	16,164,762	14,950,124	12,202,708
Lymphocyte %	-	-	1,616,430	1,531,345	1,508,370	12,195,026
Monocyte count	16,047,916	1,601,135	7,117	16,047,958	14,687,897	12,068,644
Monocyte %	-	-	1,625,740	1,537,874	1,510,233	12,076,052

Neutrophil count Neutrophil %	16,235,819 -	1,586,187 -	9,391 1,609,549	16,235,860 1,540,087	15,061,611 1,521,541	12,306,548 12,299,431
Not FBC		638,650	314,654			
Total (% of 228,707,454)	**217,752,448 (95.2%)**	**10,955,006 (4.8%)**	**10,955,006 (4.8%)**	**227,515,914 (99.5%)**	**185,982,456 (81.3%)**	**241,685,233 (105.7%)**

Abbreviations: FBC=full blood count; RBC=red blood cell count; WBC=white blood cell count; MCV=mean corpuscular volume; MCH=mean corpuscular haemoglobin; MCHC=mean corpuscular haemoglobin concentration; RDW=red blood cell distribution width; MPV=mean platelet volume.
[1]Consistent records are those where the medical code and entity code correspond to the same parameter. Those where this does not apply were considered to have mismatched coding.
* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Of the mismatched pairs, 44,349 were in strata 1, where existing medical and entity codes were available for both suggested parameters that could have been assigned. There were 10,272,007 in strata 2, where six parameters do not have an existing entity code and an alternative was assigned. The remaining 638,650 were in strata 3, where one code suggested a FBC parameter but the other suggested the measurement was not a FBC parameter. See Table 3.3 for further details.

Table 3.3: Description of mismatched pairs

FBC parameter (medical code)	**FBC parameter (entity code)**	**Number of mismatched pairs**
Strata 1[1]:		
White blood cell count (38198)	Lymphocyte count (208)	43,869
Neutrophil count (18)	Haemoglobin (173)	425
Mean corpuscular haemoglobin concentration (30)	Mean corpuscular haemoglobin (180)	48
Mean corpuscular volume (10)	White blood cell count (207)	4

Haemoglobin (795)	Mean corpuscular volume (182)	3
Strata 2[2]:		
Mean platelet volume (14166)	Platelets (189)	2,267,404
Basophil proportion (14096)	Basophil count (313)	1,647,073
Eosinophil proportion (19760)	Eosinophil count (168)	1,579,993
Lymphocyte proportion (17621)	Lymphocyte count (208)	1,590,215
Monocyte proportion (13775)	Monocyte count (183)	1,601,135
Neutrophil proportion (17622)	Neutrophil count (184)	1,586,187
Strata 3[3]:		
One indicates FBC but the other indicates anything other than FBC		638,650

[1]Where existing medical or entity codes for both parameters could have been assigned
[2]Where the medical code suggested one of the six FBC parameters without an existing entity code
[3]Where one code suggested a FBC parameter but the other suggested the measurement was not FBC-related
* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

There were 609 mismatched parameters where the entity code suggested a FBC parameter but the medical code was 0, which represents missing data (see Appendix B3).

3.5.3 Step 3: explore mismatches further

The 44,349 mismatched pairs in strata 1 belonged to 44,221 FBC tests. Almost all FBCs had only one mismatched pair in them (99.7%, n=44,098). Both parameters suggested by the entity and medical code were already available in that FBC for 96.7% (n=42,891) of mismatched pairs, one of the two available for 0.2% (n=68), and neither were available for 3.1% (n=1,390).

The 10,272,007 mismatched pairs in strata 2 belonged to 3,615,271 FBC tests. The majority of tests had only one mismatched pair (56.5%, n=2,042,743). The parameter suggested by the entity code was already available in that FBC for 94.5% (n=9,708,930) of mismatches and neither parameter suggested by the medical and entity were available for 5.5% (n=563,077). The parameter suggested by the medical code was not already present in any FBC.

The 638,650 mismatched pairs in strata 3 were among 357,315 FBC tests. Most tests had only one mismatched pair among them (64.3%, n=229,836). Among the mismatches, the parameter suggested by the medical code was already available in the same FBC for 2.0% (n=12,826) and by the entity code already available for 47.2% (n=301,292). Neither parameter suggested by the entity or medical code was already available in the same FBC test for 50.8% (n=324,532).

3.5.4 Step 4: rectify mismatches

In strata 1, 51 parameters suggested by the entity code and four suggested by the medical code were chosen. For 34,059 mismatches, it was unclear which of the two parameters the blood level and unit represented. The remaining 10,235 mismatched pairs had missing values so were not assessed.

In strata 2, 153,590 parameters suggested by the entity code and 9,471,831 by the medical code were chosen. For 602,633 mismatches, it was unclear which of the two parameters the blood level and unit represented. The remaining 43,953 had missing values.

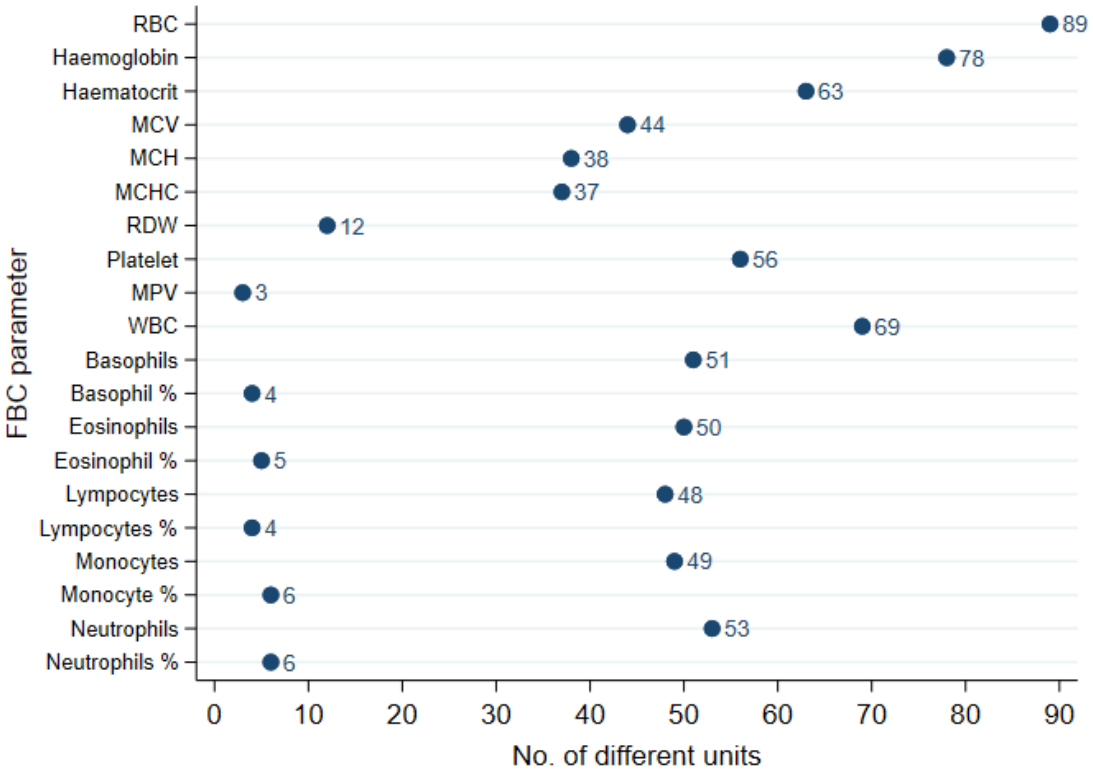
In strata 3, 24,925 parameters suggested by the entity code and 113,065 by the medical code were chosen. For 344,438 mismatches, it was unclear what the blood level and unit represented. The remaining 156,222 mismatches had missing values.

Of the 10,955,006 mismatched pairs overall, mean platelet volume was the most commonly rectified parameter (n=2,041,360) and 1,191,540 could not be rectified (see Appendix B4 for full details). The resulting dataset contained 227,515,914 consistently coded or rectified FBC parameters among 2,914,589 patients (see Table 3.2 for the number per parameter).

3.5.5 Step 5: standardise FBC data to a single unit

Red blood cell count had the most variability in number of units (n=89) (Figure 3.6). No apparent differences in blood levels and units of measurement were observed by practice or over time for each FBC parameter.

Figure 3.6: Number of different units entered for each FBC parameter (n=769 in total)



Abbreviations: FBC=full blood count, RBC=red blood cell count, WBC=white blood cell count, MCV=mean corpuscular volume, MCH=mean corpuscular haemoglobin, MCHC=mean corpuscular haemoglobin concentration, RDW=red blood cell distribution width, MPV=mean platelet volume.
* This figure is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Most units of measurement were unreliable in the resulting dataset – often missing, partially entered, or clearly wrong. For example, red blood cell count values reported in seconds (Appendix B5). For extreme blood levels, it was not possible to assess whether the unit was correct or if the level was measured in an alternative unit of measurement. Standardisation was therefore not performed. Alternatively, only parameters with the required units entered in the database were included in the dataset (Table 3.4 shows the final units of measure). The resulting dataset contained 2,870,006 patients with 81.3% (n=185,982,456) of the original 228,707,454 parameters.

Table 3.4: Descriptive statistics of each FBC parameter in the final derived dataset

FBC parameter	No. records	Mean (SD)	Median (min-max)	No. (%)[1] missing values
RBC (10^{12}/L)	14,989,333	4.3 (5.3)	4.4 (0-8577)	1,547,684 (9.4%)
WBC (10^9/L)	13,391,464	7.6 (46.6)	6.9 (0-123000)	3,145,553 (19.0%)
Haemoglobin (g/dL)	15,956,531	14.7 (14.2)	13.5 (0-13728.7)	580,486 (3.5%)
Haematocrit (L/L)	14,929,673	30.6 (49)	38.7 (0-121881.6)	1,607,344 (9.7%)
MCV (fL)	15,508,063	90.7 (15.4)	91 (0-10730)	1,028,954 (6.2%)
MCH (fL)	14,978,027	45.4 (214.4)	30.3 (0-33378.5)	1,558,990 (9.4%)
MCHC (g/dL)	14,758,966	318.3 (973.6)	33.5 (0-352127.7)	1,778,051 (10.8%)
RDW (%)	312,052	14 (1.7)	13.7 (0-43.8)	16,224,965 (98.1%)
Platelets (10^9/L)	15,409,814	266 (175)	256 (0-319000)	1,127,203 (6.8%)
MPV (fL)	1,964,340	9.6 (7.3)	9.4 (0-726)	14,572,677 (88.1%)
Basophil count (10^9/L)	11,532,657	0.1 (2.4)	0.1 (0-6500)	5,004,360 (30.3%)
Basophil proportion (%)	11,577,467	0.9 (1.7)	0.8 (0-2694.1)	4,959,550 (30.0%)
Eosinophil count (10^9/L)	11,609,391	0.2 (0.7)	0.2 (0-1218)	4,927,626 (29.8%)
Eosinophil proportion (%)	11,619,040	3.1 (3.4)	2.6 (0-5869.8)	4,917,977 (29.7%)

Lymphocyte (10^9/L)	12,202,709	2.1 (78.8)	1.9 (0-275000)	4,334,308 (26.2%)
Lymphocyte proportion (%)	12,195,027	28.5 (43.4)	28.1 (0-95486.1)	4,341,990 (26.3%)
Monocyte count (10^9/L)	12,068,645	0.6 (0.6)	0.5 (0-1218)	4,468,372 (27.0%)
Monocyte proportion (%)	12,076,053	7.8 (5.5)	7.5 (0-8590.6)	4,460,964 (27.0%)
Neutrophil count (10^9/L)	12,306,549	4.5 (7.1)	4.1 (0-12200)	4,230,468 (25.6%)
Neutrophil proportion (%)	12,299,432	60.4 (44.2)	60 (0-59623.5)	4,237,585 (25.6%)

Abbreviations: FBC=full blood count, RBC=red blood cell count, WBC=white blood cell count, MCV=mean corpuscular volume, MCH=mean corpuscular haemoglobin, MCHC=mean corpuscular haemoglobin concentration, RDW=red blood cell distribution width, MPV=mean platelet volume.
[1]Percentages are calculated out of the total number of FBC tests (n=16,537,017).

To identify the extent of extreme or implausible values, each parameter was categorised into deciles, except basophils, eosinophils, lymphocytes, monocytes, and neutrophils, which were categorised into quartiles. Summary statistics for each percentile were derived. From the lowest percentiles, there were 171 parameters with negative values and 6,327,555 entered as zero. The highest percentiles for each parameter showed a plausible median and interquartile range, suggesting relatively few extreme values (Appendix B6).

3.5.6 Step 6: derive missing FBC levels using functional equations

Firstly, 332 parameters with no test date were removed, as these parameters could not be matched to other parameters in the same FBC. Secondly, 8,091,664 parameters were removed, as multiple blood tests taken on the same day in a single patient could not be distinguished. The resulting dataset contained 177,890,460 parameters, 16,589,428 FBC tests, and 2,856,400 patients.

Before applying the equations in Appendix B2, approximately 11 parameters were available on average among the 16,589,428 tests. After application, approximately 16 were available on average. If a derived parameter was negative, the parameter and all mathematically associated

parameters were deleted due to implausibility, resulting in 23,541,367 parameters deleted. At a blood test level, I deleted 52,411 FBC tests that had no available data for any individual parameter.

3.5.7 Step 7: describe the resulting data

Of the original 2,914,589 patients with FBCs, 2.0% (n=58,569) were excluded after all amendments. Excluded patients had mean age 42.2 years at study entry and 47.2% were male. The final dataset contained 241,685,233 parameters from 16,537,017 FBCs among 2,856,020 patients. In the final dataset, patients had mean age 42.3 years at the start of the study period and 44.5% were male.

Approximately 15 parameters were available per FBC on average. The dataset contained more parameters than were originally available in the CPRD Test dataset, likely due to derivation (step 6) (see Table 3.2). All nine measured parameters were available for 67.1% (n=11,102,834) of FBCs and all 11 descriptive parameters for 0.3% (n=48,046). All 20 parameters were available in 0.3% (n=47,999) of FBCs.

Summary statistics for each parameter are in Table 3.4. All parameters had extreme or implausible values, as indicated by their minimum and maximum values. Missing data was most common for red blood cell distribution width (98.1%, n=16,224,965 tests) and least common for haemoglobin (3.5%, n=580,486 tests).

3.5.8 Discussions with haematologists to understand any issues

Four of seven haematologists approached responded (based in Oxford, Coventry, Barton, and Bristol). One of nine GPs approached responded (based in Oxford).

All haematologists gave similar answers to each question. In summary, they described that the FBC report was always sent electronically to the GP practice and was labelled with only the patient's name and NHS number. Haematologists always included units of measurement in the report and in full. Clinical codes for each blood level were always sent to the GP. Standard practice has been to supress the red blood cell distribution width from the FBC report because it was considered unhelpful to GPs, although no haematologist could explain who made this decision and why. Similarly, mean platelet volume had historically been supressed from FBC reports as standard practice, which changed over the last decade and is now included in FBC reports.

The GP who responded gave consistent responses to the haematologists. In summary, the FBC report is sent electronically by the haematologists and labelled with the patient's name and NHS number. All FBC levels are nearly always available except red blood cell distribution width, which was hardly ever included in the report. Units of measurement are also provided in this report. Once received, the electronic FBC report is 'dragged' into a patient's electronic record so there is no manual entry except in extreme cases, such as there being multiple people with the same name and GP staff referring back to the NHS number to find out to whom the report belongs.

3.6 Discussion

Research studies using laboratory data from EHR databases often do not report assessing data quality – three studies performed quality checks using FBC data in the systematic review (Chapter 2). This data preparation highlighted that EHR data quality should be assessed to ensure a fundamental understanding of the data and to derive a reliable dataset for analysis. I emphasise this further because the use of these databases for research was not the primary reason for their development. The findings support other studies that make similar statements, such as one study discussing opportunities and challenges of using EHR data to test prediction

models stating that data quality needs to be considered[167] and a systematic review highlighting the need for standardised methodology to perform EHR data quality checks[147]. My methodological approach could also help researchers adhere to the RECORD guidelines for observational studies using EHR data, which are an extension to the STROBE guidelines and include additional items regarding data cleaning and quality checks[142].

3.6.1 Summary of main findings

My methodological approach identified that around 5% of the FBC data haves mismatches in coding, with medical codes (translated from Read codes) and entity codes (from the EHR system) each suggesting different tests measurements. Mismatches were not common to particular practices, with around 94% of practices in the dataset having a FBC test with an inconsistently coded parameter. Additionally, no apparent differences in mismatches were observed over time. The procedure to assign entity codes to patient records within the EHR is unknown and the process of handling FBC results is automated (described in Section 3.3), so it is unclear why some mismatched. It is possible this is a mapping error in the CPRD database. I identified no other study that explored the consistency of these two CPRD coding systems. Often, one parameter suggested by either code was already available in that FBC. However, this did not necessarily mean that the second suggested parameter was the correct one, as other mismatched pairs within that same FBC could suggest that same parameter. Around 17% of the mismatched pairs were rectified based on the FBC value and unit of measurement. Standardising each blood level to a conventional unit was not possible due to many units being unreliable, such as details partially entered. Despite this, the majority of parameters were already recorded in standard units so the proportion of parameters dropped was relatively low.

Researchers may not be aware of the mathematical relationships between parameters. Only one study in the systematic review (Chapter 2) used these equations to derive missing blood

levels. A prior study compared various missing data imputation methods for clinical laboratory measurements including the full blood count, but did not consider these mathematical relationships[168]. My approach to resolve missing data resulted in a dataset that contained more known FBC parameters than originally available. It was anticipated that this was due to new parameters being introduced over time, but I found no difference in availability of parameters over time (from 2000 to 2015).

Around 2% of patients with FBCs were removed from the original CPRD Test dataset through this data quality check, resulting in a relatively large sample of FBCs. Age and sex were the only patient characteristics available and were balanced between patients included and excluded. I therefore consider the final dataset to be representative of the overall sample.

3.6.2 Causes of missing data

When exploring causes of missing data, I identified that all 20 FBC parameters are derived automatically from haematology analysers, but this has historically not always been the case. This could explain why the original CPRD dataset had on average 11 of 20 parameters available per FBC, with technology catching up to changes in practice as new parameters become available. Reasons for missing data are not recorded in CPRD so this cannot be confirmed.

Missing data was most common for red blood cell distribution width and mean platelet volume, unknown for approximately 98% and 88% of FBC tests, respectively. Historically, these parameters were derived by laboratory analysers but the output suppressed before the FBC report is sent to the GP practice. Discussions with haematologists identified that these parameters were not considered helpful or meaningful and standard practice was to exclude these blood levels from FBC reports. This is likely why many FBC tests in CPRD have missing data for these parameters. In recent years (after the 2015 data cut date), standard haematological

practice changed and the two parameters are more commonly available in primary care patient records. Discussions with Michael Jacobs, a senior haematologist, identified that missing data in other parameters may be a result of analysers malfunctioning or deriving strange blood levels that require active intervention from haematologists, who may supress that blood level from the FBC report to the GP.

3.6.3 Conclusion

My methodological approach could help tackle barriers to assessing data quality from EHRs identified in existing systematic reviews and help researchers improve the quality of research findings[145, 147]. My approach identified many considerations when preparing EHR data and I advise researchers to perform data quality checks to understand the extent of any issues. Although routine datasets provide a large sample size for analysis, I emphasise that the reliability of the data should be prioritised.

Box 1: Tips for future researchers on performing EHR data quality assessments

1. Use a methodological approach (such as my statistical programming, which is available on request) as a basis for EHR data validation, adapting to other EHRs as necessary.

2. Perform (or have close involvement with EHR staff who perform) data preparation to better identify the accuracy of the dataset and develop a fundamental understanding of the acquisition and representativeness of EHR datasets.

3. Use existing functional relationships to derive missing FBC levels.

4. Invest in powerful laptops or advanced computing facilities that are efficient for processing large datasets and purchase external hard drives for data storage.

5. Incorporate data quality checking into study timelines. The process can take many months but is important to derive a reliable dataset for analysis.

3.7 Handling missing data for subsequent analyses

Missing data can follow three mechanisms[169]:

1. Missing completely at random (MCAR), where there are no systematic differences between the available and missing data
2. Missing at random (MAR), where systematic differences between the available and missing data can be explained by other available data
3. Missing not at random (MNAR), where systematic differences between the available and missing data cannot be explained by the available data

The reasons described in Sections 3.6.1 and 3.6.2 indicate that missing FBC test results are missing not at random (MNAR), as the available data cannot explain why some FBC results are known and others are missing. For example, missing FBC results may be due to haematologists actively supressing results due to standard practice or concerns of reliability of the results, which is not information available in the CPRD database.

Multiple imputation and complete-case analysis are common approaches to resolve missing data but are not ideal for this dataset. Multiple imputation combines results from several plausible values generated using appropriate imputation distributions[170]. Software that performs multiple imputation implements the MAR mechanism and therefore relies on other available data to impute missing values. Using multiple imputation can therefore introduce bias when the data are MNAR[171]. Additionally, in a longitudinal setting, multiple imputation is generally preferred where measurements are taken regularly with few time points and may not be practical with large sample sizes and tests performed sporadically over many time points. A complete-case analysis, which restricts the analysis to patients with complete data on all variables, is generally preferred when missing data are MNAR[169]. However, this approach is inefficient due to data loss of known variables.

Maximum likelihood estimation (using expectation-maximisation) is an alternative approach and is incorporated into linear models, such as mixed-effects models. When developing linear models, maximum likelihood estimation uses the full, incomplete dataset to give parameter estimates that most likely reflect the observed data[172]. The estimates are computed separately for patients with complete data on some variables and patients with complete data on all variables and the likelihoods are maximised together to find the likeliest estimate. Analyses using maximum likelihood estimation are often favoured because they adjust the analysis for missing data without performing any imputation or limiting to complete cases[173-176]. Therefore,

primary analyses performed in subsequent chapters will use analytic methods that incorporate maximum likelihood estimation.

4

Trajectories in the FBC towards bowel cancer diagnosis

4.1 Introduction

The ColonFlag algorithm, developed by Medial EarlySign, is a machine-learning algorithm that uses changes in FBC levels over time to give a monotone score of 0-100, where higher levels may indicate the presence of undiagnosed bowel cancer[64] (further details are in Section 1.2.2). Due to the nature of machine-learning, the computational basis for its score outputs is not known in any detail. For example, the changes it relies on was considered an evidence gap (further details in Section 1.2.3). The relevance of changes over time in FBC levels for diagnosis was also highlighted in my systematic review (Chapter 2). The review identified that patients diagnosed have significantly lower red blood cell count, haemoglobin, and mean corpuscular volume and higher red blood cell distribution width, white blood cell count, and platelets within 6 months pre-diagnosis compared to patients not diagnosed. Smaller differences were reported among FBCs measured earlier than six months. This suggests differences in the FBC levels between patients with and without a diagnosis grow larger as diagnosis approaches.

The aim of this study was to identify and compare trajectories (i.e. trends over time) in FBC parameters between patients with and without a bowel cancer diagnosis. To identify potential

opportunities for early detection, I also assessed whether trajectories differ by tumour stage at diagnosis.

4.2 Methods

Study reporting follows the STROBE guidelines[141].

4.2.1 Study design

I performed a retrospective case-control study to explore changes in repeated FBC measures before a diagnosis of bowel cancer compared to patients without a diagnosis. The CPRD dataset derived in Chapter 3 was used, which contains FBC blood tests. The CPRD data was linked to the UK National Cancer Registration and Analysis Service (NCRAS), which provided details of bowel cancer diagnoses for patients in the CPRD. Data access was approved and ethical approval obtained by the CPRD Independent Scientific Advisory Committee (protocol no.: 14_195RMn2A2R). The CPRD and (most recent) NCRAS data cut dates were 28th April 2015 and 14th January 2014, respectively. The clinical codes (medical and entity codes) used to extract FBC results are the same used in Chapter 3 (Section 3.5.1). The clinical codes (ICD-10) to extract diagnoses from the NCRAS are in Table 4.1.

Table 4.1: ICD-10 codes for bowel cancer to identify cases in the NCRAS

ICD-10 code	Term
C18	Malignant neoplasm of colon
C18.0	Caecum (incl.: Ileocaecal valve)
C18.1	Appendix
C18.2	Ascending colon
C18.3	Hepatic flexure

C18.4	Transverse colon
C18.5	Splenic flexure
C18.6	Descending colon
C18.7	Sigmoid colon (Incl.: Sigmoid (flexure); Excl.: rectosigmoid junction (C19))
C18.8	Overlapping lesion of colon
C18.9	Colon, unspecified (Incl.: Large intestine NOS)
C19	Malignant neoplasm of rectosigmoid junction (Incl.: Colon with rectum, Rectosigmoid (colon))
C20	Malignant neoplasm of rectum (Incl.: Rectal ampulla)
C21	Malignant neoplasm of anus and anal canal
C21.0	Anus, unspecified (Excl.: anal: margin (C43.5, C44.5), skin (C43.5, C44.5) perianal skin (C43.5, C44.5))
C21.1	Anal canal (Incl.: Anal sphincter)
C21.2	Cloacogenic zone
C21.8	Overlapping lesion of rectum, anus and anal canal (Incl.: Anorectal junction, Anorectum, and Malignant neoplasm of rectum, anus and anal canal whose point of origin cannot be classified to any one of the categories C20-C21.2)

For each patient, study entry was defined as the latest of date of registration with the practice, patient's 40th birthday, or 1st January 2000. Study exit was defined as the earliest of date of leaving the practice, death, or 14th January 2014 (the NCRAS data cut date).

4.2.2 Participants

Patients were included if they had at least one FBC parameter within 10 years before index date (defined below). Patients were excluded if registered with their primary care practice for less than one year, had a history of bowel cancer before study entry, or diagnosed after study exit. Additionally, patients diagnosed with another cancer type before or simultaneously with bowel

cancer diagnosis were excluded. Patients with an available date of diagnosis but no indication of the cancer type were also excluded.

4.2.3 Clinical outcome

The primary outcome was the first diagnosis of bowel cancer in each patient's study period, taken from the NCRAS dataset, with the corresponding date of diagnosis known for all diagnosed patients included in the study. For cases (patients diagnosed), the index date was defined as the date of the first bowel cancer diagnosis in the patient's study period. For controls (patients without a diagnosis), the index date was a randomly selected date in the patient's study period (between their study entry and exit date). A random date was chosen to mimic the sporadic nature of diagnoses in the overall study period for cases and to eliminate bias introduced by patients who exited the study due to death (see Appendix C12 for further details).

4.2.4 Demographic and FBC variables

Year of birth and sex were available for all patients in the CPRD dataset. The date of each FBC was available for all tests. I included 15 of the 20 parameters that make up the FBC in this study. I excluded five: percentage basophils, eosinophils, lymphocytes, monocytes, and neutrophils because the corresponding counts (also FBC parameters) were available and used. I excluded FBC levels outside biologically plausible ranges, such as negative values (Appendix C1) – these ranges were provided by Michael Jacobs, a senior haematologist in the John Radcliffe Hospital. FBCs performed earlier than 10 years before index date or after index date were excluded.

Each FBC parameter has a normal or reference range of values considered to reflect a healthy individual[177]. Reference ranges differ based on age and sex and may differ slightly across primary care practices, as practices often derive their own ranges using patient data in their records[178-181]. I used reference ranges applicable for patients aged 13 years or above provided

by the Department of Laboratory Haematology at Oxford University Hospitals Trust in this study (Appendix C2)[160].

Iron-deficiency anaemia warrants further investigation for bowel cancer. Anaemia was defined as haemoglobin <13 g/dL for men and <12 g/dL for women, as recommended by the National Institute for Health and Care Excellence (NICE) and World Health Organisation (WHO)[53, 54, 131]. Microcytic anaemia, commonly present in iron-deficiency anaemia[182], was defined as the presence of microcytosis (mean corpuscular volume <80 fL, as recommended by NICE[183]) with anaemia.

4.2.5 Statistical analysis

All analyses were stratified by sex. A two-sided 5% significance level was used for all statistical analyses. Analyses were conducted using Stata/SE 15.1.

I used Locally Weighted Scatterplot Smoothing (LOWESS)[184] to graphically describe the trajectories in each FBC parameter over time for ages 40, 50, 60, 70, 80, and 90 (+/- 3) years at index to assess if trajectories differ by age. Controls with many FBCs in their record are likely to have some other condition/disease, for which they are being monitored, that could affect blood levels, e.g. lowered haemoglobin. These patients would have more weighting on the trajectories, as they contribute many more tests than healthier patients do (see Appendix C3 for red blood cell count and haemoglobin, as examples). Therefore, for both cases and controls, I randomly selected three FBCs per patient (if there were more than three) to reduce the influence of these FBCs with abnormal levels on LOWESS trajectories.

Mixed-effects models for each FBC parameter were developed for males and females separately (using all available FBCs per patient). To model differences in FBC levels between cases and

controls over time, bowel cancer status (yes/no) and time to index date (years) were included as fixed effects together with an interaction between them. Each model also included age at index date (years) as a fixed effect. For red blood cell count, haemoglobin, haematocrit, and neutrophil count, an interaction between age and time was included as a fixed effect, as trajectories differed by age group upon graphical inspection of LOWESS plots (Appendix C4). Additionally, for all models except for mean platelet volume, basophil count, lymphocyte count, and monocyte count, an interaction between age and bowel cancer status was included as a fixed effect, as trends over ages differed between cases and controls upon graphical inspection of LOWESS plots (Appendix C5). In every model, each patient was modelled using a random intercept and time using a random slope with an unstructured covariance matrix to account for correlation in repeated measures.

Non-linearity of continuous variables was first assessed for time to index and, when chosen, subsequently for age at index. Non-linearity was based on visual inspection of LOWESS curves and Akaike and Bayesian information criteria, which compared piecewise linear splines, restricted cubic splines, and fractional polynomials, and varying number of knots and knot locations (Appendix C4 and Appendix C5)[125, 185, 186]. For both time and age, linear splines gave better fitting models for almost all parameters for both males and females. For consistency, linear splines were therefore used. Where non-linear (on visual inspection), time to index date was modelled using linear splines with three knots: at 1, 2, and 4 years before index date. Where non-linear, age at index date was modelled using linear splines with knots at ages 60, 70, and 80 years for red blood cell-related parameters and platelet count and were variable for white blood cell count-related parameters.

To identify opportunities for early detection, I assessed differences in trajectories between cases diagnosed at Dukes' tumour Stage A (earliest stage) and D (latest stage). Mixed-effects models

were developed using the same methods described above but were limited to cases alone and included Dukes' tumour stage (A versus D) as a fixed effect instead of bowel cancer status. To explore the association between microcytic anaemia and diagnosis, I calculated the proportion of patients with microcytic anaemia per 6-month (+/- 3 months) time band, up to five years before index date. Microcytic anaemia presence was based on any FBC in the time band, if a patient had multiple, and proportions were calculated out of the number of patients in the time band. Additionally, I derived age-adjusted odds ratios (95% CI) for microcytic anaemia presence using logistic regression for each time band separately.

4.2.6 Missing data

Age and sex were available for all patients. The methods for handling missing FBC data are described in detail in Section 3.7. In summary, the mixed-effects models use maximum likelihood estimation to derive parameter estimates, which adjusts for missing data.

4.2.7 Sensitivity analysis

I performed three sensitivity analyses to assess how the trajectories were influenced by the study design.

1. I used a cohort study design instead of a case-control design. Patient eligibility criteria were the same as in the case-control design, except patients without FBCs before the randomly selected index date (controls only), without any FBC within 10 years prior to index, or diagnosed after study exit were excluded from the case-control study but included in the cohort study. Index date for the cohort design was the date of the first FBC in the patient's record. Patients were followed for 10 years from index date until bowel cancer diagnosis. Patients not diagnosed in this 10-year period were censored at the earliest of date of leaving the practice, death, 14th January 2014, or 10 years after

index. LOWESS smoothing was used to examine the trajectories graphically, as described in Section 4.2.5.

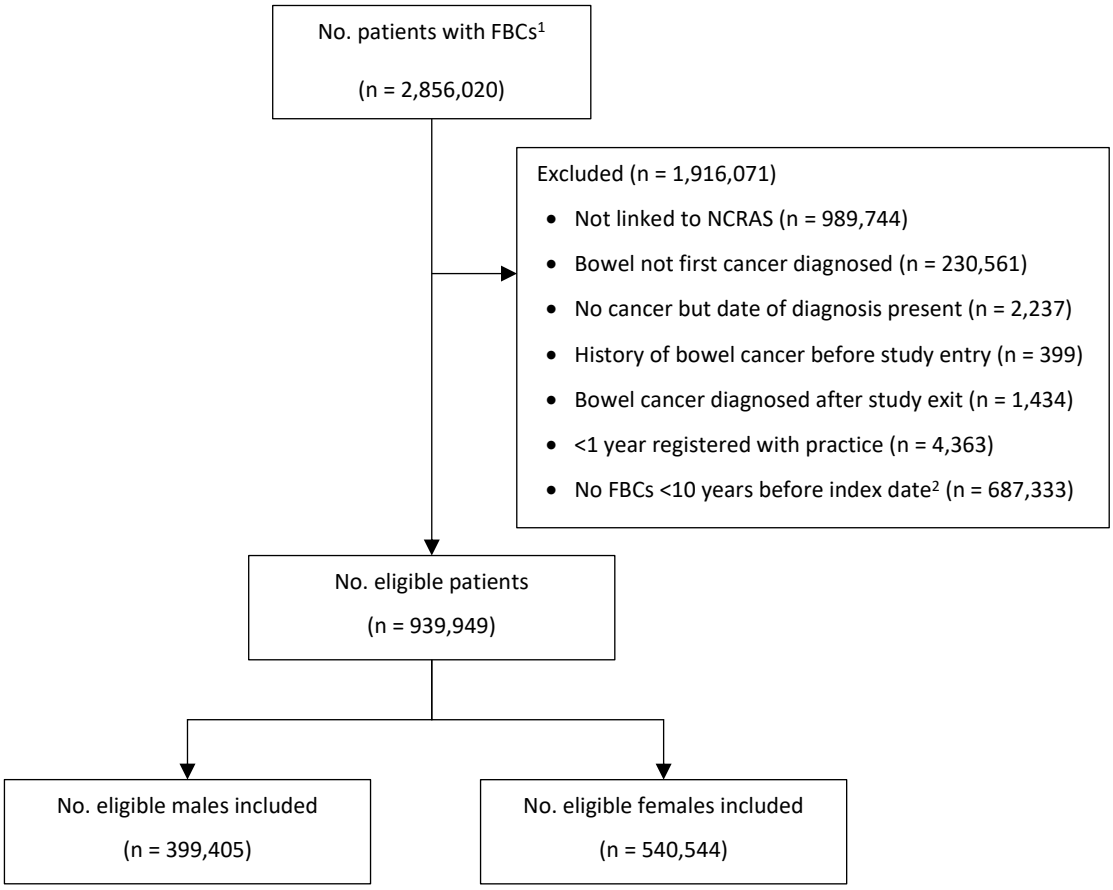
2. Index date was defined as study exit instead of a randomly selected date in the control group. All other details were kept the same. LOWESS smoothing was used to examine the trajectories graphically, as described in Section 4.2.5.

3. Matching was used, with cases and controls matched 1:5 on age at index and follow-up time. Follow-up was time (years) from first FBC to index date, converted into 6-monthly bands for matching. A random index date within the control's study period was used instead of the index date of their matched case because the latter heavily reduced the sample size. For example, many controls had an index date after study exit or had no FBCs before index. LOWESS smoothing was used to examine the trajectories graphically, as described in Section 4.2.5.

4.3 Results

There were 939,949 patients with at least one FBC within 10 years before index date who satisfied the eligibility criteria and were included in the study (Figure 4.1). There were 399,405 males with 1,193,619 FBC tests among them and 540,544 females with 1,874,609 FBC tests among them. Cases were on average older than controls at index date (see Table 4.2 for a description of the patient sample).

Figure 4.1: Patient flow diagram (case-control analysis to identify trajectories)



[1]Number of patients available in the CPRD data extract post-data quality check (Chapter 3).
[2]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
Abbreviations: FBC=full blood count, NCRAS=National Cancer Registration and Analysis Service.

Table 4.2: Characteristics of the patient sample

	Bowel cancer			**No bowel cancer**		
	No.	**Mean age[1] (SD)**	**Age[1] range**	**No.**	**Mean age[1] (SD)**	**Age[1] range**
Male	9,255	70.8 (10.6)	40 - 101	390,150	59.5 (12.9)	40 - 111
Female	8,153	73.0 (11.8)	40 - 101	532,391	60.6 (14.7)	40 - 109
Total	17,408	71.8 (11.2)	40 - 101	922,541	60.1 (14.0)	40 - 111

[1]Age at index date: date of diagnosis (cases) or a randomly selected date in the patient's study period (controls).

4.3.1 The FBC test

On average, 14 of the 15 parameters were available per FBC for both males and females. Red blood cell distribution width was missing for almost 100% of FBCs. This is likely because this parameter was historically not reported to the general practice by haematology laboratories, despite being automatically derived by haematology analysers (i.e. machines) – see Chapter 3 for further details. Consequently, red blood cell distribution width was excluded from further analyses. Haemoglobin had the least amount of missing data, missing for 1.9% (n=22,637) and 1.8% (n=34,678) of FBC tests for males and females, respectively. Missing data for each parameter are reported in Table 4.3.

Table 4.3: Summary of missing data for full blood count parameters

Full blood count parameter	**Plausible range[1]**	**Number (%) missing**	
		Males (n=1,193,619 FBCs)	**Females (n=1,874,609 FBCs)**
Red blood cell count (10^{12}/L)	2.5 – 7.5	103,607 (8.7%)	165,608 (8.8%)
Haemoglobin (g/dL)	0.3 – 21.0	22,637 (1.9%)	34,678 (1.8%)
Haematocrit/packed cell volume (L/L)	0.1 – 0.6	968,246 (81.1%)	1,514,243 (80.8%)
Mean corpuscular volume (fL)	53 – 125	57,807 (4.8%)	86,606 (4.6%)
Mean corpuscular haemoglobin (pg)	22 – 36	92,460 (7.7%)	150,566 (8.0%)
Mean corpuscular haemoglobin concentration (g/dL)	27 – 37	165,983 (13.9%)	262,094 (14.0%)
Red cell distribution width (%)	9 – 15	1,193,587 (100%)	1,874,530 (100%)
Platelet count (10^9/L)	0 – 1500	71,385 (6.0%)	114,575 (6.1%)
Mean platelet volume (fL)	5 – 14	1,040,399 (87.2%)	1,630,832 (87.0%)
White blood cell count (10^9/L)	0 – 850	223,059 (18.7%)	351,858 (18.8%)
Basophil count (10^9/L)	0 – 1.5	326,914 (27.4%)	535,356 (28.6%)
Eosinophil count (10^9/L)	0 – 8	322,907 (27.1%)	529,122 (28.2%)
Lymphocyte count (10^9/L)	0 – 850	289,640 (24.3%)	457,255 (24.4%)
Monocyte count (10^9/L)	0 – 40	298,940 (25.0%)	473,324 (25.2%)

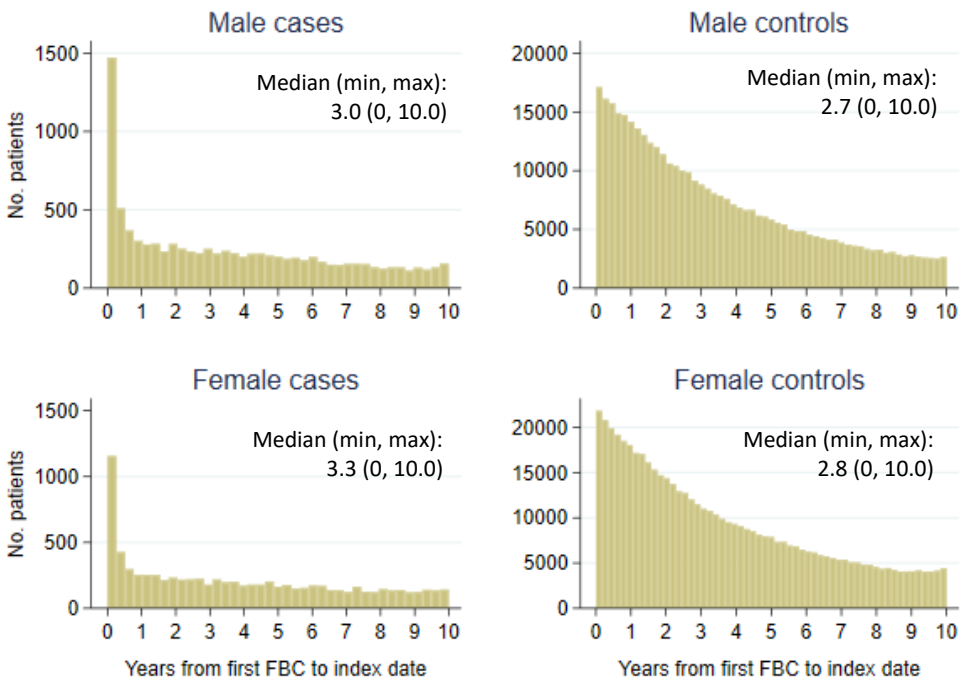
Neutrophil count (10^9/L)	0 – 850	285,027 (23.9%)	451,291 (24.1%)

[1]Biologically plausible ranges for each FBC parameter were provided by Michael Jacobs, a senior haematologist based in the John Radcliffe Hospital, Oxford.

4.3.2 Bowel cancer and FBC trajectories

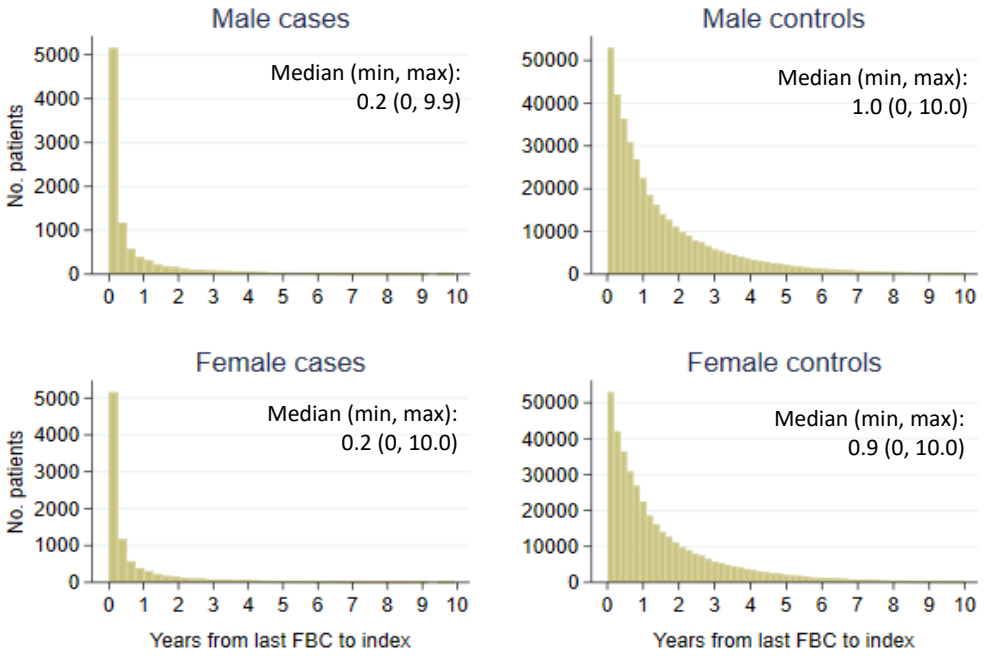
Of 399,405 males and 540,544 females, 2.3% (n=9,255) and 1.5% (n=8,153) had a bowel cancer diagnosis, respectively. Median (min, max) follow-up (years) from the earliest FBC in the 10-year period to index date was comparable between cases and controls: 3.0 (0, 10.0) for male cases vs. 2.7 (0, 10.0) for male controls and 3.3 (0, 10.0) for female cases vs. 2.8 (0, 10.0) for female controls (Figure 4.2). Median (min, max) time (years) between the most recent FBC and index date was 0.2 (0, 9.9) years for male cases, 1.0 (0, 10.0) years for male controls, 0.2 (0, 10.0) years for female cases and 0.9 (0, 10.0) years for female controls (Figure 4.3). Although both cases and controls show a positively skewed distribution in both figures, the spikes at zero reflect many cases with FBCs close to diagnosis. This is expected because FBCs are more commonly performed in diagnosed patients and close to diagnosis. This will not influence the comparison of trajectories between groups, as this analysis is exploratory by nature, but will need to be considered for the prediction model developed in Chapter 5.

Figure 4.2: Histogram of follow-up time from earliest FBC in the 10-year period to index date[1]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.

Figure 4.3: Histogram of time between latest FBC in the 10-year period and index date[1]

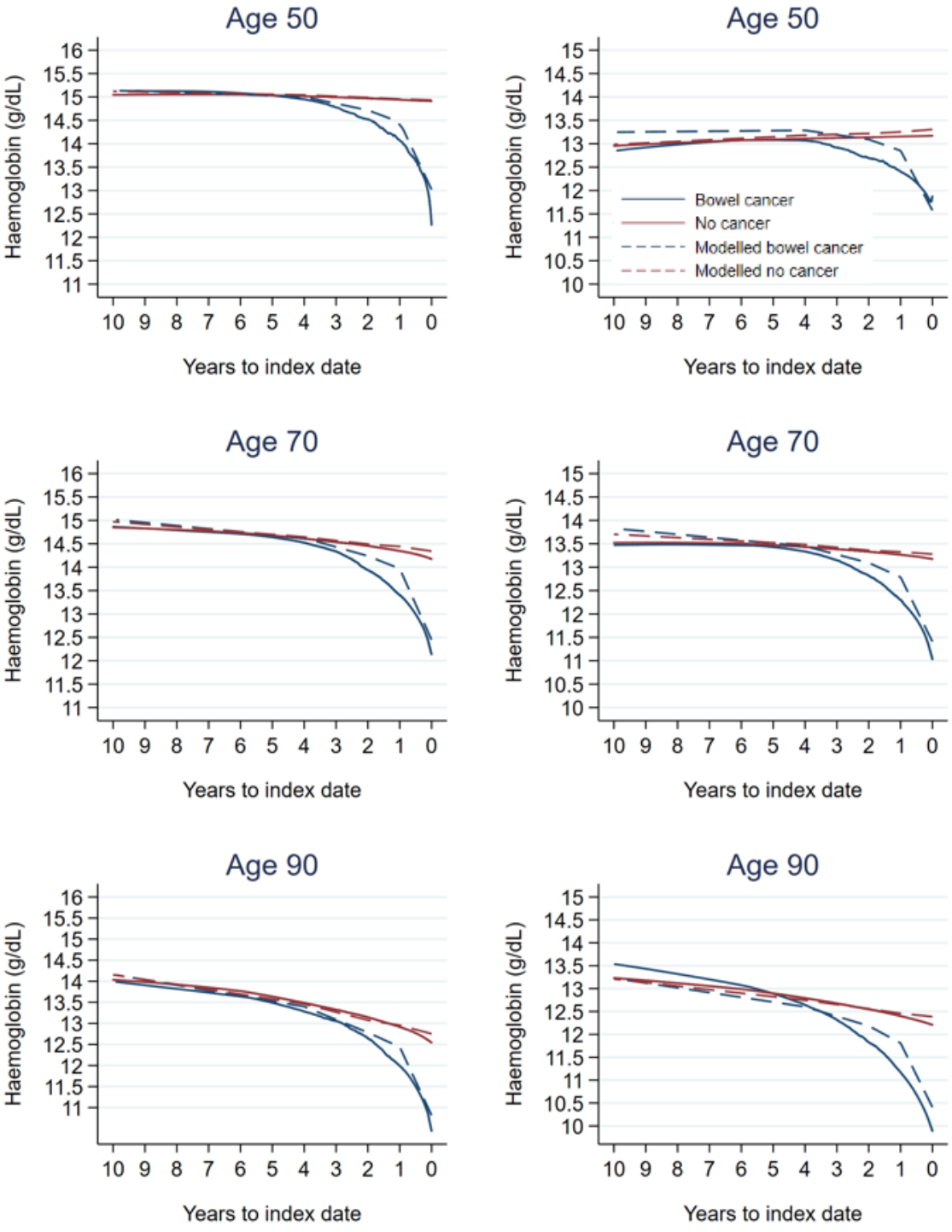


[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.

Mixed-effects models for each parameter are in Appendix C6. The presence of bowel cancer was statistically significantly associated with all parameter levels ($p<0.05$ for each model), except white blood cell count and eosinophil count for both males and females. LOWESS and modelled (fixed-effects) trajectories in haemoglobin and platelets are provided in Figure 4.4 and Figure 4.5, respectively, and for all FBC parameters in Appendix C7. There was no apparent difference in trends measured from 10 to 4 years before index date between cases and controls for both males and females. Within 4 years before index date, levels changed steadily over time in patients without a diagnosis, such as the reduction in haemoglobin over time that may be due to increasing age. However, cases had trends in FBC levels that diverged from controls for all parameters, with statistically significant interaction terms between the presence of bowel cancer and time to index date (including splines terms) ($p<0.05$ for each FBC parameter model), except basophil count in males (coefficient = -0.0002 (95% CI = -0.0005-0.0001)). Additionally for cases, the rate of change in many FBC parameters increased as the time to diagnosis approached.

Individualised trends (fitted values from mixed-effects models), which consider population-level trajectories (fixed-effects) and patient-specific contributions (random-effects), are given for four cases and four controls in Appendix C8.

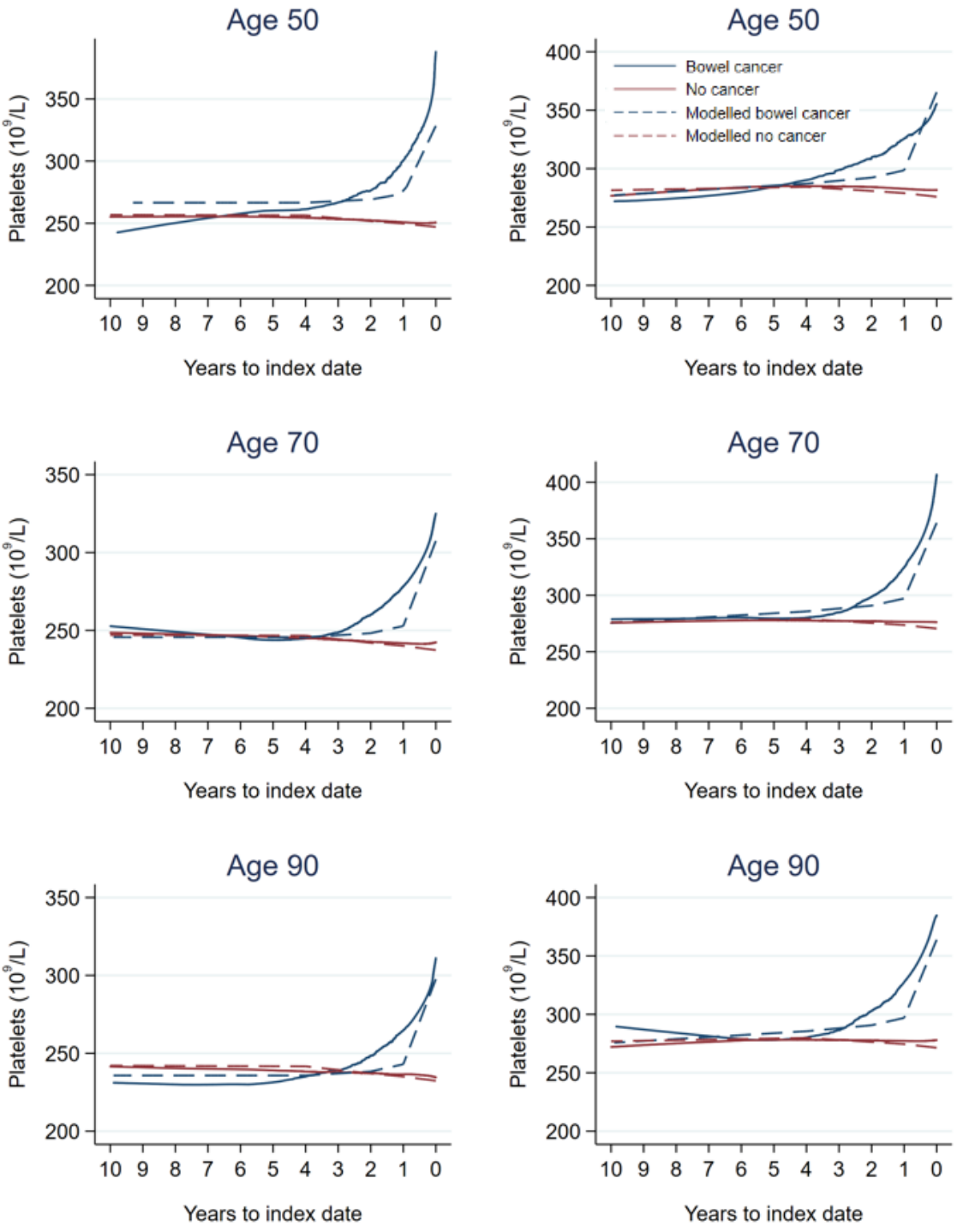
Figure 4.4: Trajectories (LOWESS and fixed-effects) in haemoglobin levels (g/dL) for males (left) and females (right) with and without a diagnosis of bowel by age at index date[1, 2]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.

[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure 4.5: Trajectories (LOWESS and fixed-effects) in platelet count (10^9/L) for males (left) and females (right) with and without a diagnosis of bowel by age at index date[1, 2]

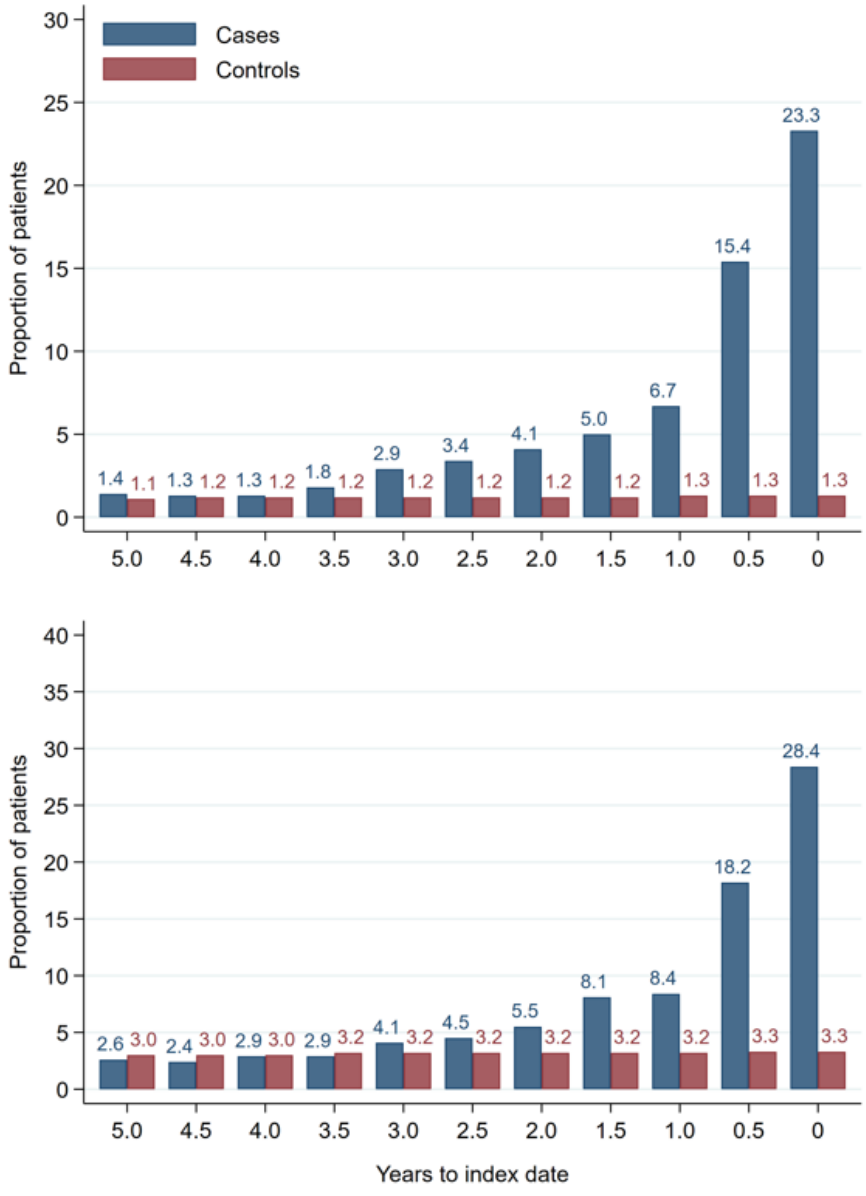


[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

4.3.3 Bowel cancer and microcytic anaemia

Anaemia was present in 48.8% of male cases and 54.3% of female cases on any FBC within one year of diagnosis. At each 6-monthly interval up to five years before index date, the proportion with microcytic anaemia was higher in cases than controls (Figure 4.6). In cases, the proportion increased as the time to diagnosis approached and was highest at 0-3 months prior to diagnosis: 23.3% (n=1,188) of 5,107 males and 28.4% (n=1,286) of 4,521 females with FBCs in that period.

Figure 4.6: Microcytic anaemia prevalence over 6-month time bands before index date between patients with and without bowel cancer – males (top) and females (bottom)



Presence of microcytic anaemia statistically significantly increased odds of diagnosis at each time band within 3 years before index date for both males (3-year OR=2.2 (95% CI=1.5-3.1)) and females (3-year OR=1.7 (95% CI=1.3-2.3)) (Table 4.4). Odds of diagnosis increased as time to index date approached. No odds ratio achieved statistical significance at earlier time points.

Table 4.4: Age-adjusted odds ratios for diagnosis corresponding to presence of microcytic anaemia at 6-monthly (+/- 3 months) time intervals prior to index date[1]

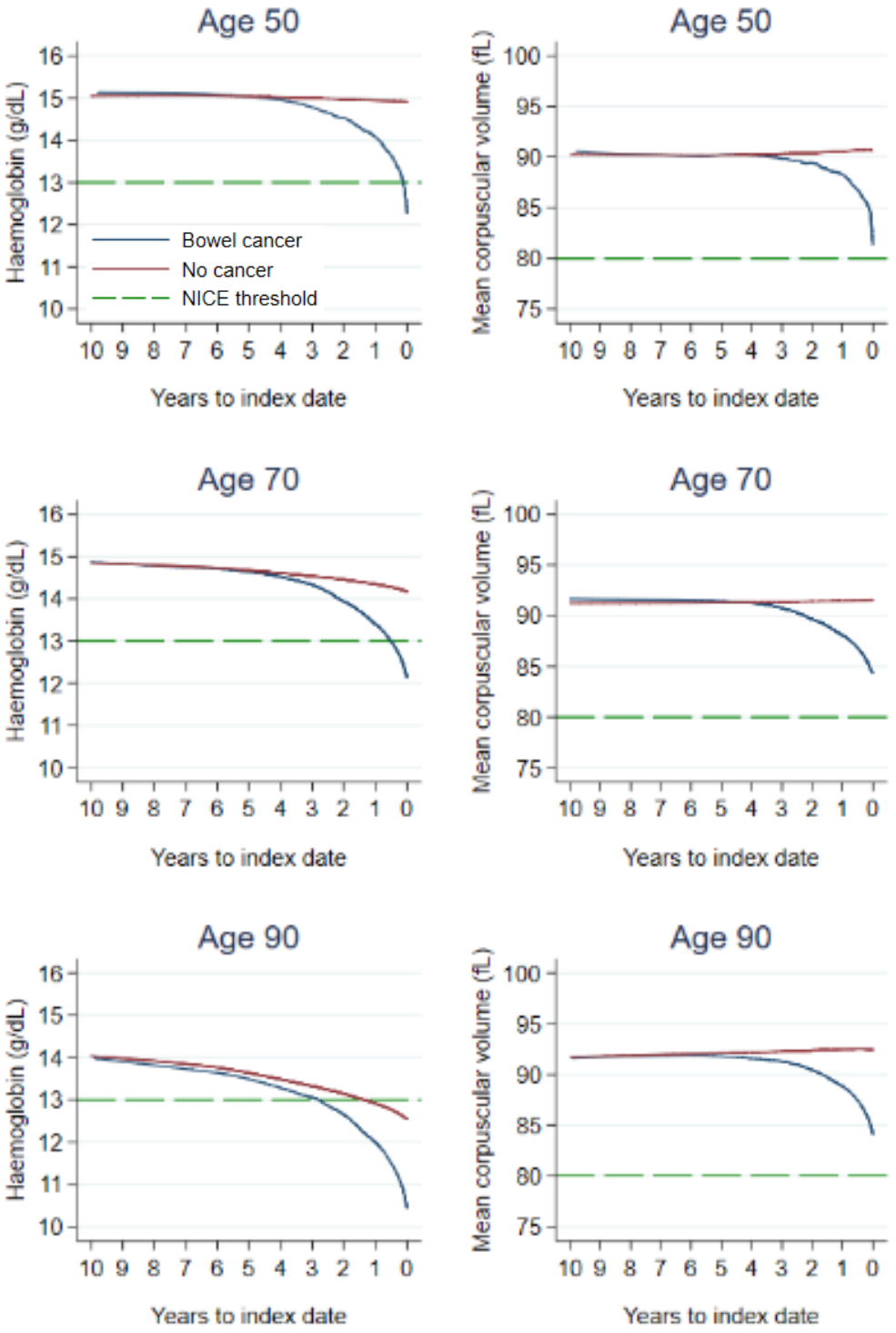
Time interval (years)	No. cases		No. controls		Odds ratio (95% CI) (reference: no microcytic anaemia)
	Microcytic anaemia[2]	Total	Microcytic anaemia[2]	Total	
Males:					
0	1188	5107	937	69648	19.6 (95% CI=17.7-21.7)
0.5	529	3424	1456	115123	10.4 (95% CI=9.1-11.8)
1	171	2560	1301	103281	4.2 (95% CI=3.4-5.1)
1.5	115	2280	1058	91084	3.5 (95% CI=2.7-4.4)
2	83	2025	930	79130	2.4 (95% CI=1.8-3.3)
2.5	63	1837	864	70678	2.3 (95% CI=1.7-3.1)
3	49	1662	757	61840	2.2 (95% CI=1.5-3.1)
3.5	29	1593	648	54331	0.9 (95% CI=0.6-1.5)
4	18	1380	549	47470	0.9 (95% CI=0.5-1.6)
4.5	15	1186	505	41655	1.0 (95% CI=0.6-1.7)
5	15	1063	419	36578	1.0 (95% CI=0.5-1.8)
Females:					
0	1286	4521	3473	105359	12.2 (95% CI=11.3-13.3)
0.5	588	3237	5684	171034	6.6 (95% CI=5.9-7.4)
1	208	2483	4948	152754	2.9 (95% CI=2.4-3.4)

1.5	178	2194	4360	136087	3.0 (95% CI=2.5-3.6)
2	108	1976	3822	119565	2.0 (95% CI=1.6-2.6)
2.5	82	1814	3376	106114	1.5 (95% CI=1.1-2.0)
3	66	1600	2963	93266	1.7 (95% CI=1.3-2.3)
3.5	44	1492	2597	82236	1.1 (95% CI=0.7-1.5)
4	40	1359	2233	73537	1.3 (95% CI=0.9-1.9)
4.5	31	1266	1920	64388	0.8 (95% CI=0.5-1.3)
5	29	1097	1722	56759	0.9 (95% CI=0.5-1.4)

[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]NICE threshold: 13 g/dL (males) or 12 g/dL (females) haemoglobin and 80 fL mean corpuscular volume.

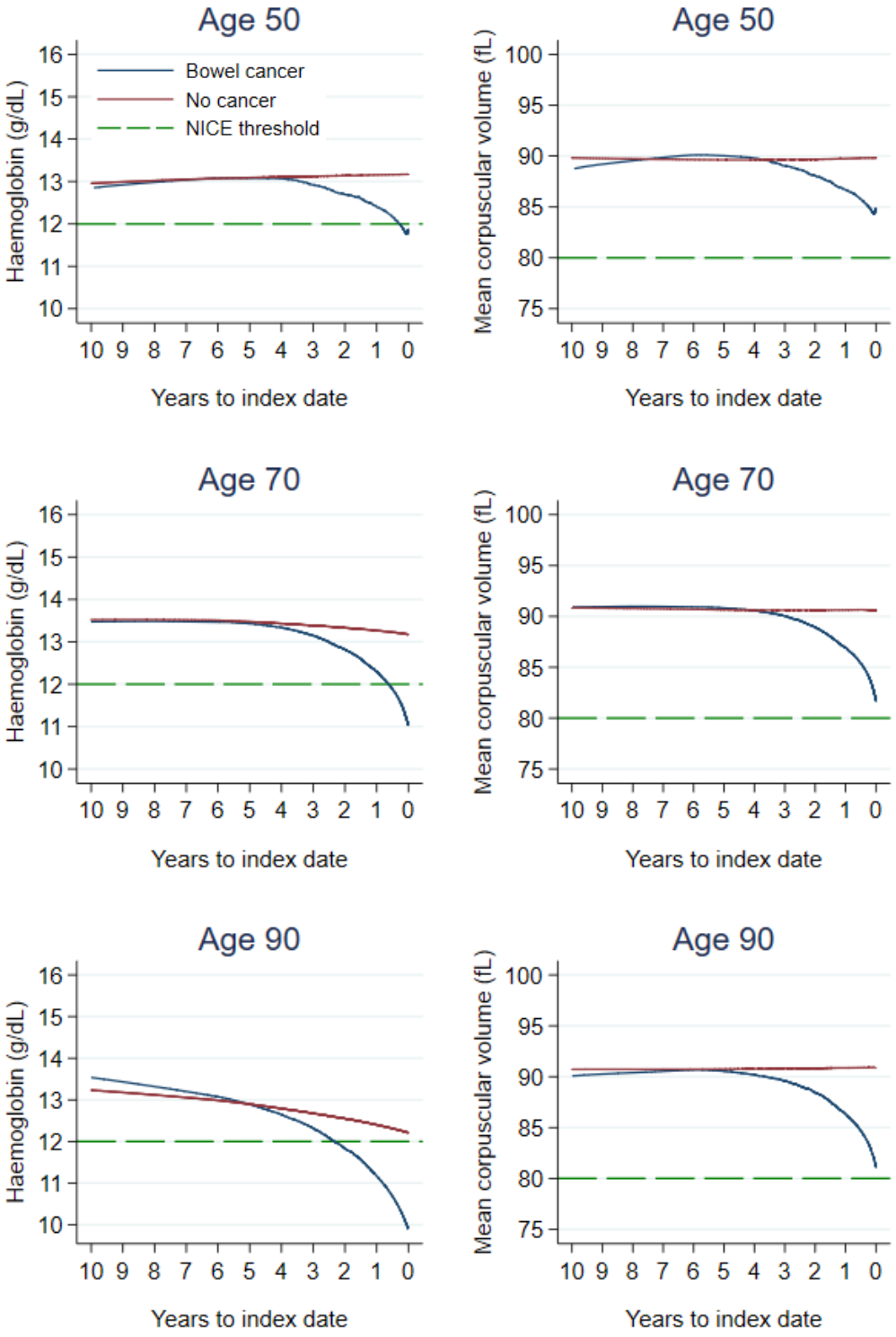
I compared the NICE thresholds for microcytic anaemia with haemoglobin and mean corpuscular volume trajectories (Figure 4.7 for males and Figure 4.8 for females). Haemoglobin levels often only reached the threshold for anaemia within 6-12 months of diagnosis in younger-moderate aged cases but reached the threshold years earlier in older cases. However, in these older age groups, even controls were often anaemic. The mean corpuscular volume threshold was not reached for any age group, suggesting a minority of patients show microcytosis on the FBC.

Figure 4.7: Trajectories (LOWESS) in haemoglobin (left) and mean corpuscular volume (right) by microcytic anaemia threshold for males, by bowel cancer status and age at index[1, 2, 3]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index.
[3]NICE threshold: 13 g/dL haemoglobin and 80 fL mean corpuscular volume.

Figure 4.8: Trajectories (LOWESS) in haemoglobin (left) and mean corpuscular volume by microcytic anaemia threshold for females, by bowel cancer status and age at index[1, 2, 3]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index.
[3]NICE threshold: 12 g/dL haemoglobin and 80 fL mean corpuscular volume.

4.3.4 Bowel cancer and FBC reference ranges

For all FBC parameters, the trajectories showed that levels remained in the reference range for both cases and controls, except red blood cell count, haemoglobin, haematocrit, mean corpuscular volume, and mean platelet volume (Appendix C2). In these five parameters, the trajectories in younger cases suggest blood levels often only go outside the reference range within approximately six months of diagnosis. However, in older cases, levels remain outside the normal range for approximately three years before diagnosis, which was also observed for older controls.

4.3.5 Bowel cancer staging and FBC trajectories

The number of cases diagnosed per Dukes' tumour stage is in Table 4.5. There were 39% and 41% of male and female cases, respectively, with Dukes' stage at diagnosis missing.

Table 4.5: Number of cases per Dukes' tumour stage at diagnosis

Dukes' stage	Males		Females	
	Number	**% of cases**	**Number**	**% of cases**
A	850	9.2%	688	8.4%
B	2,049	22.1%	1,778	21.8%
C	2,038	22.0%	1,797	22.0%
D	737	7.8%	532	6.5%
Unknown	3,581	38.7%	3,358	41.2%
Total	9,255	100%	8,153	100%

Trajectories between Stage A and D tumours are in Appendix C9. There was no apparent difference in trajectories between Stage A and Stage D tumours among older patients. However,

in younger patients, changes in blood levels started up to one year earlier in patients with Stage D than Stage A in all FBC parameters, except mean platelet volume, basophil count, eosinophil count, and lymphocyte count for both males and females. These four parameters showed no apparent difference between tumour stages.

Mixed models including Dukes' stage at diagnosis, developed in cases alone, are provided in Appendix C10. Dukes' D vs. A was statistically significantly associated with red blood cell count and monocyte count in males and platelets, white blood cell count, eosinophil count, monocyte count, and neutrophil count in females. However, many FBC parameters showed different trajectories upon visual inspection so lack of statistical significance may be due to a small sample size, as only a small proportion of patients had known staging.

4.3.6 Sensitivity analyses

Three sensitivity analyses were performed to assess how study design characteristics influence trajectories. The sensitivity analyses showed different trajectories in FBC levels and highlight the appropriateness of a case-control design and choice of a random index date in controls, as done in the primary analysis. For example, a cohort study did not provide a standardised time point to assess trajectories from in cases, as diagnoses were occurring sporadically over the 10-year follow-up period. Additionally, censoring controls at study exit in a case-control design meant trajectories were influenced by the confounding effect of death (a reason for study exit). Furthermore, a matched design did not influence the trajectories, as these were already stratified by sex and age, which were adjusted for in the mixed-effects models in the primary analysis. Full details are reported in Appendix C11 (sensitivity analysis 1 – cohort study), Appendix C12 (sensitivity analysis 2 – study exit as index date in controls), and Appendix C13 (sensitivity analysis 3 – matched design).

4.4 Discussion

4.4.1 Summary of main findings

I assessed whether FBC parameters differed between patients with and without a bowel cancer diagnosis and whether these differences changed over time. I identified trajectories (adjusted for age and sex) that highlighted a difference between cases and controls in many individual FBC parameters within approximately four years before diagnosis. In this four-year period, levels in cases began to diverge from those measured in controls and these differences grew larger in the run up to diagnosis, with levels in patients without a diagnosis changing less rapidly over time.

4.4.2 Comparison with existing literature

Two prior studies assessed changes over time in haemoglobin between patients with and without bowel cancer: one in an Israeli population[95] and one in a combined Swedish and Danish population[187]. Both studies reported that haemoglobin levels in diagnosed patients diverge from levels in patients without a diagnosis, one at four years (Israeli study) and the other at two years (Swedish and Danish study) before diagnosis, with the rate of decline increasing in the run up to diagnosis. I identified similar findings in a UK population, with haemoglobin levels that diverge around four years before diagnosis and a greater decline in the run up to diagnosis. I also identified relevant trajectories for many other FBC parameters. Another study used machine-learning to develop an algorithm called the ColonFlag, which assesses change over time in various FBC parameters (at 18 and 36 months before index FBC) from a single patient to derive a monotone score of 0-100 for diagnosis in an Israeli population[64, 188]. It is unclear what trends over time are considered related to bowel cancer in the machine-learning algorithm. A fourth study used logistic regression to test whether the difference in five parameters between the two most recent FBCs taken from the same patient was associated with diagnosis in UK primary care patient data[57]. With the two tests performed at any time in a mean follow-up period of 6.3

years, they report no statistically significant association in change in red blood cell count (p=0.13), white blood cell count (p=0.06), or haematocrit (p=0.23) but do for change in mean corpuscular volume (p=0.04) and mean corpuscular haemoglobin (p=0.02). However, my analyses suggest that red blood cell count, white blood cell count, and haematocrit do relevantly change over time due to bowel cancer.

A recent study of haemoglobin levels in newly diagnosed bowel cancer patients (mean age approximately 70 years) in Finland reported lower haemoglobin levels at diagnosis of higher stage tumours (13.2 g/dL at Stage 1 vs. 12.2 g/dL at Stage 4, on average)[189]. These levels are similar to those I identified in this study (Figure C62), where the trends in most FBC parameters are similar between Stage A and D tumours, but the divergence from controls started up to one year earlier in Stage D diagnoses than Stage A.

When compared to NICE and WHO guidelines for anaemia, which could be due to any reason including iron-deficiency, my trajectories indicated that haemoglobin levels often only reached the threshold for males (<13 g/dL) and females (<12 g/dL)[53, 54, 131] at approximately 6-12 months before diagnosis. Up to one year before diagnosis, anaemia was present in 48.8% of male cases and 54.3% of female cases. These results are similar to those reported in a previous UK primary care study of anaemia within one year prior to bowel cancer diagnosis[52]. Microcytic anaemia, commonly caused by iron-deficiency, which may warrant further investigation for bowel cancer, was present in 23% of male cases and 28% of female cases within a year prior to diagnosis. My analyses suggest there are relevant changes that occur up to three years before the presence of anaemia, including iron-deficiency anaemia. Additionally, the mean corpuscular volume threshold for microcytosis was not reached for any age group, suggesting a minority of patients show microcytosis on the FBC. Using changes over time could be more helpful to facilitate early detection than relying on low FBC parameter levels.

I also compared the FBC trajectories to normal reference ranges[160]. Abnormal FBC parameter levels are considered to represent health-related conditions or disease. Although differences in most FBC parameters between cases and controls grew larger as the time to diagnosis approached, they remained small overall and often confined to the normal reference range, except for red blood cell count, haemoglobin, haematocrit, mean corpuscular volume, and mean platelet volume. These parameters often only became abnormal close to diagnosis in younger cases but for many years before in older cases, which was also observed in older controls. Therefore, such differences in trends between cases and controls may not be obvious to a clinician in general practice, as these differences would be considered to represent little-to-no concern, and the opportunity to utilise these changes over time to identify bowel cancer missed. Our study supports the conclusions of another recent report that highlighted how the normal range does not necessarily reflect a healthy individual[190].

4.4.3 Strengths & limitations

Strengths:

This study uses a large patient sample, increasing the precision of estimates and reliability of trajectories by sex and for different age groups. This study is the first to identify relevant trajectories in many FBC parameters that could facilitate earlier detection of bowel cancer and highlights how the relevant differences occur within the reference range, meaning they are likely being missed in practice. Additionally, this study highlights that using repeated FBC measures may be more useful than relying on a single FBC, which is what happens in current general practice. Another strength is that this study fills two grey areas for future research identified in the systematic review: researching changes over time in FBC levels and assessing whether the FBC can assist earlier detection by performing analyses by tumour stage at diagnosis.

Limitations:

FBC blood tests are ordered for many reasons in primary care, not bowel cancer specifically, but these reasons are not available in CPRD. Patients with FBCs are suspected to be generally less well than patients who are not tested (except in antenatal screening, which includes a FBC), so controls included in this study may not be entirely healthy. However, many existing studies identified in my systematic review (Chapter 2) showed that in patients with FBCs, the test has the potential to distinguish patients with and without bowel cancer.

Red blood cell distribution width is a parameter used to diagnose medical conditions, especially bowel cancer[191]. This parameter was not reported to primary care practices until recently years, hence why almost all tests in the CPRD data have this parameter missing. Consequently, I excluded this relevant parameter from all analyses.

Due the nature of the case-control study design, I modelled 10 years of repeated measures data before index for each age group separately. Therefore, the trends in older age groups for some FBC parameters were subject to survivor bias. For example, in Figure C30 (Appendix C4), controls who survived to 90 years at index were generally healthier 10 years earlier than controls who survived to 80 years at index, who may or may not survive to age 90. Thus, older controls would often have FBC levels that reflected healthier individuals (e.g. higher haemoglobin) 10 years prior to index on average than patients censored at that younger age. This survivor bias could not be adjusted for in my analyses.

The data was subject to sampling bias, where patients, including controls, with a larger number of FBCs are likely to be more ill than patients who have few FBCs. The former group are therefore likely to have many FBCs with abnormal blood levels and have more weighting on the trajectories, as they contribute many more tests than healthier patients do. To eliminate this

bias and influence of these many abnormal FBC levels on LOWESS trajectories, I used a random selection of three FBCs per patient (if there were more than three). Consequently, I could not use all FBCs to identify trajectories in the raw data. Nonetheless, a large number of FBCs remained so the trajectories can be considered precise.

Other known factors for bowel cancer, such as ethnicity and family history, were not adjusted for in my mixed-effects models, as they were not available in the CPRD data extract. These factors may be considered in future work.

Tumour stage at diagnosis was missing for approximately one-third of bowel cancer diagnoses. Therefore, many cases were excluded from analyses of tumour staging, reducing sample size and precision of estimates.

4.4.4 Conclusion and further research

This research highlighted some barriers in current clinical practice that makes early detection challenging. The differences in FBC trajectories between cases and controls identified in this study would often go unnoticed in clinical practice. One reason is that it is difficult and impractical for busy clinicians and nurses filing results to notice minor changes in parameter values over time, as this involves simultaneously analysing multiple parameters from current and previous FBC reports and many FBC blood tests are performed on a daily basis. However, using changes over time may be more relevant (and predictive) than relying on the one FBC test, with the latter being what currently happens in practice, where the opportunity to identify cancer this way is missed. This is because relevant changes over time can occur for years before FBC levels reach abnormal thresholds or reach referral thresholds in national guidelines used to identify who needs further testing. A tool that utilises these changes over time in many parameters simultaneously to identify high-risk patients would be of considerable benefit.

Therefore, I have developed a dynamic statistical prediction model that uses trajectories in the FBC to derive an individual's risk of diagnosis in the future (Chapter 5). The prediction model will aim to support the current bowel cancer UK screening programme by identifying possible cases for investigation[15]. It will utilise changes across multiple FBC parameters over time to detect bowel cancer at earlier time points than is possible using single parameter thresholds. Additional future work (after the DPhil) includes deriving more appropriate, age-specific FBC reference ranges using large-scale patient data.

5

A joint model using FBC tests for bowel cancer detection

5.1 Introduction

The research described in this chapter is built on the research described in previous chapters. The systematic review (Chapter 2) identified six FBC parameters that are associated with bowel cancer diagnosis: red blood cells, haemoglobin, mean corpuscular volume, red blood cell distribution width, white blood cell count, and platelets. Differences in these FBC parameters between patients with and without a diagnosis were either statistically significant in meta-analysis or consistently statistically significantly different in each study that analysed it. These parameters are therefore considered the most relevant to identify risk of diagnosis. When critically appraising existing prediction models, the review identified that poor methods were often employed, models were often inadequately reported, many models scored a high risk of bias (in the analysis domain of the PROBAST tool), and likely do not work well as reported. The latter includes the QCancer Colorectal model, which relies on symptoms to predict two-year risk of diagnosis[69], with recent evidence indicating symptoms are often first reported to primary care within six months prior to diagnosis[13].

The data quality assessment (Chapter 3) led to a more reliable dataset for analysis, having identified issues in clinical coding, units of measurement, and extent of missing data. For

example, red blood cell distribution width is nearly always missing in the CPRD database because the blood level was supressed in the FBC report by haematologists. However, this is a relevant parameter, as identified in the systematic review. Additionally, many FBC parameters are mathematically related, suggesting not all individual levels may be needed in future work (mathematical relationships are described in Appendix B2, discussed in detail in Chapter 3).

In Chapter 4, I identified different trajectories (i.e. rate of change over time) in many FBC parameters between patients with and without bowel cancer. The trajectories could improve on current practices, which relies on results from a single FBC, the current value, to make health-related decisions. These trajectories also differed between patients diagnosed with early- vs. late-stage tumours, suggesting the potential for earlier detection.

The aim of this study was to develop and internally validate a new prognostic prediction model that, unlike existing models, uses trajectories in FBC parameters to identify future risk of bowel cancer, aided by findings from previous chapters. This model would follow the same approach as the ColonFlag Israeli algorithm[64], using changes in FBC levels over time, but would use more transparent methodology. Therefore, I also compare performance of the new prediction model to the ColonFlag algorithm. The new prediction model would also improve on existing UK models, such as QCancer Colorectal, which relies on assessments at one time point, because it would use information available before symptoms are commonly reported to primary care.

5.2 Joint modelling of longitudinal and time-to-event data

5.2.1 Cox proportional hazards model:

Originally proposed by Cox 1972[192], the Cox model is commonly used for clinical risk prediction when modelling time to the event of interest. A Cox model involves measuring a set of covariates

at a single, baseline time-point and using these covariates to predict an outcome at a specified time-point in the future. For subject $i = 1, \dots, n$, the Cox model is denoted by:

$$\boldsymbol{S_i(t) = S_0(t)^{\exp(\sum_{k=1}^{K} \beta_k x_{ik})}} \quad \text{Eq. 1}$$

Where

1. $S_i(t)$ is the survival probability (i.e. risk of not having the event), at a specified time, t, after baseline (when covariates are measured).
2. $S_0(t)$ is the baseline survival estimate at time, t, and is not estimated in a Cox regression model but is required to derive risk predictions. Various methods can be used to estimate the baseline survival, such as the Breslow estimator and the cumulative hazard rate[125, 193-196].
3. $\sum_{k=1}^{l} \beta_k x_{ik}$ is referred to as the linear predictor, which is a linear combination of $k = 1, \dots, K$ x-covariates (measured at baseline) with corresponding coefficients β_k.

The Cox model was extended by Therneau & Grambsch 2000[197], such as to allow for time-varying covariates, which are covariates that change over the follow-up period (i.e. after baseline) that could affect time to the event[198]. Extended Cox models therefore do not restrict covariates to a single, baseline time-point like traditional Cox models.

Modelling repeated measures data as time-varying covariates in Cox models was common in practice[199, 200] until a recent simulation study showed this approach can result in severe under-estimation of the true association between changes in the covariate and risk of the event[201]. This is due to regression dilution, where estimates are biased towards the null. Additionally,

covariate values are assumed to be measured without error[199, 202]. Furthermore, correlation between and within repeated measurements in subjects is not taken into account[202].

5.2.2 Mixed-effects model:

Mixed-effects models were proposed by Laird & Ware[203] to model non-independent data, such as within-subject repeated measures data. A mixed-effects model can be used to identify how measurements change over time. The mixed-effects model for subject $i = 1, \dots, n$, is denoted by:

$$\boldsymbol{Y_i} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\beta_k x_{ik}} + \sum_{\boldsymbol{p=1}}^{\boldsymbol{P}} \boldsymbol{Z_{ip} b_{ip}} + \boldsymbol{\varepsilon_i} \qquad \text{Eq. 2}$$

Where

1. Y_i is the longitudinal outcome for subject i.
2. $\sum_{k=1}^{l} \beta_{jk} x_{ijk}$ is the predicted mean longitudinal response, taken as the linear combination of $k = 1, \dots, K$ x-covariates with corresponding fixed effect terms β_k.
3. $\sum_{p=1}^{q} Z_{ip} b_{ip}$ is the subject-specific contribution, taken as the linear combination of $p = 1, \dots, P$ Z-covariates with corresponding subject-specific random effect terms b_{ip}, which follow a normal distribution with zero-mean and variance-covariance matrix D. Note, $Z = 1$ for a random intercept.
4. ε_i is the error term, which follows a normal distribution with zero-mean and variance σ^2.

5.2.3 Joint models - linking mixed-effects and Cox models:

Joint modelling of longitudinal and time-to-event data was originally introduced by Wulfsohn *et al*[204] and Henderson *et al*[205]. Joint models use changes over time in repeated measurements

to identify risk of event, such as by modelling the values of a predictor, rate of change (i.e. slope/gradient) in the predictor, or both. They have recently been extended to accommodate multiple longitudinal outcomes[206-209]. A multivariate joint model is comprised of two types of sub-models: the mixed-effects sub-model and Cox sub-model.

In the multivariate joint model setting, a mixed-effects sub-model is developed for each longitudinal outcome separately. Changes over time identified in the mixed-effects models contribute to a single Cox sub-model for risk prediction through additional parameters[210]. These parameters are referred to as association parameters, captured by a multivariate latent Gaussian process[211]. In short, this process involves first extracting the best linear unbiased predictor (BLUP) of each random effect in the mixed-effects sub-model[185, 212]. BLUPs are estimates of the random effects for individual groups, depending on how the models were structured. For example, patients (i.e. the group) can be modelled using random effects to account for correlation/covariance between and within repeated measures in individual patients and the BLUPs are values that incorporate this information. The next step in the process is to take the linear combination of BLUPs for each longitudinal outcome separately and subsequently include them as time-varying covariates in the Cox model. Let:

$$\boldsymbol{\mu_i} = \sum_{\boldsymbol{p=1}}^{\boldsymbol{P}} \boldsymbol{Z_{ip} b_{ip}} \qquad \text{Eq. 3}$$

Be the latent association covariate (to include in the Cox-sub-model), taken as the linear combination of BLUPs for $p = 1, \dots, P$ random effects from the mixed-effects sub-model (Eq. 2). The resulting Cox model is denoted by:

$$\boldsymbol{S_i(t) = S_i(t \mid b_i; \theta) = S_0(t)^{\exp(\sum_{k=1}^{K} \beta_k x_{ik} + \sum_{j=1}^{J} \gamma_j \mu_{ij})}} \quad \text{Eq. 4}$$

Where

1. $S_i(t)$ is the survival probability (i.e. risk of not having the event), at a specified time, t, for subject $i = 1, \dots, n$ (Eq. 1). This can be re-written as $S_i(t \mid b_i; \theta)$ to reflect that the survival probability at time t is conditional on the random effects (BLUPs) for each patient, b_i, and a set of parameters/coefficients, θ, from a fitted joint model.
2. $S_0(t)$ is the baseline survival estimate at time t (Eq. 1).
3. $\sum_{k=1}^{K} \beta_k x_{ik} + \sum_{j=1}^{J} \gamma_j \mu_{ij}$ is referred to as the linear predictor. $\sum_{k=1}^{l} \beta_k x_{ik}$ is as in Eq. 1. $\sum_{j=1}^{J} \gamma_j \mu_{ij}$ is the i^{th} subject's latent association part of the linear predictor, taken as the linear combination of μ_{ij}-latent association covariates for $j = 1, \dots, J$ longitudinal outcomes with corresponding coefficients γ_j. These coefficients are referred to as association parameters.

The interpretation of joint model coefficients is the same as for mixed-effects and Cox models alone[210, 213-215]. Specifically, association parameters have the following interpretation: a one-unit increase in the longitudinal outcome's trajectory increases the log-hazard rate by γ_j, holding the other covariates constant. The corresponding hazard ratio (HR) is the exponential of γ_j.

5.2.4 Dynamic predictions from joint models

Under the joint model framework, the survival probability at the time-point of interest is conditional on the longitudinal data and survival up to that time-point. Therefore, the survival probability from the Cox sub-model (Eq. 4) is estimated twice: once at the time-point of interest and once at the time of the most recent longitudinal measurement. Therefore, both survival

estimates incorporate information from the mixed-effects sub-models, such as random effects estimates (BLUPs), and Cox sub-model, such as the model coefficients and survival/time-to-event function. The ratio of these two survival probabilities is referred to as the dynamic survival probability.

$$\boldsymbol{S_i(t \mid t_{obs}) = \frac{S_i(t \mid b_i; \theta)}{S_i(t_{obs} \mid b_i; \theta)}} \quad \text{Eq. 5}$$

Where

1. $S_i(t \mid b_i; \theta)$ is the survival probability at a specified time, t, of interest, derived using Eq. 4.
2. $S_i(t_{obs} \mid b_i; \theta)$ is the survival probability at the time of the most recent longitudinal measurement, t_{obs}, before t ($t_{obs} < t$), derived using Eq. 4.
3. $S_i(t \mid t_{obs})$ is therefore the dynamic survival probability at a specified time, t, for subject $i = 1, \dots, n$, conditional on survival over the longitudinal period and survival up to that time-point.

For example, hypothetically, assume the interest is to predict one-year risk of an event (time=1) following entry into the study (time=0) for a patient with longitudinal measurements at time=0, 0.5, 0.7, 1, and 1.5 years until study exit at time=2. The survival probability is estimated at one year, $S_i(1)$ (time-point of interest), and at 0.7 years, $S_i(0.7)$ (time of the most recent longitudinal measurement before the time-point of interest). The patient's one-year dynamic survival probability is then $S_i(1)/S_i(0.7)$.

The dynamic risk of having the event is derived as:

$$\boldsymbol{Dynamic\ Risk = 1 - S_i(t \mid t_{obs})} \quad \text{Eq. 6}$$

This risk therefore incorporates information from the longitudinal period, such as trajectories in repeated measures (informed by the BLUPs, i.e. random effects).

5.2.5 Why joint models?

There are many reasons for developing a joint model in this research, including logistical and methodological. A full list of reasons is provided in Appendix D4. For example, logistic and Cox regression models, which are common, rely on risk factors measured at a single, baseline time-point (time=0) to estimate risk of a future event but joint models use risk factors measured at multiple time-points before baseline to estimate this risk. They could therefore use relevant trajectories in risk factors to identify risk of a future event for each patient.

Alternatives to joint modelling exist. A recent systematic review identified seven methodological frameworks for harnessing longitudinal data for risk prediction[216]. The authors identified many advantages of joint models compared to other approaches, such as machine-learning and landmarking[217]. For example, joint models can account for correlation between repeated measures within subjects, longitudinal outcomes, and sub-models, which is not possible in approaches like extended Cox models, where the longitudinal data is included as time-varying covariates and patient-level trajectories and correlation are not modelled. Additionally, joint modelling performs less parameter estimation compared to other approaches, such as machine-learning, and therefore can use smaller sample sizes. Furthermore, statistical models are commonly used in healthcare and are therefore easier to embed into practice compared to approaches like machine-learning, which are often considered “black box” algorithms because their methodology is not transparent and do not instil confidence in clinicians to use such algorithms to make health-related decisions. However, the main drawback of joint models is

their computational intensiveness, which could affect study aspects, such as the study design and predictors, to make the model practical and achievable.

5.3 Methods

The TRIPOD guidelines were followed for the reporting of the prediction model[143]. In this chapter, all data preparation was performed in Stata/SE V15.1 and analyses performed in RStudio (R V4.0.2).

5.3.1 Data

The dataset derived in Chapter 3 was used to develop and internally validate the model. In short, the dataset consists of FBC blood tests from the CPRD database, consisting of patient records from over 600 primary care practices in the UK[149]. The CPRD data is linked to the NCRAS, which indicated if a patient in CPRD was diagnosed with bowel cancer. The clinical codes used to extract FBC results from CPRD are those reported in Chapter 3 (Table 3.1) and diagnoses from NCRAS are those reported in Chapter 4 (Table 4.1).

5.3.2 Study design

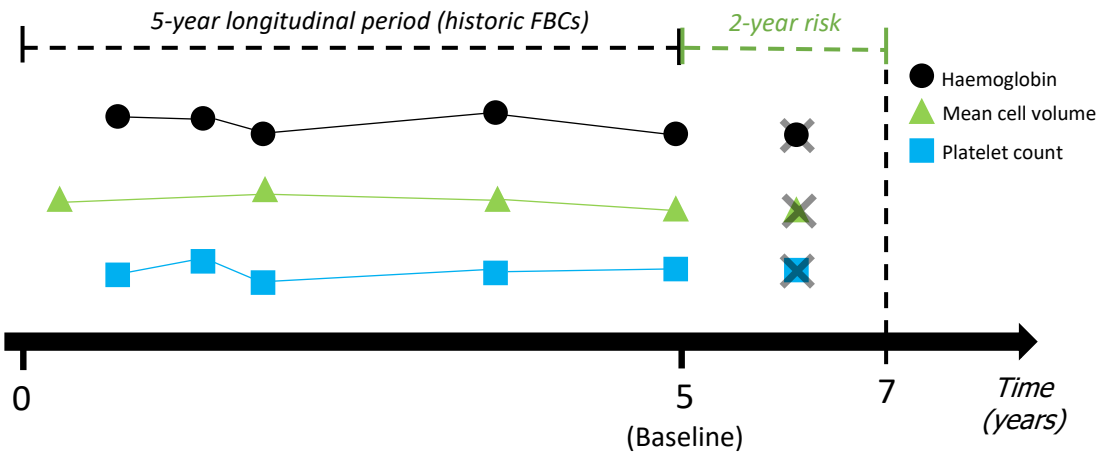
Study entry was defined as the latest of date of registration with the practice, patient's 40th birthday, or 1st January 2000. Study exit was defined as the earliest of date of leaving the practice, date of death, or 14th January 2014 (the NCRAS data-cut date). Data from eligible patients within this study period was used (eligibility criteria are in Section 5.3.3).

I performed a prospective cohort study to develop the model. This design was chosen because it better mimics the real-time patient flow in general practice so is easier to describe and

interpret compared to a case-control design. Additionally, cohort studies allow for a more appropriate assessment of risk compared to a case-control study[218].

The sensitivity analysis performed in Chapter 4 showed different trajectories in a cohort design compared to a case-control design (Appendix C11). In that sensitivity analysis, I used the date of the earliest FBC in the patient's study period as baseline and identified trajectories over 10 years following baseline that did not adequately reflect the rate of change identified in the case-control design. To ensure I derived consistent trajectories using a cohort study design as in a case-control design, I therefore did not use the date of the earliest FBC as baseline for this prediction model. Instead, baseline was the date of the most recent FBC (prior to the two-year exclusion period before diagnosis/study exit date, defined in Section 5.3.4, to ensure each patient has two years of follow-up following baseline FBC). FBCs performed earlier than baseline were considered historic (see Figure 5.1). This approach mimics patient flow in practice, where the current FBC is assessed and previous FBCs are considered historical and usually not considered. Trajectories were identified using historic FBCs over a five-year period up until baseline FBC. Risk predictions are therefore made from the baseline time-point, which incorporates information from historical FBCs.

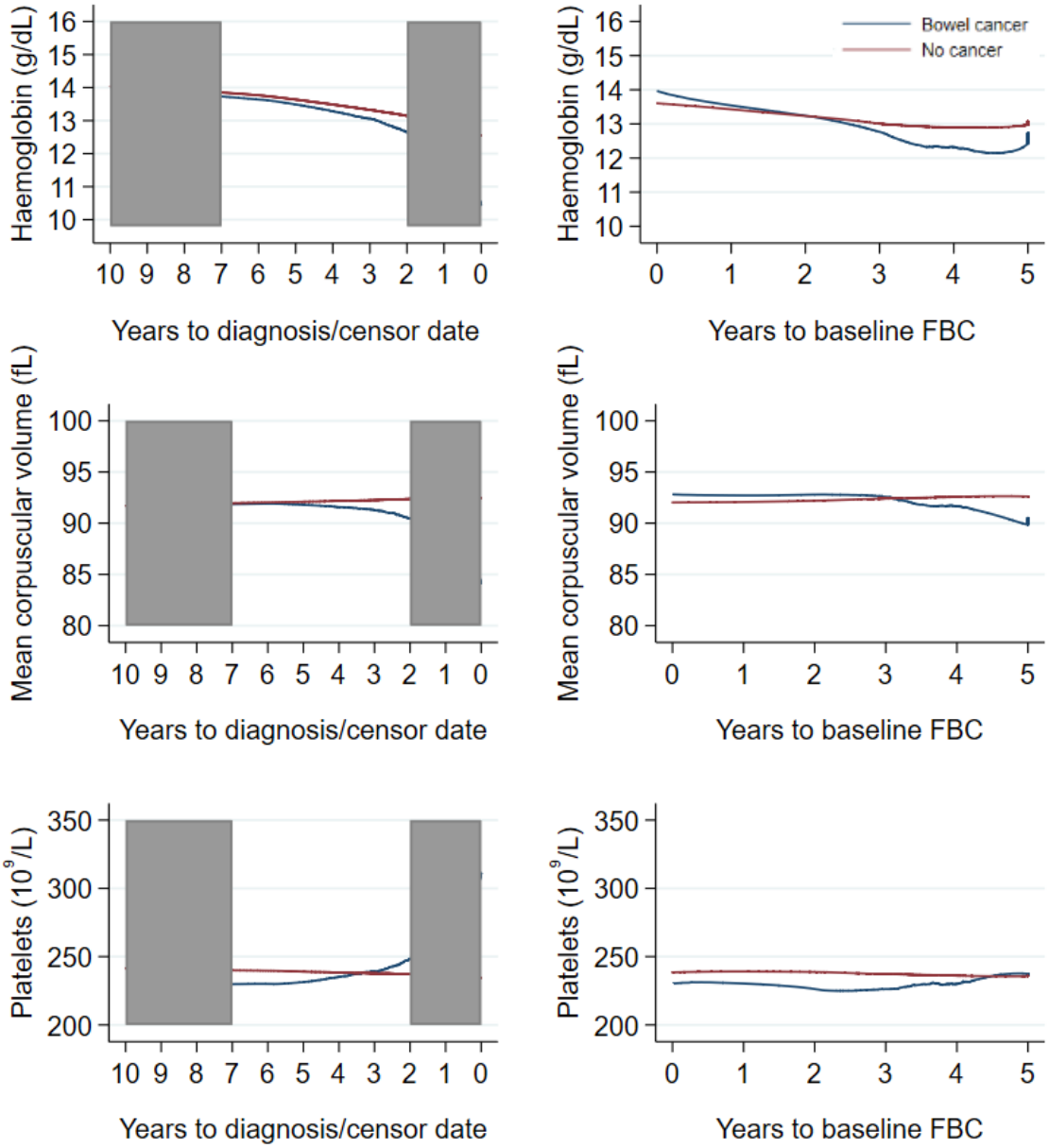
Figure 5.1: Flow of (dummy) longitudinal data for two-year risk of diagnosis



For example, assume a hypothetical patient has five FBCs in their record, with the earliest FBC measured in September 2007 and the most recent (prior to the two-year exclusion period) measured in June 2011. The most recent FBC is considered baseline and corresponds to the end of the five-year longitudinal period (June 2011, time=5). The start of the five-year longitudinal period is therefore five years earlier (June 2006, time=0). The earliest FBC (September 2007) is then measured at time=1.25 years in this five-year period (i.e. one year and three months after the start of the longitudinal period: June 2006 to September 2007). Trajectories among FBCs in this five-year period (time=0 to 5) are used to identify risk of diagnosis in two-years following baseline (time=7).

This approach replicated the trajectories identified in the case-control design but in a cohort study setting (see Figure 5.2). For example, FBC levels in patients not diagnosed were similar between study designs, such as haemoglobin around 14 g/dL and 13 g/dL at the start and end of the longitudinal period in both designs, respectively. In diagnosed patients, levels were similar to patients not diagnosed until around four years before diagnosis in the case-control and the corresponding time point of three years in the cohort design for each parameter. Patients with and without a diagnosis had the same rate of change in levels between study designs.

Figure 5.2: Replicating trajectories (LOWESS) in the prior case-control analysis (left) in this joint model development cohort study (right)[1, 2, 3]



[1]The shaded areas in the case-control (left) panel ensure only the corresponding five-year longitudinal period in the prediction model cohort study (right) panel are presented. In the cohort study, there is a two-year exclusion period pre-diagnosis/study exit date and only five years of FBCs modelled. Hence, the start of the longitudinal period in the prediction model cohort study, time=0, corresponds to time=7 in the case-control design. The end of the longitudinal period in the cohort study is at time=5, which corresponds to time=2 in the case-control design.

[2]In the cohort study (right), time=5 represents the baseline FBC (the most recent FBC available prior to the two-year exclusion period pre-diagnosis/study exit). FBCs over five years prior to baseline are used to identify trajectories that contribute to risk of diagnosis in two years following baseline.

[3]LOWESS trends are for males aged 90 (+/- 3 years) at diagnosis/censor in the case-control analysis and aged 88 (+/- 5 years, to derive a smoother curve) at baseline FBC in the cohort study. These ages ensure the trends presented correspond to the same age group.

The ColonFlag model uses FBCs over three years to derive a score for diagnosis[64] but a five-year period was chosen for this new joint model to allow for more available FBCs to be used. A longer period was not used because joint models are computationally intensive to develop and the trajectories identified in Chapter 4 show differences in FBC levels start around four years before diagnosis, suggesting earlier FBCs are not needed. Limiting to the five-year period made this model more computationally achievable.

5.3.3 Participants

Patient **inclusion** criteria were:

- Aged at least 40 years at study entry
- Had at least one haemoglobin, mean corpuscular volume, and platelet measurement among any of the FBCs in their study period (a requirement of the modelling package).

Patient **exclusion** criteria were:

- Registered with their primary care practice for less than one year
- CPRD record not linked to NCRAS
- Had a history of bowel cancer before study entry
- Diagnosed with another cancer type before or simultaneously with bowel cancer
- Available date of diagnosis but no indication of the cancer type

5.3.4 Outcome

The outcome was a diagnosis of bowel cancer at two years (+/- three months) after baseline FBC.

Firstly, FBCs within the two years (+/- three months) proceeding diagnosis date (for those diagnosed) and study exit date (for those not diagnosed) were excluded. Consequently, if a

patient only had FBCs within the two years before diagnosis/study exit, they contributed no FBCs so were excluded from the study. This two-year exclusion period ensures each patient has two years of follow-up following the baseline FBC. It also ensures only FBCs available (over a five-year period) prior to the baseline FBC (most recent FBC before the two-year exclusion window) are used to identify risk of diagnosis **in** two years, as opposed to **within** two years, to facilitate earlier detection (i.e. there would be no events from baseline until two years (+/- three months) later). This reduces bias because patients diagnosed have more FBCs close together before diagnosis compared to patients not diagnosed.

Patients diagnosed with bowel cancer two years (+/- three months) after their baseline FBC were considered to have the event. Patients not diagnosed were censored at the earliest of study exit or two years after their baseline FBC.

A two-year window was chosen for the following reasons:

1. **Clinical relevance**: within two years, cancers are likely to manifest clinically but the opportunity to modify prognosis requires an adequate time interval for intervention.
2. **Comparability with existing models**: the QCancer Colorectal model in the UK uses a two-year outcome[69]. Additionally, the ColonFlag (Israeli algorithm)[64] that uses changes in FBC parameters over time could discriminate higher-risk from lower-risk patients in the UK at 18-24 months prior to diagnosis (c-statistic = 0.78)[65].
3. **Trajectories towards diagnosis**: the analyses in Chapter 4 showed relevant differences in many FBC levels between diagnosed and non-diagnosed patients, starting around four years before diagnosis. A two-year outcome would allow FBC parameters to commence on the trajectory in diagnosed patients and continue on this trajectory for around two years, which would be identified by the prediction model.

5.3.5 Predictors

5.3.5.1 Candidate predictors:

Year of birth, sex, and FBC results were extracted from the CPRD database as the only candidates to include in the model. A detailed list of reasons explaining why only these were considered is in Appendix D1. In summary, some known risk factors, such as family history of bowel cancer and ethnicity, and confounders to FBC levels, such as diet and altitude, are not accurately recorded or available in EHR datasets. Additionally, the ColonFlag model, which relies on only age, sex, and FBC results, was shown to have good discrimination for bowel cancer detection[65]. Furthermore, joint models are computationally intensive to develop so it would likely not have been possible to include many predictors in the model (with the advanced computational facilities available).

The reasons described in Appendix D1 indicate there are likely too many prognostic factors and confounders to include in a single model, with many being inaccurately recorded in EHRs or not available at all. Too many predictors would also effect the required sample size needed to develop the model and computational capacity needed (Section 5.3.10). My supervisors and I agreed not to overcomplicate the model and that, in practice, GPs would know the patient's full situation and can take appropriate action. This is supported by the design of the ColonFlag, including only age, sex, and FBC results to derive a model that performs adequately. My supervisors and I therefore hypothesised that FBC results, age, and sex would be sufficient to consider as the only candidates for the new model described in this chapter.

5.3.5.2 Final predictors chosen:

A separate joint model was built for males and females. This was due to the computational intensiveness of joint models, allowing as many men and women to be modelled separately than in a single model with all patients, increasing the usable sample size and number of events to

build the model per sex. Additionally, fewer parameters would need estimating, such as the effect of sex and any interactions with sex, resulting in simpler models. Each model was adjusted for age at baseline FBC (years).

A single FBC can include up to 20 parameters. Of these, haemoglobin, mean corpuscular volume, and platelets were included in the model. A detailed list of why the model was limited to these three is in Appendix D2. In summary, reasons include clinical relevance, with haemoglobin and mean corpuscular volume related to iron-deficiency anaemia, the systematic review identifying these as relevant parameters (Chapter 2), weak correlation between the blood levels (Appendix D3), and less computational burden compared to when using more parameters.

5.3.6 Missing data

Age and sex were available for all patients. The methods for handling missing FBC data are described in detail in Section 3.7. In summary, the mixed-effects sub-model of joint models use maximum likelihood estimation to derive parameter estimates, which adjusts for missing data.

5.3.7 Model development

A multivariate joint model of longitudinal and time-to-event data was developed using the *mjoint()* command in the *JoineRML* package in R software, recently developed by Hickey *et al*[208]. A model was developed for males and females separately. In the mixed-effects sub-models, the choice of linear splines and interactions were as reported in Chapter 4 (Section 4.2.5). In short, age at baseline FBC (fixed effect) was modelled using linear splines, with knots at 60, 70, and 80 for haemoglobin, 55 for mean corpuscular volume, and 60 for platelets. Time to baseline FBC (fixed effect) was modelled using linear splines, with a knot at year 3 in the five-year longitudinal period for each FBC parameter. An interaction between time and age (fixed effect) was included in the haemoglobin mixed-effects model only. A random intercept for patient and random slope

for time was used for each FBC parameter, with an unstructured covariance matrix to account for correlated repeated measures. In the Cox sub-model, age at baseline FBC was modelled using fractional polynomials (powers: 2, 2, determined automatically by the software) – see Appendix D5 for a comparison of functional forms for age in the Cox-sub-model. Additionally, the model using fractional polynomials did not violate the proportional hazards assumption (assessed using plots of Schoenfeld residuals – Figure D8). Age-adjusted trajectories in the three FBC parameters were pulled automatically from the mixed sub-models and entered as covariates in the Cox sub-model, as described in Section 5.2.3.

By default, the *JoineRML* package performs 100 burn-in iterations per longitudinal parameter during maximum likelihood estimation (to derive each estimate in the model). Upon assessing model convergence using convergence plots for each estimate, 75 burn-in iterations were used in total. This reduction from 300 to 75 iterations still achieved convergence, but also sped up computation time, as fewer iterations were being performed to estimate parameters. Additionally, antithetic sampling was used, which is a type of Monte Carlo variance reduction technique to speed up convergence. A simulation study showed this approach offers the quickest computation time with large sample sizes compared to ordinary Monte Carlo and Quasi-Monte Carlo methods[219]. The study also shows that antithetic sampling gives comparable coverage rates to other methods, which represent the proportion of simulations where the estimate resides within the corresponding confidence interval[220].

The baseline survival estimate at two-years (from the baseline FBC) is required to derive a two-year risk score from a Cox model (the formula for a Cox-risk equation is in Section 5.2.3). The Breslow baseline survival estimate, which uses mean-centred covariates (to represent 'the average person')[196], is the most widely used approach and has been incorporated in major

statistical packages. The Breslow estimate was estimated by the *JoineRML* package automatically using the *baseHaz()* command.

Two-year risk predictions (for the development, i.e. apparent, cohort) were estimated using the *dynSurv()* command in the *JoineRML* package. This command identifies individual-specific trajectories from the mixed effects sub-model for each of the three FBC parameters, extracts the BLUPs (i.e. random effects) from these models, and includes them in a Cox model to identify each patient's event-free probability based on their FBC trajectories, as described in Eq. 4 (Section 5.2.3) and Eq. 5 (Section 5.2.4). The risk of diagnosis is then one minus this event-free probability (Eq. 6, Section 5.2.3).

5.3.8 Model performance

5.3.8.1 Overall performance, discrimination, and calibration

Overall performance was assessed using Royston and Sauerbrei's (pseudo) R_D^2[139] and the Brier score. R_D^2 is a measure of explained variation, designed for proportional hazards models. Unlike the conventional R^2 measurement, which is not designed for time-to-event models, R_D^2 is on the log hazard scale and may be small (i.e. close to 0). R_D^2 was estimated using the *royston()* command in the *survival* package. The Brier score summarises the difference between observed and predicted outcomes. A score of zero is therefore favoured. The Brier score was calculated using a standard formula[221]:

$$\boldsymbol{Brier\ Score} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \boldsymbol{p_i})^{\boldsymbol{2}}}{\boldsymbol{n}} \qquad \text{Eq. 7}$$

Where

1. y_i is the observed outcome for the i^{th} patient, for $i = 1, \dots, n$ patients (=1 if diagnosed, =0 if not diagnosed).
2. p_i is the probability/risk of diagnosis for the i^{th} patient.

3. $\sum_{i=1}^{n}(y_i - p_i)^2$ is therefore the sum of the squared differences between the observed outcome and risk of diagnosis over all patients, n.

Discrimination, a measure of how well the model separates patients who do and do not have the outcome, was assessed using the c-statistic (or AUC) and Royston and Sauerbrei's D-statistic. The c-statistic was estimated using the *roc()* command in the *pROC* package. The conventional rule-of-thumb of ≥0.7 was used to indicate good discrimination. The D-statistic was designed for proportional hazards models to account for censored observations. It is the log hazard ratio of two prognostic groups, derived as the linear predictor (defined in Section 5.2.3) dichotomised at the median. Therefore, a D-statistic of 0 means no discriminative ability, with higher values representing greater discriminative ability. The D-statistic was estimated using the *royston()* command in the *survival* package.

Calibration, which measures the agreement between observed and predicted outcomes, was assessed using the calibration slope and calibration plots. The calibration slope is the coefficient of the linear predictor, included as the only covariate in a separate Cox model. A calibration slope of 1 indicates perfect calibration. Calibration plots were derived following the methods by Royston[222, 223]. This involves first categorising patients into 20 equally sized groups of predicted two-year risk (i.e. 20 risk groups) and the mean of the predicted risk compared with the observed risk for each risk group separately graphically. The observed two-year risk for each group was estimated using the Kaplan-Meier survival function to account for censored observations.

5.3.8.2 Diagnostic test accuracy and risk thresholds

Risk thresholds corresponding to centiles of highest risk (from 1% to 99%) were derived for the split sample internal validation cohort only (defined in Section 5.3.9.2). For each threshold,

sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. A ROC curve was also derived.

5.3.9 Internal validation

Internal validation allows for assessment and correction of optimistic performance, where performance measures are too large, and overfitting[125, 138, 140, 224, 225]. For example, overfitting can be reduced using shrinkage techniques, which shrink the coefficients in a model by some shrinkage factor determined during internal validation[226-229]. To ensure a robust internal validation, two approaches were used: bootstrapping and a random split sample.

5.3.9.1 Bootstrapping

Bootstrapping is widely favoured for internal validation, as it uses the entire cohort to build the model and iteratively resamples from the cohort to derive validation cohorts[125, 230, 231]. The amount of optimistic performance and shrinkage factor (to reduce overfitting) can be identified through bootstrapping using a standard approach[125, 140, 232-235]:

1. Build the model in the development cohort and assess model performance in this cohort ($P_{original}$).
2. Derive a bootstrap sample by resampling from the development cohort.
3. Build the model in the bootstrap sample.
4. Assess performance of the bootstrapped model in the same bootstrap sample ($BP_{bootstrap}$) and the original development sample ($BP_{original}$).
5. Calculate optimism: $BP_{optimism} = BP_{bootstrap} - BP_{original}$
6. Repeat steps 2-5 many times, e.g. 100.
7. Calculate average optimism ($BP_{avg.optimism}$).

Average optimism is then deducted from the performance measure of the original model in the development sample ($P_{original}$) to derive optimism-corrected performance:

$$\boldsymbol{P_{optimism-correctected} = P_{original} - BP_{avg.optimism}} \quad \text{Eq. 8}$$

The optimism-corrected calibration slope is referred to as a uniform shrinkage factor. All coefficients in the model can be multiplied by this shrinkage factor to shrink the estimates and therefore reduce overfitting. However, optimism and overfitting are expected to be low when sufficiently large sample sizes are used (sample size calculation in Section 5.3.10).

In this study, I used bootstrapping with three bootstrap samples to derive optimism-corrected performance measures for each model. A rule-of-thumb is to use at least 100 samples to increase the precision of the amount of optimism but three were used here for practicality. This is still favourable to other approaches for internal validation, such as random data splitting, which is inefficient because it reduces the sample size used to build the model and provides only one sample to assess optimism so provides a less precise assessment.

5.3.9.2 Split sample

Joint models are computationally intensive and burden on computer capacity increases when combined with big datasets. A conventional working laptop offers either 8 or 16 GB RAM, which is a built-in memory slot that stores data whilst being processed. I used an advanced 'super-computer' to develop the model, which offered approximately 3,100 GB RAM. However, this was still insufficient for the entire patient sample and could facilitate approximately 226,000 patients only. Therefore, the final development sample consisted of around 226,000 randomly chosen males and females separately, which used as much of the available computer capacity as possible, whilst making modelling achievable. The remaining patients were reserved as a split

sample cohort for internal validation. Amount of optimistic performance and overfitting were not assessed using the split sample, as bootstrapping was used as the primary approach (on the development cohort).

This split sample internal validation involved deriving traditional performance measures, such as measures of discrimination and calibration, to assess how well the models work in a separate cohort to the development cohort. I assessed how the models perform overall in the split sample and for subgroups of age and by number of FBCs available in the five-year longitudinal period. Additionally, I compared performance measures between my joint models and the ColonFlag algorithm by Kinar 2016[64]. My joint models were developed because the ColonFlag, an Israeli machine-learning algorithm that uses changes in FBC levels over time to derive a monotone score for diagnosis, showed good discrimination in UK data for two-year risk (c-statistic: 0.78)[65]. Performance of my joint models is expected to be at least comparable to the ColonFlag.

The ColonFlag score was derived for each FBC by Medial EarlySign staff, who were given the UK dataset by my supervisors. They performed their own data preparation and ran the machine-learning algorithm in confidence. They did not disclose their methods or reasons for patient exclusions to anyone involved in the research. Once derived, Medial EarlySign staff returned a monotone score from 0-100 (0=lowest risk, 100=highest risk) for each FBC test. The score incorporates changes in (repeated) FBC data measured around 15 and 33 months prior to the current FBC. In this split sample, I excluded patients who had no ColonFlag score available or no ColonFlag score corresponding to the baseline FBC to ensure I tested each model on the same patient sample and risk scores from the same FBC.

The ColonFlag algorithm was originally developed for one cohort, including sex as a covariate, whereas I developed joint models for males and females separately. To allow a direct

comparison, the ColonFlag was assessed for males and females separately. Only the c-statistic was derived for the ColonFlag algorithm due to the nature of the machine-learning algorithm, which provides a monotone score and not a measure of absolute risk, hence the Brier score, R_D^2, D-statistic, calibration slope, and calibration plots could not be derived.

Performance of the joint models and ColonFlag was assessed for the overall split sample cohort. Additionally, the c-statistic and calibration plots were derived for subgroups of 10-year age bands (from 40 to 90 years) and subgroups of number of FBCs available (2 vs 4 vs 6… 14 FBCs, with limited sample sizes beyond 14 FBCs).

5.3.10 Sample size

The required sample size was calculated *a-priori* using the *pmsampsize* package in Stata software, recently developed by Riley *et al*[236] to improve on the commonly-used 10 events per variable rule-of-thumb[237]. There are not many examples yet where this new package has been used to plan a study. The package is designed for conventional modelling approaches, such as logistic and Cox models, so do not provide sample sizes for joint models, although a Cox sub-model is used in the joint model. However, it was used here to provide an indication of the minimum sample size.

As predictions from a joint model are made using a Cox sub-model, a time-to-event outcome was assumed in *pmsampsize*. The package uses the mean follow-up time in the available data, number of proposed predictor parameters in the model, and the expected event rate, Cox-Snell R^2 (can be calculated from the AUC if not reported[238]), and amount of shrinkage (to adjust for overfitting) based on existing evidence, such as existing prediction models.

The mean follow-up time in the available CPRD data was 3.8 years for males and 4.1 years for females. Five predictor parameters were planned for inclusion in the Cox sub-model: two fractional polynomial terms for age at baseline FBC and one association covariate (defined as the linear combination of BLUPs from linear mixed effects models, see Section 5.2) for each of the three FBC parameters. The internal validation of the QCancer Colorectal prediction model for two-year risk of bowel cancer reported a 0.2% two-year event rate for both males and females. The Cox-Snell R^2 was 0.003079 for males and 0.0029112 for females. Although the authors did not report assessing overfitting, performance measures were similar between development and internal validation cohorts, suggesting little overfitting. Therefore, a 0.9 shrinkage factor was assumed (shrinkage factor =1 represents no overfitting). Based on these estimates, 14,591 males and 15,433 females are required, with 123 and 130 events, respectively.

5.4 Results (development)

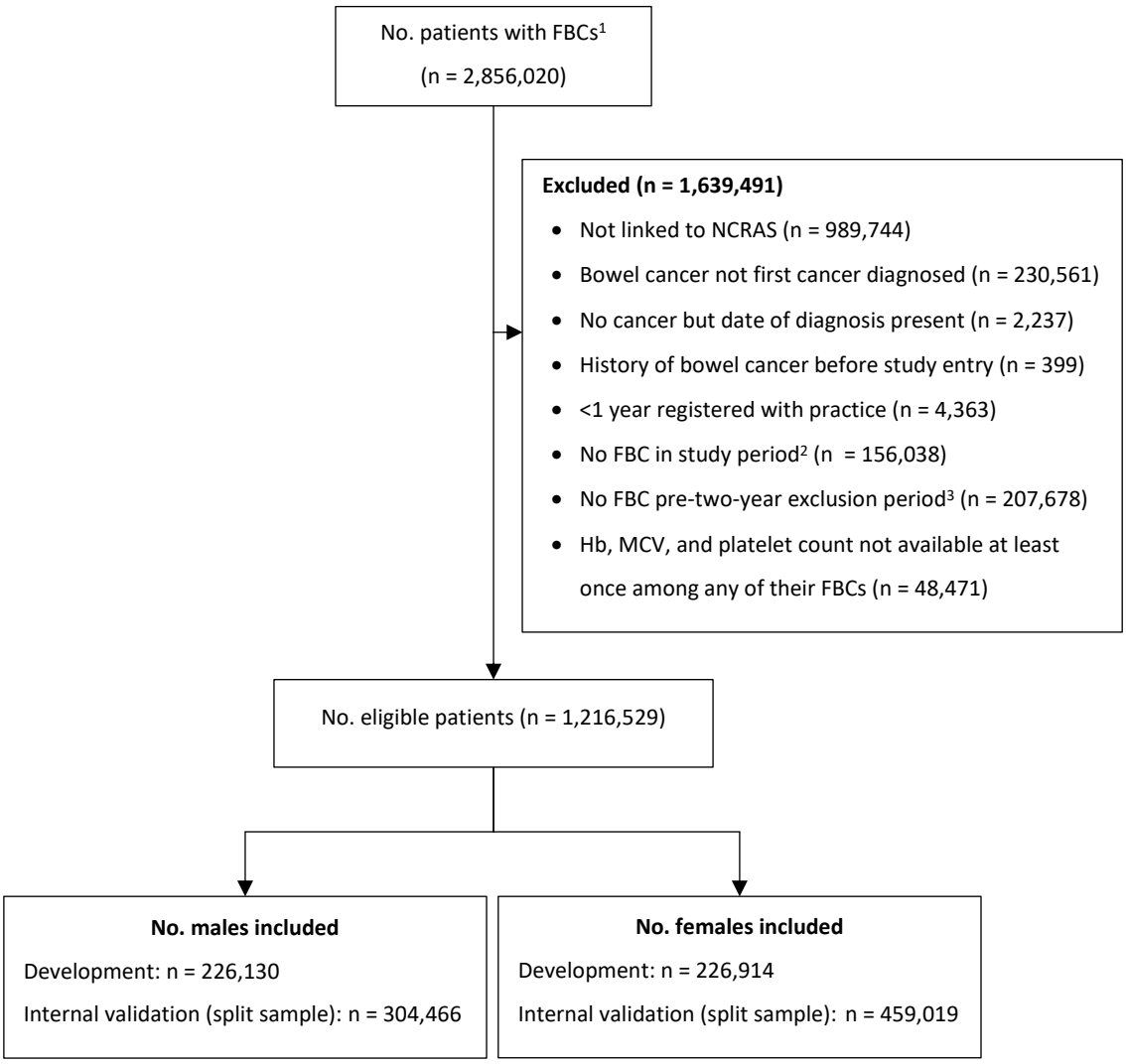
5.4.1 Summary of patient data

Figure 5.3 shows a patient flow diagram. Of 530,596 eligible males and 685,933 eligible females, 42.6% (n=226,130) and 33.1% (n=226,914) were randomly assigned to the development cohort to make modelling achievable (as described in Section 5.3.9.2). The remaining patients were considered a split sample for internal validation of the joint models (results for these patients are in Section 5.5). Of the 226,130 males and 226,914 females used to develop the model, 865 (0.4%) and 677 (0.3%) were diagnosed with bowel cancer two years (+/- three months) after their baseline FBC, respectively. This exceeded the sample size requirements (Section 5.3.10).

A summary of patient characteristics is in Table 5.1 and a summary of FBC and follow-up data is in Table 5.2. On average, patients diagnosed with bowel cancer were approximately 10 years older than patients who were not diagnosed at baseline FBC. This is expected because bowel

cancer is more common in older age groups. The median number of FBCs and follow-up time were balanced between patients with and without a diagnosis.

Figure 5.3: Patient flow diagram (cohort study to develop joint model)



[1]Number of patients available in the CPRD data extract after data validation and quality assessment (Chapter 3).
[2]Study entry was the latest of date of registration with the practice, patient's 40th birthday, or 1st January 2000. Study exit was the earliest of date of leaving the practice, date of death, or 14th January.
[3]FBCs within the two years (+/- three months) proceeding the diagnosis/study exit date were excluded. If a patient had no FBCs before this two-year exclusion period, the patient was excluded.
Abbreviations: FBC=full blood count; NCRAS=National Cancer Registration and Analysis Service; Hb=haemoglobin; MCV=mean corpuscular volume.

Table 5.1: Summary of patient characteristics (joint model development)

	Males		Females	
	Diagnosed	**Not diagnosed**	**Diagnosed**	**Not diagnosed**
No. (%)	865 (0.4%)	225,265 (99.6%)	677 (0.3%)	226,237 (99.7%)
Mean age[1] (SD)	70.9 (10.0)	59.8 (12.9)	73.3 (11.0)	61.4 (14.6)
Age[1] range	40-95	40-104	40-96	40-108

[1]Age (years) at baseline FBC (most recent FBC available prior to two years before diagnosis/censor).

Table 5.2: Summary of FBC data and follow-up (joint model development)

	Males		Females	
	Diagnosed	**Not diagnosed**	**Diagnosed**	**Not diagnosed**
No. FBCs in total	3,774	718,377	3,573	838,165
Mean no. FBCs (SD)	5.6 (7.8)	4.4 (6.8)	7.0 (9.0)	5.3 (8.0)
Age 40-49 years	*2.6 (1.7)*	*2.9 (4.9)*	*2.8 (1.8)*	*3.4 (5.4)*
Age 50-59 years	*5.7 (8.0)*	*4.1 (7.7)*	*6.5 (9.3)*	*4.9 (7.7)*
Age 60-69 years	*4.1 (4.9)*	*4.7 (7.1)*	*10.8 (13.1)*	*5.9 (9.3)*
Age 70-79 years	*5.9 (8.4)*	*5.2 (7.2)*	*5.9 (7.3)*	*6.5 (9.5)*
Age 80-89 years	*6.9 (9.1)*	*5.0 (5.5)*	*5.7 (5.8)*	*5.6 (7.3)*
Age 90+ years	*5.0 (4.5)*	*4.3 (4.1)*	*3.8 (2.3)*	*4.6 (5.5)*
Median no. FBCs (range)	3 (1-61)	2 (1-193)	3 (1-70)	4 (1-172)
Age 40-49 years	*2 (1-8)*	*2 (1-75)*	*2 (1-8)*	*2 (1-86)*
Age 50-59 years	*3 (1-41)*	*2 (1-193)*	*3 (1-41)*	*2 (1-89)*
Age 60-69 years	*3 (1-39)*	*3 (1-12)*	*5 (1-70)*	*3 (1-172)*
Age 70-79 years	*3 (1-59)*	*3 (1-105)*	*4 (1-44)*	*3 (1-149)*
Age 80-89 years	*4 (1-61)*	*3 (1-85)*	*4 (1-35)*	*3 (1-90)*
Age 90+ years	*4 (1-18)*	*3 (1-51)*	*3 (1-10)*	*3 (1-70)*

No. (%) missing Hb[1]	25 (0.66%)	4,017 (0.56%)	34 (0.95%)	4,551 (0.54%)
No. (%) missing MCV[1]	174 (4.61%)	16,697 (2.32%)	182 (5.09%)	18,144 (2.16%)
No. (%) missing platelets[1]	150 (3.97%)	15,767 (2.19%)	167 (4.97%)	19,057 (2.27%)
Median time[2] (range)	2.2 (0-5.0)	1.5 (0-5.0)	2.7 (0-5.0)	2.1 (0-5.0)
Median follow-up[3] (range)	4.18 (1.75-7.18)	3.47 (1.75-7.25)	4.67 (1.76-7.21)	4.06 (1.75-7.25)

[1]Percentages are out of total number of FBCs in that group.
[2]Time (years) between the first and last FBC in the five-year longitudinal period among patients.
[3]Follow-up per patient is the time (years) from first FBC to diagnosis/censor.
Abbreviations: FBC=full blood count; Hb=haemoglobin; MCV=mean corpuscular volume.

5.4.2 Joint model for two-year risk of bowel cancer

Among the 226,130 males, there were 718,109 haemoglobin, 705,280 mean corpuscular volume, and 706,234 platelet measurements used to build the model. For the 226,914 females, this was 837,153 haemoglobin, 823,412 mean corpuscular volume, and 822,514 platelet measurements.

The final multivariate joint model for males and females is provided in Table 5.3 (fixed effects from mixed-effects sub-model), Table 5.4 (random effects variance-covariance matrix from mixed-effects sub-model) and Table 5.5 (Cox sub-model). Each joint model took approximately 12 hours of continuous computation time to develop. Obtaining risk predictions, which are required to assess model performance (Section 5.4.4), took approximately 72 hours of continuous computation time for each model separately.

Table 5.3: Mixed-effects sub-models from the joint models (fixed effects)

Variable	Males			Females		
	Haemoglobin	**Mean corpuscular volume**	**Platelets**	**Haemoglobin**	**Mean corpuscular volume**	**Platelets**
Intercept	15.666 (95% CI=15.497-15.836)	86.165 (95% CI=85.887-86.442)	281.327 (95% CI=278.515-284.139)	11.441 (95% CI=11.311-11.57)	86.457 (95% CI=86.175-86.74)	299.518 (95% CI=296.58-302.456)
Age[1] (years)	-0.01 (95% CI=-0.014--0.007)	0.082 (95% CI=0.076-0.087)	-0.41 (95% CI=-0.462--0.359)	0.036 (95% CI=0.033-0.038)	0.066 (95% CI=0.06-0.072)	-0.217 (95% CI=-0.272--0.163)
Age[1] – knot at 55 (years)		-0.039 (95% CI=-0.046--0.032)			-0.051 (95% CI=-0.058--0.044)	
Age[1] – knot at 60 (years)	-0.014 (95% CI=-0.021--0.008)		-0.127 (95% CI=-0.204--0.049)	-0.038 (95% CI=-0.044--0.033)		0.076 (95% CI=-0.004-0.157)
Age[1] – knot at 70 (years)	-0.026 (95% CI=-0.033--0.019)			-0.03 (95% CI=-0.036--0.023)		
Age[1] – knot at 80 (years)	-0.002 (95% CI=-0.01-0.006)			0.005 (95% CI=-0.001-0.011)		
Time[2] (years)	-0.026 (95% CI=-0.083-0.031)	0.035 (95% CI=0.024-0.045)	-2.019 (95% CI=-2.2--1.837)	0.067 (95% CI=0.024-0.11)	0.053 (95% CI=0.042-0.064)	-1.524 (95% CI=-1.69--1.359)
Time[2] – knot at 3 (years)	0.11 (95% CI=0.033-0.188)	0.169 (95% CI=0.156-0.183)	-0.374 (95% CI=-0.639--0.11)	0.203 (95% CI=0.142-0.264)	0.209 (95% CI=0.196-0.223)	-1.296 (95% CI=-1.533--1.059)
Time[2] by age[1] interaction	-0.001 (95% CI=-0.002-0)			-0.002 (95% CI=-0.002--0.001)		
Time[2] (knot at 3) by age[1] interaction	-0.001 (95% CI=-0.002-0)			-0.003 (95% CI=-0.004--0.002)		
Time[2] by age[1] (knot at 60) interaction	-0.004 (95% CI=-0.006--0.002)			-0.004 (95% CI=-0.006--0.003)		

Time[2] (knot at 3) by age[1] (knot at 60) interaction	0.005 (95% CI=0.002-0.007)			0.007 (95% CI=0.005-0.01)		
Time[2] by age[1] (knot at 70) interaction	0.002 (95% CI=0-0.005)			0.004 (95% CI=0.002-0.006)		
Time[2] (knot at 3) by age[1] (knot at 70) interaction	-0.008 (95% CI=-0.011--0.005)			-0.007 (95% CI=-0.01--0.004)		
Time[2] by age[1] (knot at 80) interaction	-0.004 (95% CI=-0.007--0.002)			-0.001 (95% CI=-0.003-0.001)		
Time[2] (knot at 3) by age[1] (knot at 80) interaction	0.008 (95% CI=0.004-0.011)			0.002 (95% CI=0-0.005)		

[1]Age (years) at baseline FBC (most recent FBC available prior to two years before diagnosis/censor).

[2]Time (years) from start of the five-year longitudinal period to the baseline FBC. The baseline FBC corresponds to the end of the five-year longitudinal period.

Table 5.4: Mixed-effects sub-models from the joint-models (random effects: variance-covariance matrix[1])

			Haemoglobin		Mean corpuscular volume		Platelets	
			Intercept[2]	**Slope[2]**	**Intercept[2]**	**Slope[2]**	**Intercept[2]**	**Slope[2]**
Males	**Haemoglobin**	**Intercept[2]**	1.59520	-0.144660	2.04870	-0.318870	-23.3640	4.0977
		Slope[2]		0.044209	-0.33267	0.085664	3.4132	-1.0402
	Mean corpuscular volume	**Intercept[2]**			26.40100	-1.780500	-55.2080	11.2740
		Slope[2]				0.569270	10.8570	-3.1867
	Platelets	**Intercept[2]**					4101.4000	-242.3500
		Slope[2]						70.7910
	Residual variance		*0.7519*		*2.3297*		*39.6766*	
Females	**Haemoglobin**	**Intercept[2]**	1.4936	-0.1431	3.6875	-0.5019	-24.3680	4.3766
		Slope[2]		0.0418	-0.5266	0.1316	3.7666	-1.1432
	Mean corpuscular volume	**Intercept[2]**			34.2360	-2.7352	-100.6200	18.0650
		Slope[2]				0.7784	16.7710	-4.8491
	Platelets	**Intercept[2]**					5034.1000	-320.9600
		Slope[2]						94.4680
	Residual variance		*0.7294*		*2.6459*		*41.0801*	

[1]Diagonals represent variance and non-diagonals represent co-variance

[2]Random intercept for patient and random slope for time over the five-year longitudinal period prior to the baseline FBC

The fixed effects estimates from mixed-effects sub-models represent how the blood levels change by age and over time, on average. Fixed-effects have been modelled using a non-linear functional form and, in the haemoglobin model, with an interaction so are not straightforward to interpret.

Table 5.5: Cox sub-models from the joint models

Variable	**Males**	**Females**
	HR (95% CI)	**HR (95% CI)**
Age^2 (years)[1]	1.016 (1.014, 1.018)	1.015 (1.013, 1.016)
$Age^2 \times \log(Age)$ (years)[1]	0.997 (0.996, 0.997)	0.997 (0.997, 0.997)
Increase in haemoglobin (g/dL)	0.874 (0.827, 0.922)	0.873 (0.814, 0.936)
Increase in mean cell volume (fL)	0.996 (0.983, 1.009)	0.985 (0.971, 0.999)
Increase in platelets (10^{12}/L)	1.001 (1.000, 1.002)	1.002 (1.001, 1.003)
***Baseline two-year survival*[2]**	***0.9999407***	***0.9999606***

[1]Age (years) at baseline FBC (most recent FBC available prior to two years before diagnosis/censor).
[2]Breslow estimate.
Abbreviations: HR=hazard ratio; CI=confidence interval.

The Cox sub-model (Table 5.5) provides hazard ratios that reflect how much the hazard rate changes by as the blood level increases by one unit from the average population trajectory (patients without a diagnosis). These estimates are combined with the baseline survival estimate to give a risk of diagnosis (see Section 5.2.3). A (patient-level) decline in haemoglobin and mean corpuscular volume and rise in platelet count from the average population trajectory (non-cases), identified from mixed-effects sub-models, increases the two-year risk of diagnosis.

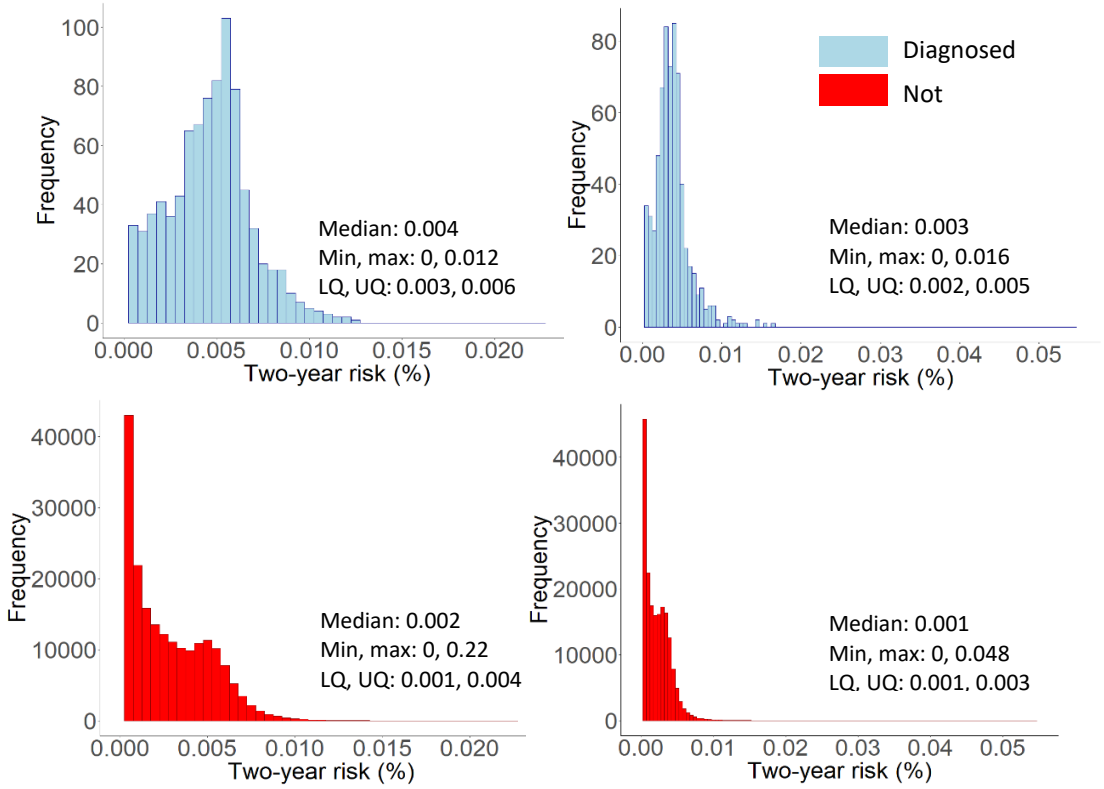
5.4.3 Dynamic predictions from the joint model

Dynamic risk predictions, which consider the trajectories in FBC levels, are described in Figure 5.4. The range of values the predictions take is small, which is likely due to the nature of dynamic

predictions, as these are a ratio of two event-free probabilities: one at the outcome time-point of interest and one at the time of the most recent longitudinal measurement before the outcome time-point of interest (Eq. 6, Section 5.2.4). This ratio would be close to 1 when event-free probabilities at both time-points are similar, which was the case in this sample. Specifically, two-year event-free probabilities were close to 1 for each patient, likely due to the nature of Cox models, which rely on the number of and time-to-events – there was a low event rate (0.4% males and 0.3% females diagnosed) and events only occurred at two years (+/- 3 months from baseline FBC, as per the nature of the study design. Event-free probabilities at the time of the most recent longitudinal measurement before the outcome time-point of interest (i.e. at baseline FBC) equalled 1 for all patients, as no events were included at that time-point. As event-free probabilities at two-years and at the most recent longitudinal measurement before then (i.e. at baseline FBC) were similar, dynamic event-free predictions were relatively close to 1 and did not span the entire range of possible values (0-1). Hence, dynamic risk predictions (1 – dynamic event-free predictions) are close to 0.

On average, patients diagnosed had higher risks than patients not diagnosed (median: 0.004 diagnosed vs. 0.002 not diagnosed males and 0.003 diagnosed vs. 0.001 not diagnosed females). These differences are small on an absolute scale, likely for the reasons described above, but are substantial on a relative scale.

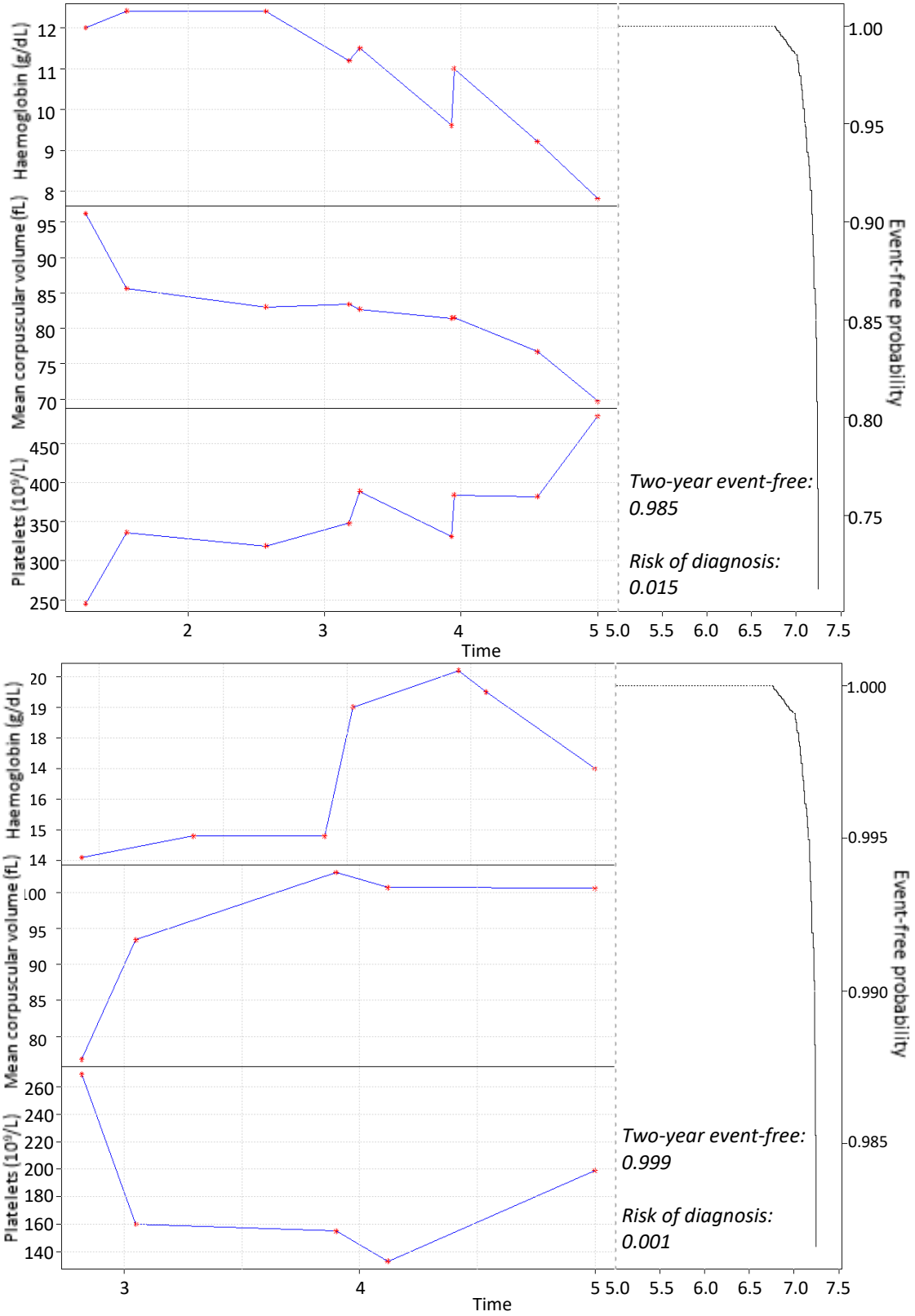
Figure 5.4: Distribution of dynamic risk predictions for males (left) and females (right) in the derivation cohort



Patient examples:

Dynamic plots showing risk of diagnosis based on the patient's longitudinal trajectories are in Figure 5.5 for one diagnosed and one not diagnosed patient, as examples. Both patients are randomly selected out of all females. The diagnosed female was aged 72 years and not diagnosed female was aged 75 years at baseline FBC (most recent FBC). The diagnosed patient had declining haemoglobin and mean corpuscular volume and rising platelets over the five-year longitudinal period, which follow trajectories previously reported in bowel cancer patients (Chapter 4). The two-year risk of diagnosis for this patient is 1.5%. However, the patient without a diagnosis has trajectories that do not follow those observed in diagnosed patients so this patient has a lower two-year risk of 0.1%. These predicted risks are similar (1.5% vs. 0.1%) on an absolute scale but differ substantially on a relative scale.

Figure 5.5: Dynamic prediction plots for a female diagnosed (top) and not diagnosed (bottom)[1]



[1]Time=5 is the baseline FBC. Trajectories over five years before baseline contribute to two-year risk of diagnosis (time=7) from baseline FBC.

5.4.4 Performance of the joint models

Model performance statistics are provided in Table 5.6. Both apparent performance and optimism-corrected performance are reported. The former is the performance in the development sample and the latter is corrected for optimistic performance and overfitting derived using bootstrapping (see Section 5.3.9). Optimism was low so optimism-corrected performance is very similar to apparent performance. The Brier score for overall performance, where a lower score represents a better fitting model, was close to zero, suggesting little difference between observed and predicted risks for both males and females. The c-statistic, a discrimination measure of how well the model separates patients who do and do not have the outcome, was approximately 0.76 in both cohorts for both males and females. This exceeds the 0.7 rule-of-thumb, widely used to represent a good model, suggesting the models can discriminate high-risk from low-risk patients using only earlier data available up to two years before diagnosis. Additionally, the calibration slope was close to 1, suggesting very little overfitting.

Table 5.6: Performance measures of the joint model (joint model development)

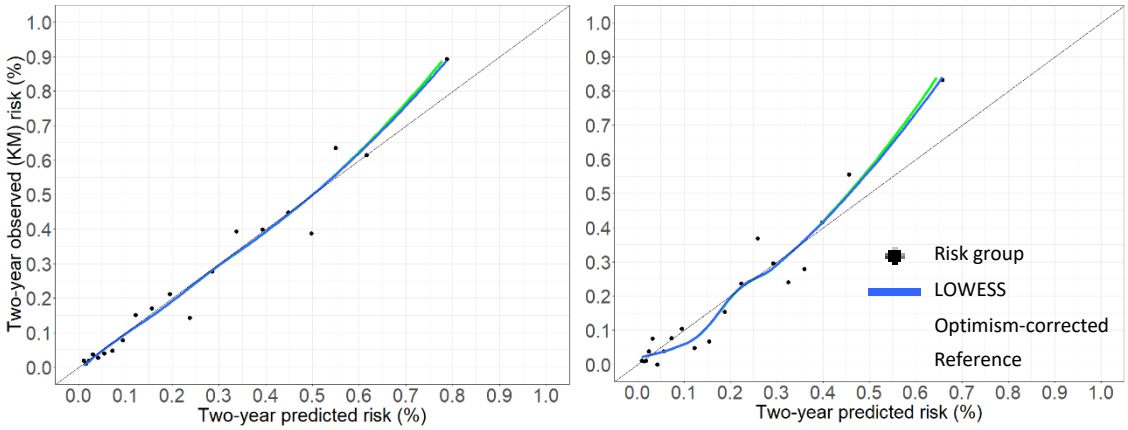
Performance measure	**Males**			**Females**		
	Apparent[1]	**Average optimism[2]**	**Optimism-corrected[3]**	**Apparent[1]**	**Average optimism[2]**	**Optimism-corrected[3]**
Brier score	0.0038	<0.0001	0.0038	0.0030	<0.0001	0.0030
R_D^2	0.31	<0.01	0.31	0.33	<0.01	0.33
C-statistic	0.758	-0.004	0.762	0.761	<0.001	0.761
D-statistic	1.37	-0.02	1.39	1.43	-0.02	1.44
Calibration slope	1.00	-0.02	1.02	1.00	-0.02	1.02

[1]Apparent measures describe performance in the development sample used to build the model.
[1]Average optimism reflects whether the performance measures are too large and is calculated via bootstrapping.

[3]Optimism-corrected measures are apparent performance measures corrected for optimism, i.e. are usually shrunk to correct for performance being too large.

Calibration plots comparing observed and predicted two-year risk are in Figure 5.6, with a reference (y=x) line that represents perfect calibration included. Calibration was good for both males and females, with observed and predicted two-year risk closely matching (overlay the reference line). The curves diverge slightly from the reference line at the highest predicted risk, but this is likely because there is only one risk group in that region. There was weak overfitting, as the optimism-corrected curve matches closely with the apparent curve (i.e. in the development sample).

Figure 5.6: Calibration plot for the joint models for males (left) and females (right) in the development cohort[1]



Abbreviations: KM=Kaplan-Meier

5.5 Results (split sample internal validation)

5.5.1 Summary of patient data

There were 304,466 eligible males and 459,019 eligible females included in this study. Of these, 5.3% (n=16,026) of males and 5.3% (n=24,267) of females were excluded because they had no ColonFlag score available (derived by Medial EarlySign). Patient characteristics and amount of FBC data were similar between patients with and without a ColonFlag score (Table 5.7). Of the

remaining 288,440 males and 434,752 females, 3,894 males and 5,804 females were removed, as they had no ColonFlag score at baseline FBC. The final dataset included 284,546 males and 428,948 females, with 0.4% (n=1,040) and 0.3% (n=1,200) diagnosed with bowel cancer two years after their baseline FBC. This event rate is consistent to the development cohort.

Table 5.7: Characteristics of patients with and without a ColonFlag score (split sample internal validation)

	Males (n=304,466)		**Females (n=459,019)**	
	ColonFlag present	**ColonFlag not present**	**ColonFlag present**	**ColonFlag not present**
No. (%)	288,440 (94.7%)	16,026 (5.3%)	288,440 (94.7%)	16,026 (5.3%)
Mean age[1] (SD)	59.9 (12.9)	59.8 (13.1)	59.9 (12.9)	59.8 (13.1)
Median no. FBCs (range)	2 (1, 141)	2 (1, 78)	2 (1, 141)	2 (1, 78)
Median time[2] (range)	1.5 (0, 5)	1.1 (0, 5)	1.5 (0, 5)	1.1 (0, 5)

[1]Age (years) at baseline FBC (most recent FBC available prior to two years before diagnosis/censor).
[2]Time (years) between the first and last FBC in the five-year longitudinal period.

A summary of patient characteristics is in Table 5.8 and a summary of FBC and follow-up data is in Table 5.9. On average, patients diagnosed with bowel cancer were approximately 10 years older at baseline FBC than patients not diagnosed. This is expected because bowel cancer is more common in older age groups. The median number of FBCs and median follow-up time were balanced between patients with and without a diagnosis. On average, patient characteristics, event rate, number of FBCs, amount of missing FBC data, and amount of follow-up were similar to the development cohort.

Table 5.8: Summary of patient characteristics (split sample internal validation)

	Males		Females	
	Diagnosed	**Not diagnosed**	**Diagnosed**	**Not diagnosed**
No. (%)	1,040 (0.4%)	283,506 (99.6%)	1,040 (0.4%)	283,506 (99.6%)
Mean age[1] (SD)	71.6 (10.2)	59.8 (12.9)	71.6 (10.2)	59.8 (12.9)
Age[1] range	40-95	40-109	40-95	40-109

[2]Age (years) at baseline FBC (most recent FBC available prior to two years before diagnosis/censor).

Table 5.9: Summary of FBC data and follow-up (split sample internal validation)

	Males		Females	
	Diagnosed	**Not diagnosed**	**Diagnosed**	**Not diagnosed**
No. FBCs in total	4,513	907,361	6,370	1,584,250
Mean no. FBCs (SD)	5.3 (7.3)	4.5 (7.0)	8.6 (12.1)	5.3 (8.3)
Median no. FBCs (range)	3 (1-58)	2 (1-141)	4 (1-79)	3 (1-129)
No. (%) missing Hb[1]	33 (0.73%)	4,451 (0.49%)	51 (0.80%)	7,788 (0.49%)
No. (%) missing MCV[1]	104 (2.30%)	15,271 (1.68%)	209 (3.28%)	26,355 (1.66%)
No. (%) missing platelets[1]	119 (2.64%)	17,622 (1.94%)	229 (3.59%)	32,591 (2.06%)
Median time[2] (range)	2.3 (0-5.0)	1.5 (0-5.0)	2.5 (0-5.0)	2.1 (0-5.0)
Median follow-up[3] (range)	4.25 (1.75-7.19)	3.46 (1.75-7.25)	4.45 (1.76-7.19)	4.07 (1.75-7.25)

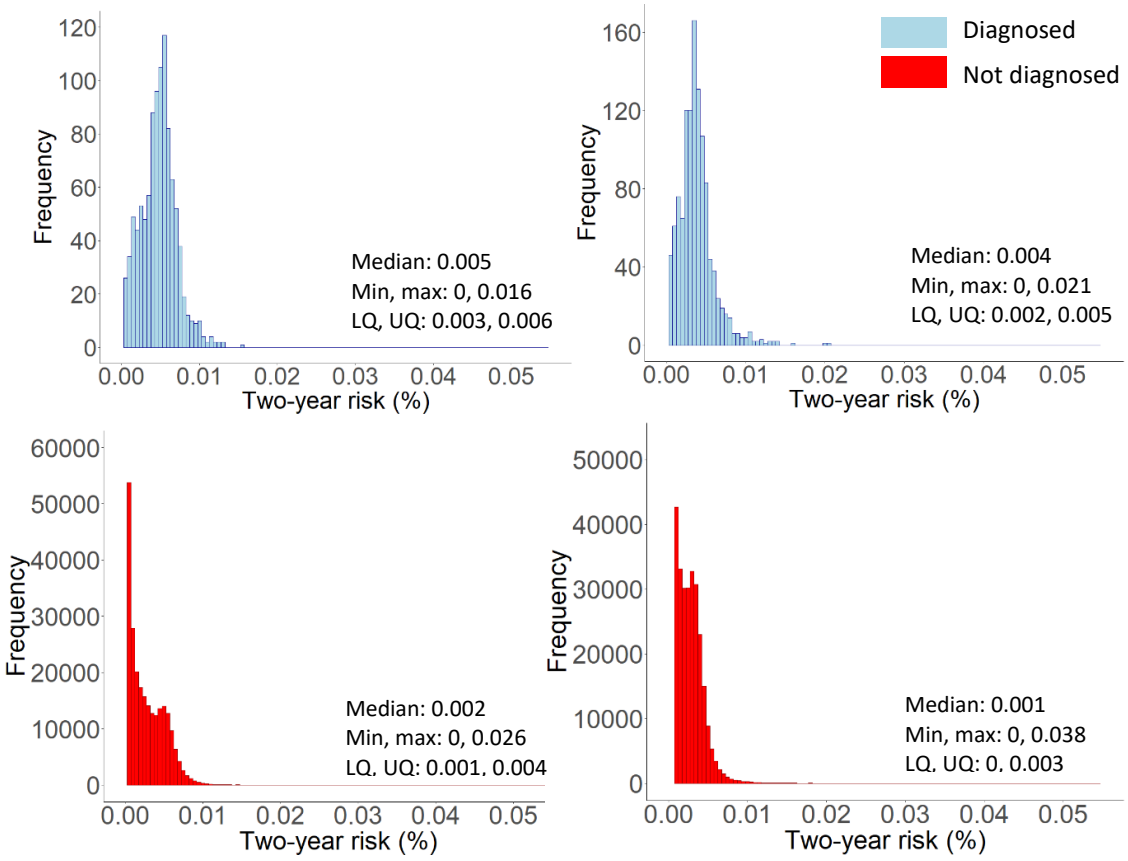
[1]Percentages are out of total number of FBCs in that group.
[2]Time (years) between the first and last FBC in the five-year longitudinal period.
[3]Follow-up per patient is the time (years) from first FBC to diagnosis/censor.
Abbreviations: FBC=full blood count; Hb=haemoglobin; MCV=mean corpuscular volume.

5.5.2 Summary of predictions from the joint model and ColonFlag algorithm

Dynamic risk predictions from the joint models are described in Figure 5.7 and the ColonFlag score in Figure 5.8. On average, patients diagnosed had higher risks than patients not diagnosed from the joint models (e.g. median: 0.005 diagnosed vs. 0.002 not diagnosed males and 0.004

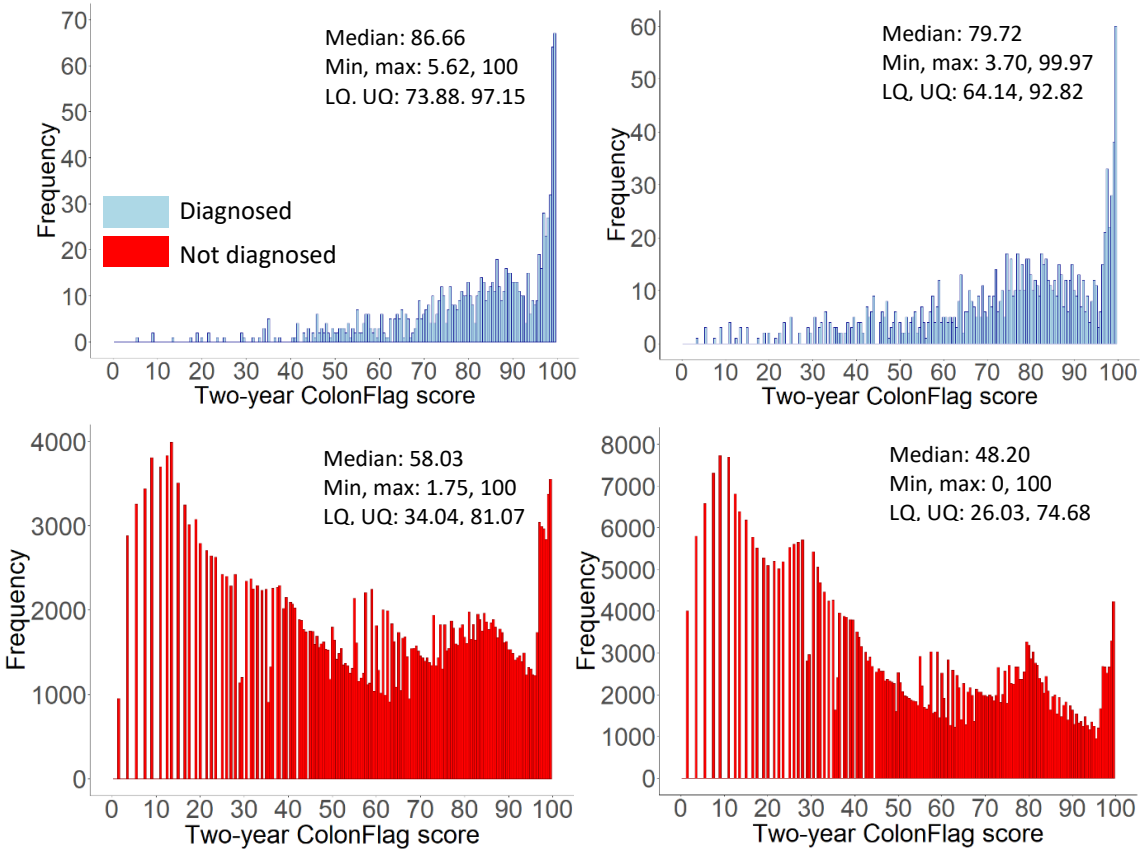
diagnosed vs. 0.001 not diagnosed females). Diagnosed patients also had higher ColonFlag scores (median: 86.66 diagnosed vs. 58.03 not diagnosed males and 79.72 diagnosed vs. 48.20 not diagnosed females). Predictions from the ColonFlag algorithm spanned a greater set of values than predictions from the joint models. This is likely because of the nature of dynamic predictions from joint models (details in Section 5.2.4) and because the joint models provide a measure of absolute risk from a Cox sub-model and therefore depend on the observed event rate, whereas the ColonFlag is a monotone score so does not reflect risk/probability of diagnosis.

Figure 5.7: Distribution of dynamic two-year risk predictions from the joint models for males (left) and females (right) in the split sample internal validation cohort



Abbreviations: LQ=lower quartile; UQ=upper quartile.

Figure 5.8: Distribution of ColonFlag two-year score for males (left) and females (right) in the split sample internal validation cohort



Abbreviations: LQ=lower quartile; UQ=upper quartile.

5.5.3 Performance of the joint models and ColonFlag

Performance measures for the joint models are in Table 5.10. Performance was similar to the development cohort. For example, the Brier score, where a lower value is better, was 0.0038 (development) and 0.0036 (split sample internal validation) for males. This was expected because the average optimism, identified using bootstrapping, was low in the development cohort (Table 5.6).

Table 5.10: Performance measures of the joint models and ColonFlag (split sample internal validation)

Performance measure	Males		Females	
	Joint model	**ColonFlag[1]**	**Joint model**	**ColonFlag[1]**
Brier score	0.0036		0.0027	
R_D^2	0.327		0.337	
C-statistic (95% CI)	0.767 (0.755-0.779)	0.776 (0.764-0.788)	0.764 (0.752-0.776)	0.761 (0.750-0.773)
D-statistic (95% CI)	1.428 (1.327-1.528)		1.461 (1.368-1.553)	
Calibration slope	1.052		1.025	

[1]The ColonFlag provides a monotone score and not a measure of absolute risk. Therefore, many performance measures could not be assessed for this machine-learning algorithm.

The c-statistic (for discrimination) was 0.767 for males and 0.764 for females, suggesting the models can adequately separate patients with and without a diagnosis using only age and earlier FBC data up to two years beforehand. The c-statistic was similar to the ColonFlag algorithm (0.776 males, 0.761 females). Additionally, ROC curves for the joint models and ColonFlag were superimposed, indicating similar diagnostic ability for two-year risk (Figure 5.9). Furthermore, diagnostic test accuracy measures, including sensitivity and specificity, are in Table 5.11 for thresholds corresponding to risk percentiles of 75% and above and in Appendix D6 for every 1% risk percentile.

Figure 5.9: ROC curves for two-year risk from the joint model and ColonFlag for males (left) and females (right) in the split sample internal validation cohort

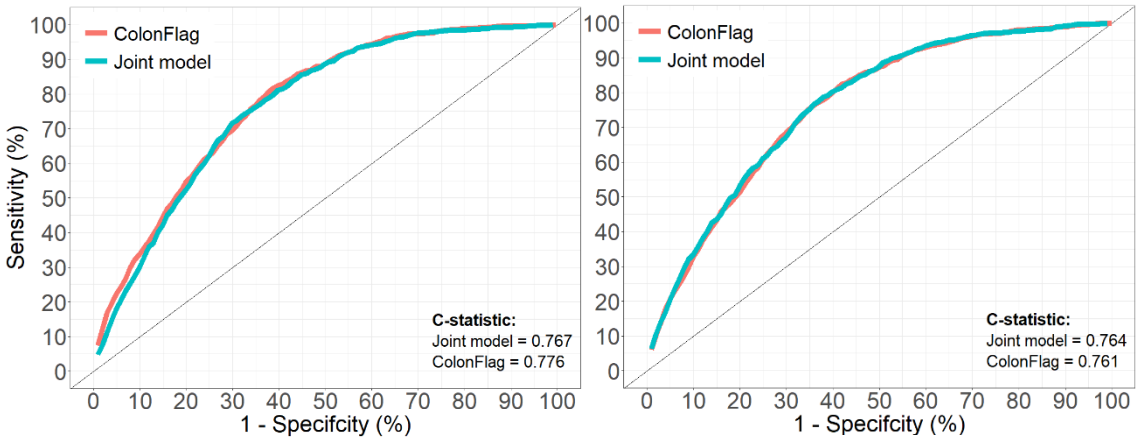


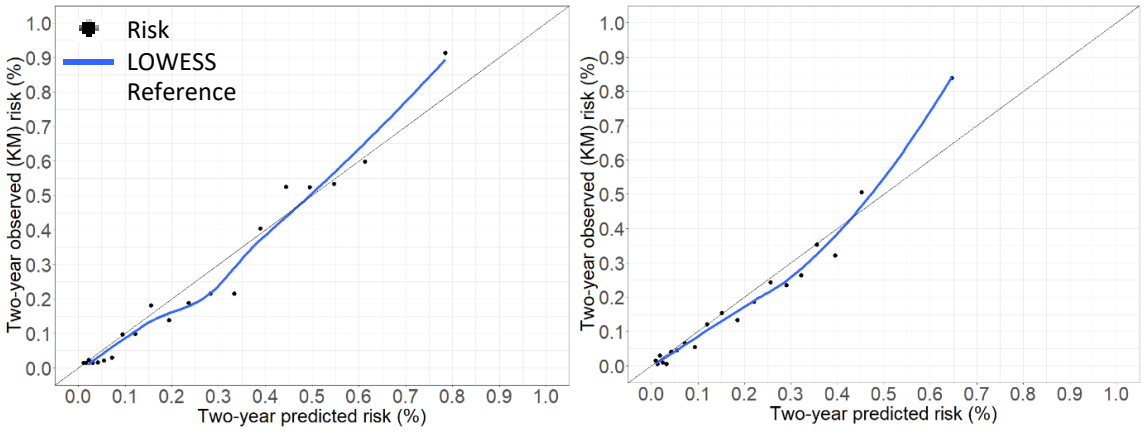
Table 5.11: Diagnostic test accuracy measures of the joint models for risk percentiles of 75% and above (split sample internal validation)

Risk centile	Risk (%) cut-off	True positives	False positives	True negatives	False negatives	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Males:									
75%	0.4175%	644	70493	213013	396	61.92	75.14	0.91	99.81
80%	0.4701%	542	56368	227138	498	52.12	80.12	0.95	99.78
85%	0.5203%	438	42244	241262	602	42.12	85.10	1.03	99.75
90%	0.5756%	312	28143	255363	728	30.00	90.07	1.10	99.72
95%	0.6583%	186	14042	269464	854	17.88	95.05	1.31	99.68
99%	0.8590%	50	2796	280710	990	4.81	99.01	1.76	99.65
Females:									
75%	0.31%	732	106505	321243	468	61.00	75.10	0.68	99.85
80%	0.34%	637	85153	342595	563	53.08	80.09	0.74	99.84
85%	0.37%	521	63822	363926	679	43.42	85.08	0.81	99.81
90%	0.42%	400	42495	385253	800	33.33	90.07	0.93	99.79
95%	0.49%	242	21206	406542	958	20.17	95.04	1.13	99.76
99%	0.72%	76	4214	423534	1124	6.33	99.01	1.77	99.74

Abbreviations: PPV=positive predictive value; NPV=negative predictive value

Calibration plots comparing observed and predicted two-year risk from the joint models are in Figure 5.10. Calibration was good in the internal validation cohort, as the predictions hover over the reference (y=x) line, which represents perfect calibration. The curves diverge slightly from the reference line at the highest predicted risk, but this is likely because there is only one risk group in that region. Calibration can only be assessed for models that provide absolute risks, hence could not be assessed for the ColonFlag because this machine-learning algorithm provides a monotone score, not an absolute risk score. This is demonstrated in Figure D9 (Appendix D6), where the predicted ColonFlag scores do not coincide with the observed risks.

Figure 5.10: Calibration plot for the joint models in males (left) and females (right) in the split sample internal validation cohort



Abbreviations: KM=Kaplan-Meier

5.5.4 Performance of the joint models and ColonFlag by age group

A summary of FBC and follow-up data is provided by age group in Table 5.12. The median number of FBCs and follow-up time and amount of missing data was similar between patients with and without a diagnosis.

Table 5.12: Summary of FBC data and follow-up by age (at baseline) group (split sample internal validation)

	Males		Females	
	Diagnosed	**Not diagnosed**	**Diagnosed**	**Not diagnosed**
No. (%) patients per group				
Age 40-49 years	*22 (0.1%)*	*74,097 (99.9%)*	*39 (0.1%)*	*115,994 (99.9%)*
Age 50-59 years	*115 (0.2%)*	*74,631 (99.8%)*	*104 (0.1%)*	*99,093 (99.9%)*
Age 60-69 years	*257 (0.4%)*	*66,987 (99.6%)*	*239 (0.3%)*	*333,046 (99.7%)*
Age 70-79 years	*401 (0.9%)*	*42,676 (99.1%)*	*425 (0.6%)*	*65,852 (99.4%)*
Age 80-89 years	*224 (1.0%)*	*22,182 (99.0%)*	*331 (0.6%)*	*50,848 (99.4%)*
Age 90+ years	*21 (0.7%)*	*2,933 (99.3%)*	*62 (0.5%)*	*11,917 (99.5%)*
No. FBCs in total				
Age 40-49 years	*106*	*146171*	*98*	*295924*
Age 50-59 years	*388*	*207186*	*334*	*331387*
Age 60-69 years	*1086*	*239377*	*1270*	*333046*
Age 70-79 years	*1759*	*190923*	*2617*	*317750*
Age 80-89 years	*1056*	*109790*	*1756*	*253864*
Age 90+ years	*118*	*13914*	*295*	*52279*
Median no. FBCs (range)				
Age 40-49 years	*4 (1-21)*	*1 (1-141)*	*2 (1-8)*	*2 (1-134)*
Age 50-59 years	*2 (1-18)*	*2 (1-131)*	*2 (1-20)*	*2 (1-119)*
Age 60-69 years	*3 (1-58)*	*3 (1-123)*	*5 (1-70)*	*4 (1-189)*
Age 70-79 years	*3 (1-58)*	*3 (1-110)*	*4 (1-53)*	*4 (1-219)*
Age 80-89 years	*3 (1-48)*	*3 (1-84)*	*3 (1-52)*	*4 (1-144)*
Age 90+ years	*4 (1-22)*	*3(1-78)*	*4 (1-31)*	*3 (1-87)*
No. (%) missing Hb[1]				
Age 40-49 years	*1 (0.94%)*	*577 (0.39%)*	*2 (2.04%)*	*1186 (0.40%)*

Age 50-59 years	*2 (0.52%)*	*965 (0.47%)*	*1 (0.30%)*	*1483 (0.45%)*
Age 60-69 years	*8 (0.74%)*	*1152 (0.48%)*	*17 (1.34%)*	*1532 (0.46%)*
Age 70-79 years	*14 (0.80%)*	*976 (0.51%)*	*19 (0.73%)*	*1604 (0.50%)*
Age 80-89 years	*7 (0.66%)*	*687 (0.63%)*	*12 (0.68%)*	*1593 (0.63%)*
Age 90+ years	*1 (0.85%)*	*94 (0.68%)*	*0 (0.00%)*	*390 (0.75%)*
No. (%) missing MCV[1]				
Age 40-49 years	*2 (1.89%)*	*1767 (1.21%)*	*1 (1.02%)*	*4012 (1.36%)*
Age 50-59 years	*1 (0.26%)*	*3028 (1.46%)*	*8 (2.40%)*	*5020 (1.51%)*
Age 60-69 years	*26 (2.39%)*	*3798 (1.59%)*	*31 (2.44%)*	*4739 (1.42%)*
Age 70-79 years	*35 (1.99%)*	*3696 (1.94%)*	*120 (4.59%)*	*5948 (1.87%)*
Age 80-89 years	*28 (2.65%)*	*2681 (2.44%)*	*45 (2.56%)*	*5563 (2.19%)*
Age 90+ years	*12 (10.17%)*	*301 (2.16%)*	*4 (1.36%)*	*1073 (2.05%)*
No. (%) missing platelets[1]				
Age 40-49 years	*2 (1.89%)*	*2026 (1.39%)*	*3 (3.06%)*	*5588 (1.89%)*
Age 50-59 years	*10 (2.58%)*	*3713 (1.79%)*	*15 (4.49%)*	*6352 (1.92%)*
Age 60-69 years	*27 (2.49%)*	*4446 (1.86%)*	*51 (4.02%)*	*5921 (1.78%)*
Age 70-79 years	*37 (2.10%)*	*4135 (2.17%)*	*105 (4.01%)*	*6896 (2.17%)*
Age 80-89 years	*32 (3.03%)*	*2953 (2.69%)*	*52 (2.96%)*	*6430 (2.53%)*
Age 90+ years	*11 (9.32%)*	*349 (2.51%)*	*3 (1.02%)*	*1404 (2.69%)*
Median time[2] (range)				
Age 40-49 years	*1.0 (0-4.6)*	*0 (0-5.0)*	*0.8 (0-4.8)*	*0.5 (0-5.0)*
Age 50-59 years	*1.5 (0-4.9)*	*1.2 (0-5.0)*	*1.0 (0-4.9)*	*2.1 (0-5.0)*
Age 60-69 years	*2.5 (0-5.0)*	*2.3 (0-5.0)*	*2.0 (0-5.0)*	*2.6 (0-5.0)*
Age 70-79 years	*2.4 (0-5.0)*	*3.0 (0-5.0)*	*2.7 (0-5.0)*	*3.1 (0-5.0)*
Age 80-89 years	*2.4 (0-5.0)*	*3.1 (0-5.0)*	*2.8 (0-5.0)*	*3.1 (0-5.0)*
Age 90+ years	*2.9 (0-5.0)*	*2.8 (0-5.0)*	*2.4 (0-5.0)*	*2.6 (0-5.0)*
Median follow-up[3] (range)				

Age 40-49 years	*3.2 (1.8-6.5)*	*2.0 (1.8-7.2)*	*2.8 (1.9-6.8)*	*2.5 (1.8-7.2)*
Age 50-59 years	*3.5 (1.8-6.8)*	*3.2 (1.8-7.2)*	*3.1 (1.8-7.0)*	*4.1 (1.8-7.2)*
Age 60-69 years	*4.4 (1.8-7.2)*	*4.3 (1.8-7.2)*	*3.8 (1.8-7.1)*	*4.6 (1.8-7.2)*
Age 70-79 years	*4.3 (1.8-7.1)*	*5.0 (1.8-7.2)*	*4.7 (1.8-7.2)*	*5.1 (1.8-7.2)*
Age 80-89 years	*4.5 (1.8-7.2)*	*5.0 (1.8-7.2)*	*4.7 (1.8-7.0)*	*5.1 (1.8-7.2)*
Age 90+ years	*4.8 (1.8-6.9)*	*4.7 (1.8-7.2)*	*4.3 (1.8-7.0)*	*4.6 (1.8-7.2)*

[1]Percentages are out of total number of FBCs in that age group.
[2]Time (years) between the first and last FBC in the five-year longitudinal period.
[3]Follow-up per patient is the time (years) from first FBC to diagnosis/censor.
Abbreviations: FBC=full blood count; Hb=haemoglobin; MCV=mean corpuscular volume.

The c-statistic by age group is in Figure 5.11 for males and Figure 5.12 for females. The c-statistic for the joint model was similar for all age groups in females, except lower in the 40-50 age group, which had the least number of events and therefore widest confidence intervals. However, for males, the c-statistic was more variable by age group. The c-statistic was similar between the joint models and ColonFlag algorithm for all age groups, suggesting discriminative performance between the models is balanced, although slightly higher for the ColonFlag for most age groups.

Figure 5.11: C-statistic for the joint model and ColonFlag algorithm by age (at baseline) group in males (split sample internal validation cohort)

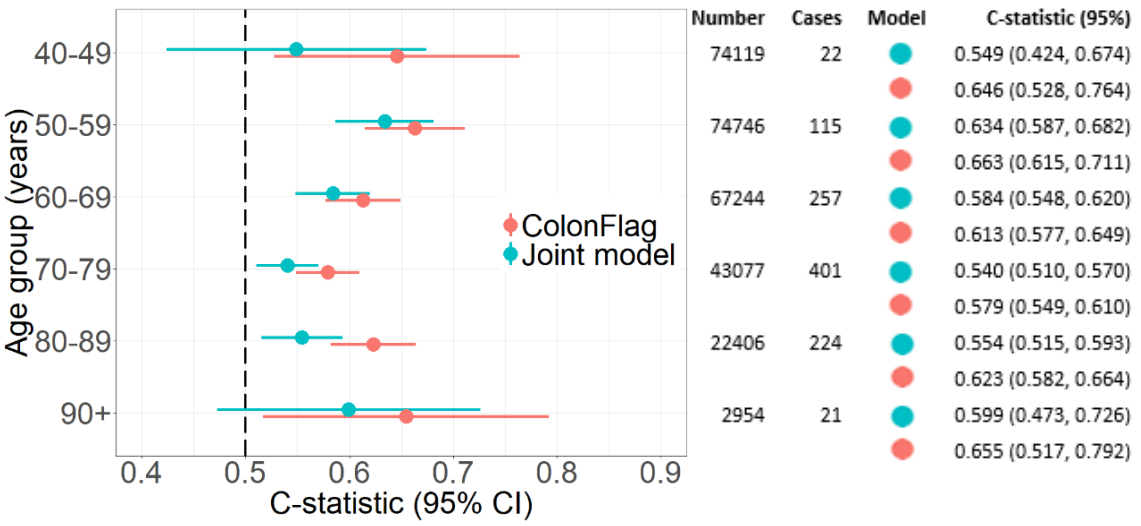
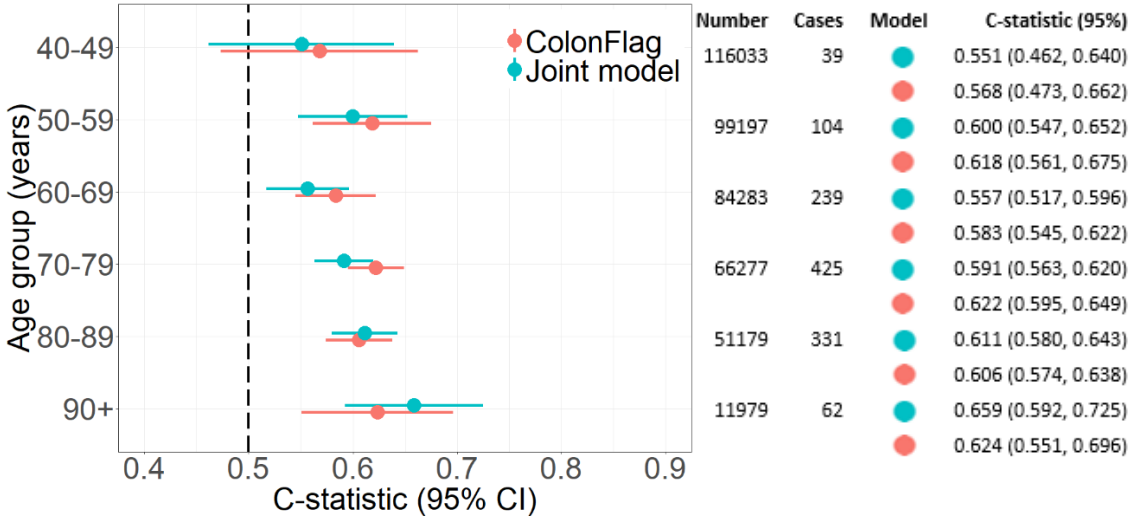
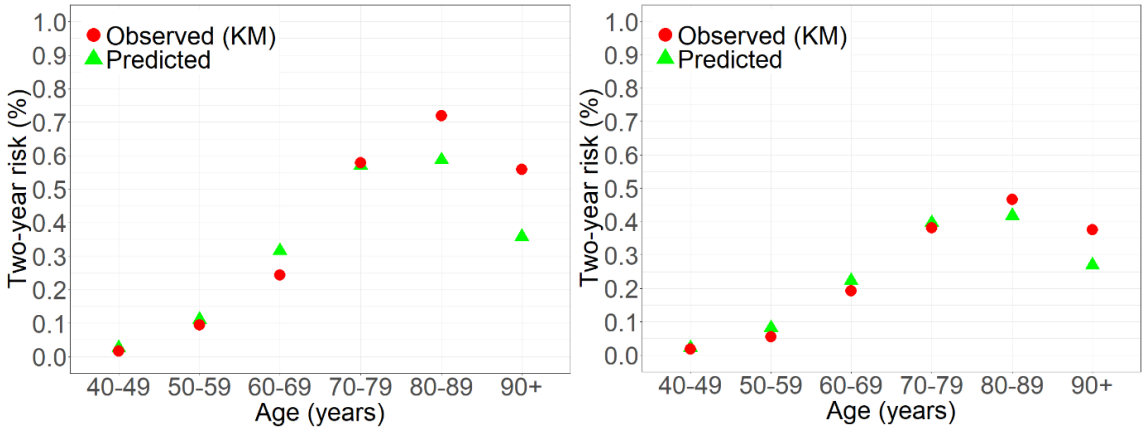


Figure 5.12: C-statistic for the joint model and ColonFlag algorithm by age (at baseline) group in females (split sample internal validation cohort)



Calibration plots by age group are in Figure 5.13 for the joint models. The models were well calibrated in all age groups, although they under-predicted risk of diagnosis in males aged 90+ years at baseline FBC, which may be due to the small number of events in the subgroup (21 diagnosed, Table 5.12) and other health reasons.

Figure 5.13: Calibration plot for the joint models by age (at baseline) group in males (left) and females (right) in the split sample internal validation cohort



Abbreviations: KM=Kaplan-Meier

5.5.5 Performance of the joint models and ColonFlag by number of FBCs

A summary of FBC and follow-up data is provided by number of FBCs in Table 5.13. Patients diagnosed were on average approximately 10 years older at baseline FBC than patients not diagnosed, as expected. The median duration of the five-year longitudinal period, follow-up time, and amount of missing data was similar between patients with and without a diagnosis.

Table 5.13: Summary of patient characteristics, FBC data, and follow-up by number of FBCs (split sample internal validation)

	Males		**Females**	
	Diagnosed	**Not diagnosed**	**Diagnosed**	**Not diagnosed**
No. (%) patients per group				
2	*189 (0.3%)*	*61,011 (99.7%)*	*190 (0.2%)*	*89,995 (99.8%)*
4	*127 (0.5%)*	*25,300 (99.5%)*	*129 (0.3%)*	*44,852 (99.7%)*
6	*64 (0.5%)*	*11,907 (99.5%)*	*82 (0.4%)*	*22,033 (99.6%)*
8	*38 (0.8%)*	*5,061 (99.3%)*	*39 (0.4%)*	*9,791 (99.6%)*
10	*26 (1.0%)*	*2,465 (99.0%)*	*21 (0.5%)*	*4,655 (99.5%)*
12	*13 (1.1%)*	*1,203 (98.9%)*	*12 (0.5%)*	*2,239 (99.5%)*
14	*4 (0.6%)*	*665 (99.4%)*	*4 (0.3%)*	*1,248 (99.7%)*
Mean age[1] (SD)				
2	*70.0 (10.0)*	*58.6 (12.0)*	*71.6 (12.6)*	*59.3 (13.9)*
4	*73.9 (9.8)*	*63.7 (12.1)*	*74.7 (10.0)*	*63.5 (14.1)*
6	*72.6 (8.1)*	*67.3 (11.8)*	*76.4 (11.9)*	*66.9 (13.9)*
8	*72.7 (10.8)*	*69.1 (11.6)*	*76.3 (9.3)*	*68.7 (13.9)*
10	*72.9 (7.7)*	*70.4 (11.7)*	*77.6 (8.8)*	*70.3 (13.7)*
12	*77.0 (8.9)*	*70.7 (12.2)*	*76.5 (10.2)*	*70.9 (13.6)*
14	*82.3 (4.4)*	*71.5 (12.0)*	*75.0 (2.9)*	*70.9 (13.5)*
No. FBCs in total				

2	*378*	*122022*	*380*	*179990*
4	*508*	*101200*	*516*	*179408*
6	*384*	*71442*	*492*	*132198*
8	*304*	*40488*	*312*	*78328*
10	*260*	*24650*	*210*	*46550*
12	*156*	*14436*	*144*	*26868*
14	*56*	*9310*	*56*	*17472*
No. (%) missing Hb[2]				
2	*1 (0.26%)*	*479 (0.39%)*	*3 (0.79%)*	*724 (0.40%)*
4	*4 (0.79%)*	*559 (0.55%)*	*12 (2.33%)*	*909 (0.51%)*
6	*6 (1.56%)*	*453 (0.63%)*	*1 (0.20%)*	*657 (0.50%)*
8	*0 (0.00%)*	*215 (0.53%)*	*0 (0.00%)*	*423 (0.54%)*
10	*2 (0.77%)*	*137 (0.56%)*	*1 (0.48%)*	*314 (0.67%)*
12	*0 (0.00%)*	*137 (0.95%)*	*0 (0.00%)*	*180 (0.67%)*
14	*0 (0.00%)*	*60 (0.64%)*	*0 (0.00%)*	*78 (0.45%)*
No. (%) missing MCV[2]				
2	*7 (1.85%)*	*1379 (1.13%)*	*6 (1.58%)*	*2127 (1.18%)*
4	*8 (1.57%)*	*1585 (1.57%)*	*10 (1.94%)*	*2692 (1.50%)*
6	*3 (0.78%)*	*1241 (1.74%)*	*12 (2.44%)*	*2249 (1.70%)*
8	*6 (1.97%)*	*905 (2.24%)*	*6 (1.92%)*	*1632 (2.08%)*
10	*8 (3.08%)*	*675 (2.74%)*	*11 (5.24%)*	*1166 (2.50%)*
12	*8 (5.13%)*	*445 (3.08%)*	*5 (3.47%)*	*677 (2.52%)*
14	*1 (1.79%)*	*359 (3.86%)*	*1 (1.79%)*	*518 (2.96%)*
No. (%) missing platelets[2]				
2	*6 (1.59%)*	*1951 (1.60%)*	*10 (2.63%)*	*3075 (1.71%)*
4	*10 (1.97%)*	*2115 (2.09%)*	*23 (4.46%)*	*3980 (2.22%)*
6	*17 (4.43%)*	*1684 (2.36%)*	*10 (2.03%)*	*3180 (2.41%)*

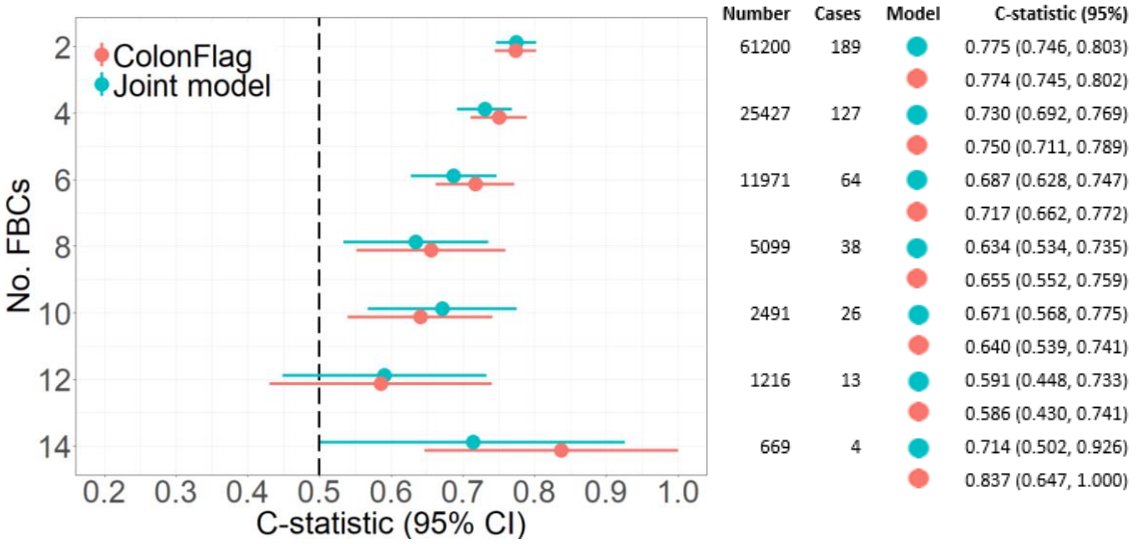
8	*8 (2.63%)*	*1143 (2.82%)*	*19 (6.09%)*	*2064 (2.64%)*
10	*2 (0.77%)*	*700 (2.84%)*	*1 (0.48%)*	*1387 (2.98%)*
12	*16 (10.26%)*	*464 (3.21%)*	*14 (9.72%)*	*729 (2.71%)*
14	*1 (1.79%)*	*323 (3.47%)*	*3 (5.36%)*	*560 (3.21%)*
Median time[3] (range)				
2	*1.3 (0.0, 4.9)*	*1.6 (0.0, 5.0)*	*1.1 (0.0, 5.0)*	*1.6 (0.0, 5.0)*
4	*3.1 (0.2, 5.0)*	*3.4 (0.0, 5.0)*	*2.8 (0.3, 5.0)*	*3.5 (0.0, 5.0)*
6	*3.7 (0.4, 5.0)*	*4.2 (0.1, 5.0)*	*3.6 (0.7, 5.0)*	*4.2 (0.0, 5.0)*
8	*4.3 (1.2, 5.0)*	*4.4 (0.3, 5.0)*	*4.3 (1.8, 4.9)*	*4.4 (0.3, 5.0)*
10	*4.4 (1.5, 5.0)*	*4.5 (0.2, 5.0)*	*4.6 (2.1, 5.0)*	*4.5 (0.2, 5.0)*
12	*4.2 (2.9, 4.9)*	*4.6 (0.8, 5.0)*	*4.3 (2.5, 4.9)*	*4.5 (0.5, 5.0)*
14	*4.7 (3.7, 4.9)*	*4.6 (1.0, 5.0)*	*4.7 (3.9, 4.9)*	*4.6 (0.8, 5.0)*
Median follow-up[4] (range)				
2	*3.3 (1.8, 7.1)*	*3.6 (1.8, 7.2)*	*3.2 (1.8, 7.2)*	*3.6 (1.8, 7.2)*
4	*5.1 (2.1, 7.0)*	*5.4 (1.9, 7.2)*	*4.8 (2.1, 7.1)*	*5.5 (1.8, 7.2)*
6	*5.6 (2.3, 7.1)*	*6.2 (2.0, 7.2)*	*5.5 (2.4, 7.2)*	*6.1 (1.9, 7.2)*
8	*6.3 (2.9, 7.0)*	*6.4 (2.2, 7.2)*	*6.3 (3.6, 6.9)*	*6.4 (2.2, 7.2)*
10	*6.2 (3.6, 7.0)*	*6.5 (2.2, 7.2)*	*6.7 (3.9, 7.1)*	*6.5 (2.2, 7.2)*
12	*6.3 (4.7, 6.9)*	*6.5 (2.6, 7.2)*	*6.3 (4.4, 7.0)*	*6.5 (2.2, 7.2)*
14	*6.5 (5.6, 6.8)*	*6.6 (2.8, 7.2)*	*6.5 (5.8, 7.1)*	*6.6 (2.6, 7.2)*

[1]Age (years) at baseline FBC (most recent FBC available prior to two years before diagnosis/censor).
[2]Percentages are out of total number of FBCs in that age group.
[3]Time (years) between the first and last FBC in the five-year longitudinal period.
[4]Follow-up per patient is the time (years) from first FBC in the longitudinal period to diagnosis/censor.
Abbreviations: FBC=full blood count; Hb=haemoglobin; MCV=mean corpuscular volume.

The c-statistic by number of FBCs is in Figure 5.14 for males and Figure 5.15 for females. The c-statistic was similar between the joint models and ColonFlag algorithm regardless of the number of FBCs used to identify risk, suggesting discriminative performance between the models is balanced. Surprisingly, the c-statistic declined as the number of tests increased. This may be

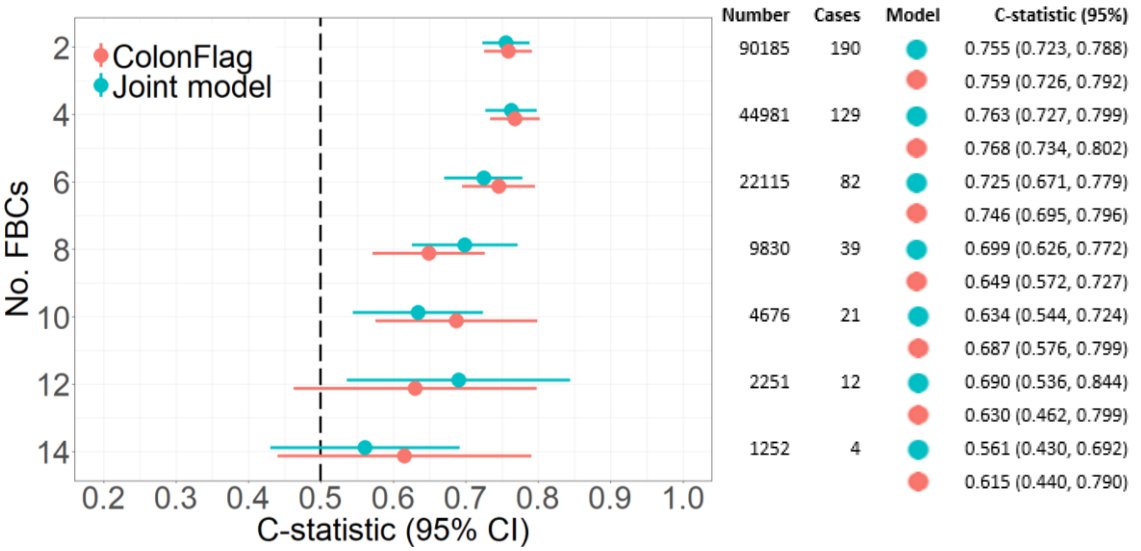
because FBCs are more common in older age groups (Table 5.2), where model performance was not the best (Section 5.5.4). Additionally, the sample size was small among patients with higher numbers of FBCs available, producing wide confidence intervals. C-statistics by number of FBCs should be derived for each age group separately to eliminate the effect of age on number of FBCs. As an example, these are provided for males aged 70-89 years at baseline FBC, as this age group had the largest sample size and number of events (Figure 5.16). In this age group, the c-statistic generally increased as the number of FBCs increased, as expected. However, the small sample size for each group produced wide confidence intervals.

Figure 5.14: C-statistic for the joint model and ColonFlag algorithm by number of FBCs in males (split sample internal validation cohort)



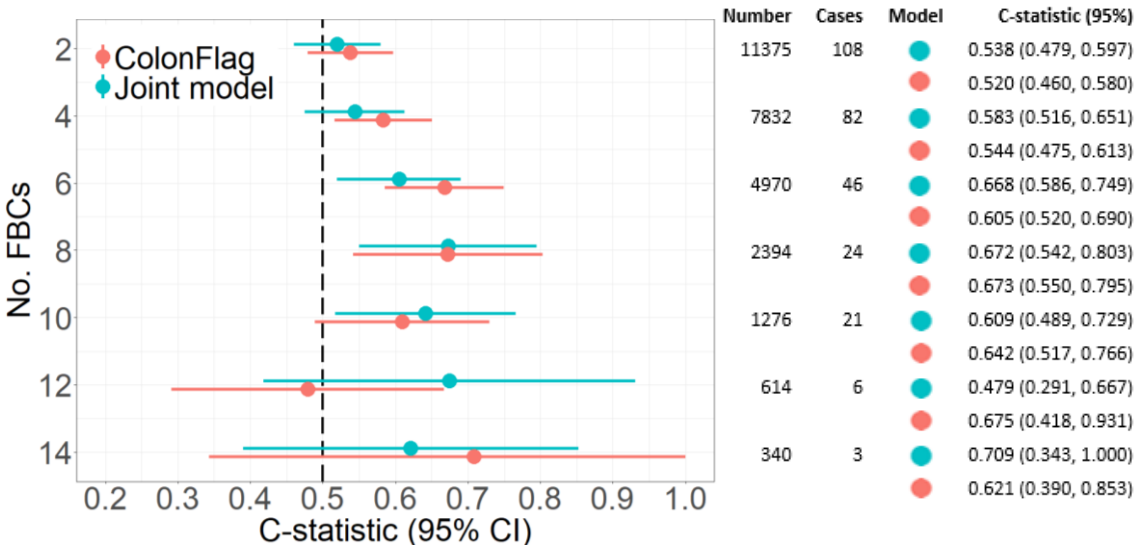
Abbreviations: FBC=full blood count.

Figure 5.15: C-statistic for the joint model and ColonFlag algorithm by number of FBCs in females (split sample internal validation cohort)



Abbreviations: FBC=full blood count.

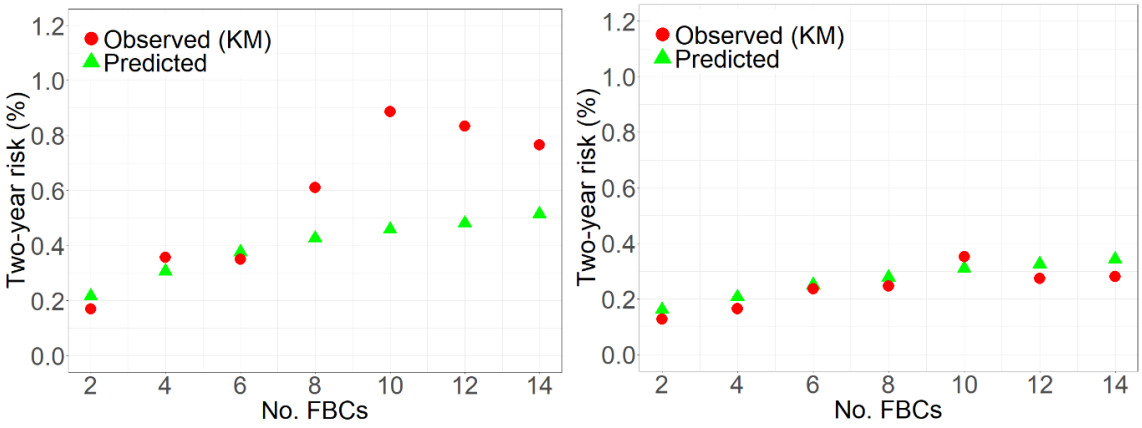
Figure 5.16: C-statistic for the joint model and ColonFlag algorithm by number of FBCs in females aged 70-89 years at baseline FBC (split sample internal validation cohort)



Abbreviations: FBC=full blood count.

Calibration plots by number of FBCs are in Figure 5.17 for the joint models. The models were well calibrated regardless of the number of FBCs used to identify two-year risk of diagnosis, except they under-predicted risk for males with ≥ 8 FBCs, which was not observed in females. As a check, I assessed calibration in the development cohort, where this under-prediction reduced heavily (Figure D10), suggesting this under-prediction may be specific to this internal validation cohort. This will need re-assessing in future external validation cohorts.

Figure 5.17: Calibration plot for the joint models by number of FBCs in males (left) and females (right) in the split sample internal validation cohort



Abbreviations: KM=Kaplan-Meier; FBC=full blood count.

5.6 Discussion

5.6.1 Summary of main findings

Many FBC parameters change over time due to bowel cancer. The opportunity to use these relevant changes to detect bowel cancer is missed in practice because blood levels often remain in the normal reference range (so are not flagged as abnormal) and clinicians only assess the most recent FBC. I utilised these relevant changes in a dynamic clinical risk prediction model that harnesses repeated FBC measures to identify risk of diagnosis in two years for males and females separately. A (patient-level) decline in haemoglobin and mean corpuscular volume and rise in platelet count from the average population trajectory increased the risk of diagnosis in two

years, as expected based on trajectories identified in Chapter 4. Both multivariate models have good predictive performance and are well calibrated using only age and trajectories in haemoglobin, mean corpuscular volume, and platelet count up to two years before diagnosis. Optimism and overfitting, assessed using bootstrapping, were very low so further refinements to the model were not needed, e.g. shrinkage of coefficients to reduce overfitting. This is likely due to the large sample size used and small number of variables included in the models.

To ensure a robust internal validation, I also assessed performance of the joint models in a split sample internal validation cohort. Performance of the joint models showed comparable results to the development cohort. Performance of the joint models was good in the overall cohort and could improve further with additional risk factors in the model, which will be considered as future work. For example, the NPV for the joint models were above 99.7% for almost all thresholds, suggesting the models have high performance in identifying low-risk patients and could therefore help avoid unnecessary referrals in practice. This would ultimately reduce burden on healthcare services, such as staff, time, and cost. Performance of the joint models was good for almost all subgroups of age at baseline FBC and by number of FBCs used to identify risk. For example, calibration was good, suggesting the models are accurate, although under-predicted risk in males aged 90+ years or males with 8 or more FBCs, which may be consequence of small sample sizes and number of events and other health-related considerations in those subgroups, although this under-prediction was not observed in the development data. Additionally, lower c-statistics may have been derived because these are known to be sensitive to the sample size[239, 240]. Discrimination and calibration were good for all age groups, although was more variable for men than women, suggesting the models could work well in practice. Performance was best for younger adults, where diagnoses are less commonly made. Performance statistics in subgroups may improve with larger samples sizes and the addition of further risk factors for bowel cancer, which will be considered as future work. Furthermore, the

c-statistic reduced as number of tests increased, although the opposite was expected. This might be because FBCs were more commonly available in older patients, in whom performance was not as high (compared to younger patients). This subgroup analysis may need stratifying by age in larger samples.

I also compared the joint models and ColonFlag score. Both models performed very similarly for two-year risk, both overall (joint models 0.767 males, 0.764 females; ColonFlag 0.776 males, 0.761 females) and in relevant subgroups. This was expected because the joint models and ColonFlag algorithm ultimately use the same data (age, sex, changes over time in FBC levels) to identify risk. Discrimination was only slightly better for the ColonFlag, which may be because it uses up to all 20 FBC parameters to identify risk[64], whereas my joint models use only haemoglobin, mean corpuscular volume, and platelets. However, discrimination remained very similar regardless, suggesting these additional parameters may not be needed. Despite similar performance, my joint models can be considered an improvement on the ColonFlag because they use transparent methods, are easier to interpret and embed into clinical practice, and provide a measure of absolute risk.

5.6.2 Comparison with existing literature

In this study, 0.4% males and 0.3% females were diagnosed with bowel cancer two years after their baseline FBC (latest FBC prior to the two-year exclusion period before diagnosis/study exit). This two-year event rate is similar to those reported in other UK primary care studies, the QCancer Colorectal prediction model[69] and the ColonFlag external validation[65]. Additionally, the low event rate identified in this study was expected, consistent with national statistics for England reported by Cancer Research UK[241].

Many prediction models for bowel (or colorectal) cancer exist. I identified 13 models in my systematic review that use some FBC data to identify risk of bowel cancer. All but one (ColonFlag by Kinar 2016[64]) are static models, meaning they use data from one baseline time-point. This includes the most commonly used bowel cancer prediction model in the UK, the QCancer Colorectal model by Hippisley-Cox 2013[69]. QCancer Colorectal relies on symptoms to identify two-year risk of diagnosis, but recent studies identified that symptoms are commonly reported close to the time of diagnosis[13, 112], suggesting the model may not perform well for early detection. My joint models are an improvement on these static models, including QCancer Colorectal, because they use repeated measures from a single patient to provide a more individualised risk prediction. Additionally, compared to the QCancer Colorectal model, my joint models rely only on earlier data available up to two years before diagnosis and have good performance, suggesting they could facilitate earlier detection. Predictive performance between my joint models and the QCancer Colorectal model on the same patient cohort is yet to be formally compared.

The one, non-static model is the ColonFlag (machine-learning algorithm) by Kinar 2016[64]. This model can be considered dynamic because it uses repeated FBC measures to identify risk. Predictive performance for two-year risk in UK patients is similar between my joint models (c-statistic: ~0.76 for both models) and the ColonFlag model (c-statistic: 0.78), based on an existing external validation study[65]. This existing external validation study also reported a ROC curve for 18-24-month risk from the ColonFlag model, to which the ROC curves I present here are very comparable. My joint models use methodology that perform less parameter estimation than machine-learning, which requires much larger sample sizes and more computational capacity[236], and employs statistical methods, which are more accepted in healthcare, easier to understand, and therefore easier to embed into practice. Additionally, my joint models provide a measure of absolute risk, whereas the ColonFlag provides a monotone score between 0 and

100, which is not an absolute risk score because no functional form (or equation) to predict risk was derived. The ColonFlag score was derived using only FBC data and subsequently assessed for diagnosis.

The ColonFlag algorithm has been externally validated many times[64, 65, 87, 98, 101, 105]. Two studies relied on a small number of events (n=21 for an unspecified risk window in Ayling 2019[87] and n=60 for one-year risk in Hilsden 2018[98]), whereas the number of events was larger in this study (n=1,040 in males, n=1,200 in females) and coincided with two-year risk to assess the potential for earlier detection.

Discrimination was assessed in three studies only[64, 65, 101]. Birks 2017 was the only study to assess performance at various times before diagnosis/censor date[65]. I identified a similar c-statistic for two-year risk (0.77 in males, 0.76 in females) as the 18-24 month c-statistic in Birks 2017 (0.78). I also assessed discriminative ability in relevant subgroups of age, which was not previously assessed in any other study.

Kinar 2016[64] was the only study to assess calibration of the ColonFlag algorithm. However, the authors used unorthodox methods, using calibration plots to compare predicted risk in one cohort to observed risk in different cohorts, which does not reflect the accuracy (and therefore calibration) of the model. Traditional measures of calibration cannot be assessed for models that give no absolute risk. Therefore, calibration cannot be assessed for the ColonFlag, a monotone score derived using changes in FBC levels that was subsequently assessed for diagnosis (as the predictor using logistic regression).

5.6.3 Implications for clinical practice

The joint models are designed to provide patients with an up-to-date risk prediction when an FBC is added to their record. For example, the most recent FBC was considered baseline, which mimics practice because the most recent FBC is considered when examining a patient and historical FBCs are often not considered, but these would be considered by the joint models. Additionally, the models are designed to use routinely available data and my supervisors and I envisage they would be programmed into GP/CPRD software to run automatically when a new FBC becomes available. Therefore, there will likely be no additional work for patients or GP staff to identify a patient's risk of diagnosis from my joint models.

Current clinical practice has some limitations, which are barriers to early detection. The opportunity to identify cancer earlier is likely being missed in practice because only a single FBC is often assessed, and FBC reference ranges and national guidelines for referral are sub-optimal for early detection. My models use trajectories among repeated FBC measures that can occur before the onset of microcytic anaemia, which is related to iron-deficiency anaemia and is used as national criteria for referral for further testing[53, 54, 131, 183] and before FBC parameters reach abnormal levels[177-181]. My joint models could therefore improve on current clinical practice, facilitate early detection, and improve survival.

I have a PPI Advisory Group, who met routinely to review the work throughout the DPhil. From their perspective, a blood test provides a simple yet valuable opportunity to refine the group of patients referred for colonoscopy, reduce the clinical problem, and potentially help save many lives. One PPI representative stated, "*This model can really help tackle the problem and wished it was around when I was diagnosed*". The Advisory Group agreed that my new models could improve on existing assessments in clinical practice because they use data from earlier time points that could help with the difficulties of early detection.

Over the last 12 months, FIT testing, which examines stool samples for traces of blood, has proved a useful test outside the screening programme for use in patients with low-risk symptoms. Negative findings on FIT avoids the need for colonoscopy, based on a 98% negative predictive value in a recent primary care study[26]. However, these patients are nonetheless symptomatic, which is problematic for early detection because symptoms often become overt when the cancer is late-stage and diagnoses are often made soon (within six months) of reporting the symptom(s) to GPs [13, 112]. My joint models use only data available earlier than two years before diagnosis, a time where patients are likely asymptomatic[13]. Patients identified as high-risk from my models could be referred for FIT testing, which is much more practical than colonoscopy, such as it is an accurate test (85% sensitivity[17]), much cheaper and less invasive than colonoscopy, and carries less burden on the patient and healthcare workers. However, such negative predictive values in high-risk asymptomatic patients identified from my models would need to be investigated.

5.6.4 Strengths & limitations

Strengths:

A large sample size and duration of follow-up was used to develop the joint models, which exceeded the sample size requirements. The joint models therefore had little-to-no overfitting and performance measures can be considered precise.

Multivariate joint models were developed, which use repeated measures data. This is a key strength and improvement on existing prediction models in the UK, as joint models provide a more individualised risk assessment. The models use changes in three blood levels commonly available in primary care that have little-to-no correlation.

The models were developed using routinely available data and mimic patient flow in practice. For example, the latest FBC was considered baseline, which mimics practice because the most recent FBC is considered when examining a patient and historical FBCs are usually not considered (particularly when the FBC parameters are within their reference ranges).

The models are designed to use routinely available data and therefore would not require any additional work for patients or clinicians. They use all FBCs available in the five-year period, which maximises precision and reflects clinical practice. Testing is more likely to be common in the run up to diagnosis but the two-year exclusion period before diagnosis/study exit ensured the median number of FBCs modelled and total follow-up up was comparable between patients with and without a diagnosis in two years from baseline.

One strength of the split sample internal validation study is that it uses a large sample size to make an overall assessment of the joint models. This increases the precision of estimates and reliability of performance measures. Another strength is that predictive performance of these joint models were compared to the ColonFlag algorithm using the same patient cohort. This eliminates heterogeneity among patient samples to ensure a direct comparison of models. Many models for bowel cancer detection exist, but no studies have compared these to the ColonFlag algorithm, which is receiving much attention in UK general practices, on the same cohort until now.

Limitations:

The joint models rely on routinely available FBCs to identify risk of diagnosis. Their impact is therefore limited to patients who are referred for blood testing by their primary care GP. However, bowel cancer is more common in older age groups, where FBCs are also more

commonly performed, so it is likely the model will have higher impact in the patient group most at risk of diagnosis.

FBC blood tests are ordered for many reasons in primary care, not bowel cancer specifically, but these reasons are not available in CPRD. It is possible that patients without bowel cancer who have many FBCs in the five-year period have another disease or condition that influences blood levels over time. Therefore, some false positives (patients determined to be high risk who are not diagnosed with bowel cancer) may have another illness. Data on other gastrointestinal cancers is available in my NCRAS extract and will be considered as future work.

Existing systematic reviews have identified many risk factors for bowel cancer, including alcohol consumption levels, ethnicity, and family history of bowel cancer[136, 242]. Not all risk factors are available or accurately recorded in CPRD or were available in my dataset so were not included in the model. However, I have previously provided detailed reasons for why age, sex, and FBC parameters are appropriate as the only covariates in the model at this stage (Appendix D1). As future work, further risk factors may be obtained and considered as covariates.

Patients diagnosed with another cancer type before or simultaneously with bowel were excluded. This reduces generalisability of the models to the general population. However, removing such patients was important to eliminate the confounding effect of other cancers on FBC results[187]. Ideally, a competing risks analysis would have been performed to account for other cancer types in the model, which is possible conceptually[209] but programming and statistical packages are not yet developed to make this doable for multivariate joint models in practice. Future work involves expanding the outcome to include other cancers, where this would be explored further.

Bootstrapping is widely favoured for internal validation, as it uses the entire cohort to build the model and iteratively resamples from the cohort to derive validation cohorts[125, 230, 231]. Bootstrapping allows for an appropriate assessment of overfitting and optimistic performance, usually with many bootstrapped samples, i.e. 100+, and adjustments to correct for these. However, I limited to three bootstrap samples for males and females separately due to practicality. Although using fewer bootstrap samples reduces precision of measures of optimism and overfitting, this approach is still favoured to other approaches, such as data splitting, which reduces the sample size used to build the model and provides only one sample to assess optimism so provides a less precise assessment.

Although a large sample and number of events were used for the overall assessment of the joint models in the split sample internal validation, the sample size and number of events were small for some subgroups of age and number of FBCs used to identify risk. Performance of the models in relevant subgroups should therefore be assessed in larger samples to increase precision.

5.6.5 Conclusion and further research

I developed a multivariate joint model for males and females separately. The models use routinely available FBC data earlier than two years before diagnosis to identify relevant trajectories that contribute to two-year risk of bowel cancer. The models perform well and similarly to the ColonFlag algorithm, which is receiving much attention in clinical practice, but my joint models use transparent methods, are easier to interpret and embed into clinical practice, and provide a measure of absolute risk. Performance of the joint models in relevant subgroups was also assessed but were limited to small sample sizes. Further work is therefore required. For example, the models require external validation using patients from primary care practices not involved in the development or internal validation process. Such studies would include other data sources, such as THIN[243], RCPG Research and Surveillance Centre[244], Care

and Health Information Exchange (Hampshire Health Record)[245], and QResearch[246]. Future work includes exploring the addition of further risk factors for bowel cancer and expanding the outcome to include other cancers (using trajectories specific to those cancers). External validation comparing several models for bowel cancer detection in similar populations is also planned. Additionally, I plan to perform decision curve analyses, which are becoming increasingly popular to balance the benefits with drawbacks of intervention in clinical practice due to the prediction model[247-253], explore risk groups from my joint models further, and assessing performance of FIT screening in high-risk asymptomatic patients identified by my models.

6

Discussion

A detailed discussion for each individual research objective has been provided in the corresponding chapters (Chapters 2-5). These discussions include a detailed and comprehensive summary of findings, comparison with other literature, strengths and limitations, and implications for clinical research. In this Chapter, I will provide an overview of the research and thesis, novelties of this work, and further research ideas.

6.1 Summary of the thesis

6.1.1 Motivation

Bowel cancer is common in the UK[1]. Early detection is crucial to improve chances of survival but this remains challenging in practice. Reasons include the late onset, and therefore late reporting, of symptoms to GPs[13] and issues with the bowel cancer screening programme, such as low sensitivity and uptake of some assessments and ethical considerations[44]. Blood testing (the FBC) is common in the UK and provides a means to use potentially relevant data that may facilitate early detection. The ColonFlag machine-learning algorithm reported by Kinar 2016[64] (previously called the MeScore), was developed for the Israeli population and uses changes in the FBC blood test over time and current values to assist bowel cancer detection. This model has good discrimination in UK data[65] but many important details of this commercially developed model remain unknown. For example, the particular FBC parameters that change due

to bowel cancer and how they change was unknown. Ultimately, the FBC test may play a role in early detection but there was limited evidence, which I sought to obtain in this research.

6.1.2 A review of the literature: is the FBC helpful and are existing models adequate?

The first objective was to perform a systematic review of the FBC test for bowel cancer detection (Chapter 2). I identified many relevant FBC parameters for bowel cancer detection: red blood cell count, haemoglobin, mean corpuscular volume, red blood cell distribution width, white blood cell count, and platelets. Studies reported statistically significant differences for years before diagnosis, suggesting the FBC could facilitate earlier detection.

My systematic review identified that existing prediction models relying on FBC data for bowel cancer detection are flawed. For example, all models scored a high risk of bias in the analysis domain of the PROBAST tool. Some models rely on symptoms to predict long-term risk of diagnosis and perform well, but there are possibly likely many early events in the outcome period close to the reporting of symptoms and these may drive the good performance. Therefore, these models likely would not help with early detection but need further independent testing.

6.1.3 New methodology to help assess EHR data quality

The second objective was to derive a methodological approach to assess data quality of EHR datasets (Chapter 3). Recent systematic reviews identified many barriers to performing EHR data quality checks, which include unfamiliarity with performing such assessments[145], highlight these checks to be crucial[146], and describe the need for a generalised approach to do them[147]. I sought to undertake this task and derive such a methodological approach, which would grow as researchers start using it. This research highlighted a number of issues that need to be

considered when using EHR data for research purposes. I emphasise this further because EHRs were not primarily designed for use in research.

This objective served two purposes. Firstly, to help new researchers in the field of big data to acquire, prepare, and validate EHR data. Secondly, for me to develop a fundamental understanding of the FBC and CPRD database and derive a reliable dataset for analysis in subsequent chapters.

6.1.4 Changes in FBC levels over time are relevant for bowel cancer detection

The third objective was to identify trajectories in the FBC related to presence of bowel cancer (Chapter 4). This research was guided by the systematic review findings (Chapter 2) and used CPRD data derived from the data validation and quality assessment (Chapter 3). I identified trajectories in many FBC parameters that differed between patients with and without a diagnosis. Differences began approximately four years before diagnosis and grew larger in the run up to diagnosis, with levels in patients without a diagnosis changing less rapidly over time.

This review also highlighted a number of barriers for early detection in clinical practice. Firstly, using single FBCs is sub-optimal and repeated measures is more predictive, but this is challenging in practice because clinicians are busy and routinely reviewing repeated blood tests would not be practical. Secondly, normal/reference ranges for the FBC[177-181], which reflect a healthy individual, are sub-optimal because many relevant changes in FBC levels that could help identify cases occur within this reference range. Thirdly, national guidelines for referral for further testing are sub-optimal, as microcytic anaemia, determined from the FBC[53, 54, 131, 183], which may be related to iron-deficiency, is not diagnosed early enough to facilitate early detection. These reasons highlight the need for the prediction model that could make use of relevant trajectories

to identify high-risk patients before their blood levels reach abnormal thresholds and before the onset of microcytic anaemia.

6.1.5 Developing a dynamic prediction model for bowel cancer detection

The fourth objective was to develop a clinical prediction model that uses trajectories in FBC levels to identify two-year risk of bowel cancer (Chapter 5). A multivariate joint model was developed for men and women separately using relevant FBC parameters and analytic methods identified in the systematic review (Chapter 2), CPRD data derived from the data validation and quality assessment (Chapter 3), and relevant trajectories towards bowel cancer diagnosis (Chapter 4). The models rely only on data available earlier than two years before diagnosis, where they could facilitate early detection in asymptomatic individuals.

The multivariate joint models performed well for both males and females. A preliminary assessment of the models in relevant subgroups was also performed but needs assessing in larger external validation studies. As expected, my joint models perform very similarly to the ColonFlag machine-learning algorithm, but my models are more transparent, easier to interpret and embed into clinical practice, and provide a measure of absolute risk.

6.2 Novelties of this research

The systematic review was the first review of the FBC test for bowel cancer detection. With the use of a FBC becoming increasingly popular for detection, this review was timely. The systematic review provided a comprehensive overview of existing research but also identified a number of areas for future research needed. I also provide many recommendations for researchers relating to the analysis of FBC data and developing and testing prediction models.

A novelty of the data quality assessment is that it provides a starting point (i.e. the methodological approach) for future researchers to prepare reliable EHR datasets for analysis. This starting point can help tackle barriers in performing data checks identified by many existing systematic reviews. It would also help researchers adhere to the RECORD reporting guidelines for observational studies using EHR data, which include items regarding data cleaning and quality checks[142]. Additional novelties involve the individual steps in the methodological approach, such as this being the first study to explore the consistency of clinical codes assigned to patient records in CPRD.

The third objective, to identify trajectories in the FBC related to bowel cancer presence, has many novelties. It is the first longitudinal analysis of all FBC parameters for bowel cancer diagnosis in the UK population, with only two other studies identified, which looked at haemoglobin only in non-UK populations[95, 187]. Another novelty is that I assessed trajectories before diagnosis by tumour stage separately. One other study was identified that assessed haemoglobin levels at the time of diagnosis by tumour stage[189], but I assessed various FBC parameters and used repeated FBC measures. Another novelty of this research is that it shows limitations in current clinical practice, which are barriers to early detection. For example, I identified trends to be more predictive than single (current) FBCs and that FBC reference ranges and national guidelines for referral are sub-optimal for early detection. This research identified methods to improve current clinical practice.

The prediction model is novel because it applies new advanced statistical methods, multivariate joint modelling, to big data for clinical risk prediction. In general, most existing models use statistic measures (i.e. from one baseline time-point) to identify risk and most existing joint models are univariate, whereas my models use multivariate data. Another novelty of my joint models is that they are designed for early detection, using only FBC data available earlier than

two years before diagnosis, whereas existing FBC-based models use data up to the time of diagnosis. My models could therefore bring forward the diagnosis to a time where the cancer is easier to treat and remove, whereas currently, diagnoses are made too late in most cases, such as at a late tumour stage, where chances of survival are reduced[7]. My joint models also make use of recent technological advancements, with only one statistical package available in R software at the start of this research that does multivariate joint modelling, only made available in 2018[208].

The internal validation of the joint models was the first to give an assessment of how well the joint models work in a separate patient sample. It was the first direct comparison of any model with the ColonFlag algorithm, which is receiving a lot of interest in clinical practice. It is also the first to assess the ColonFlag algorithm in relevant subgroups, although this needs further work, such as testing in larger subgroups.

6.3 Future work

There are many ideas for future work:

1. In the systematic review (Chapter 2), I identified many FBC parameters as prognostic factors for bowel cancer diagnosis. Of these, haemoglobin and mean corpuscular volume are actively used in clinical practice to aid decisions to refer patients for further testing if levels cross certain thresholds[53, 54, 131, 183]. Such thresholds do not exist for other relevant FBC parameters or combinations of parameters. These could be investigated and performance measures derived, such as AUC, sensitivity, specificity, and positive and negative predictive values.

2. My methodological approach to assess EHR data quality is specific to the CPRD database (Chapter 3). It is expected that this methodology would grow as researchers start using it. I plan to perform external validation of my joint models using other EHR databases (described further below), which would present opportunities to further develop and standardise the methodological approach.

3. When exploring trajectories in FBC levels (Chapter 4), I identified that current FBC reference ranges are sub-optimal. Reference ranges differ by practice, as most practices use only patients in their records to derive reference ranges[177-181]. Additionally, the reference range for patients aged 13+ or sometimes 15+ years would not be suitable for the entire age group. I therefore plan to derive age-adjusted FBC reference ranges for adults, derived using real world and national-level data to standardise for the UK. Ranges may also be adjusted for ethnicity, but it is likely this information is not accurately recorded in EHR databases.

4. My joint models rely on only age, sex, and FBC parameter levels to identify risk of diagnosis (Chapter 5). Although I have provided many reasons for only using these (Appendix D1), other confounders, co-morbidities, and risk factors for bowel cancer, such as BMI and inflammatory bowel disease[254], could be considered for inclusion as further work. The addition of further variables to the models would require additional computational capacity, which would be investigated and funded for in future grant proposals.

5. External validation is an essential step in the testing of prediction models[125]. External validation assesses the generalisability, applicability, and performance of models in similar settings after refinements are made during internal validation, such as to reduce

overfitting. Before a prediction model is implemented in practice, external validation should be performed, as prediction models generally perform worse in external datasets than those used in the development process, which includes internal validation[255]. They therefore instil confidence in the reliability and accuracy of models[256]. External validation work is planned, with potential data sources including THIN[243], QResearch[246], RCPG Research and Surveillance Centre[244], Care and Health Information Exchange (Hampshire Health Record)[245], UK Biobank[257], Secure Anonymised Information Linkage (SAIL)[258], and HAVEN (a secondary care dataset)[259]. These external validations would provide larger samples for testing (as no data splitting would be involved) and include comparing performance measures between multiple models, including my joint models and the QCancer Colorectal models on the same patient sample. This study would help identify whether the QCancer Colorectal model can identify cases early, which I expect it will not because symptoms often manifest late in the cancer development process, although this has not previously been studied. External validation will also allow larger sample sizes to derive reliable performance measures of the joint models in relevant subgroups, such as by age. Additional subgroups can also be assessed, such as number of FBCs within subgroups of age, subgroups of time between FBCs, and subgroups of co-morbidities. A decision curve analysis would also be performed to balance the benefits with drawbacks of intervention in clinical practice due to the prediction model. External validation using other data sources would also allow me to assess generalisability of the models to other EHR databases. For example, CPRD GOLD includes entity coding to code patient data (see Section 3.2.2), a coding system not used by other EHR databases. Therefore, such external validations could assess transportability between EHR systems and general practice.

6. A feasibility/pilot study is under consideration, where the joint models would be incorporated into a single or group of primary care practices. This study would assess the logistics of programming a dynamic prediction model into GP/EHR software and pulling the data automatically from the system, allow a prospective assessment of the models, and assess interpretability of the risk scores by clinicians, which may involve deriving risk groups to indicate what risk values should be considered high or low.

7. Haemoglobin is known to change over time due to the presence of other various other cancers[187]. Such trajectories have not been identified for other FBC parameters for these other cancer types. I would explore these trajectories and subsequently expand the outcome of the multivariate joint models to include various cancer types. The models would therefore predict risk of diagnosis of various cancer types in asymptomatic individuals.

8. One possible future study would involve a health economics and cost-effectiveness analysis to assess the burden of my joint models on healthcare. For example, the study may investigate how costs of performing routine blood tests in the general population compare to current practice, where FBCs are only performed at the GP's request. This analysis may be part of a larger feasibility study, as described earlier.

9. Over the last 12 months, FIT testing has proved a useful test outside the screening programme for patients with low-risk symptoms. FIT is accurate (85% sensitivity[17]) and negative findings avoids the need for colonoscopy, based on a 98% negative predictive value in a recent primary care study[26]. However, these patients are still symptomatic, which, as described throughout the thesis, is problematic for early detection. Patients identified as high-risk from my models could be referred for FIT testing, which is cheaper

than colonoscopy and carries less burden on patients and healthcare workers. However, such negative predictive values in high-risk asymptomatic patients identified from my models would need to be investigated.

6.4 How this research benefited me: a personal note

DPhils can enhance career progression in many ways, such as helping to develop the required skillset to become an independent researcher. This DPhil offered me many opportunities to grow academically and develop my skills to an extent that would not have been possible without it. I learnt many things throughout this research and provide a list of some of the ways in Appendix E1. For example, this research gave me the opportunity to prepare and analyse complex, big EHR datasets and develop prediction models using advanced statistical methods. This research has ensured I am better equipped to perform my own research studies in the future. I am very thankful to have been able to do this DPhil. Ultimately, this research showed me that I have the potential to improve patient well-being.

6.5 Conclusion

This research was performed to identify whether the FBC can optimally be utilised for bowel cancer detection. This involved performing various research objectives to learn from existing research studies and to eliminate some key uncertainties in existing approaches for detection. I developed dynamic prediction models that could facilitate earlier detection and improve survival rates. The models use routinely available data and would not require additional work for healthcare workers or patients. These models perform well based on my initial testing, but further testing is required before use in general practice.

Bibliography

[1] Cancer Research UK. *Bowel cancer incidence statistics*. 2020 [Accessed 7 December 2020]; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-Zero.

[2] Cancer Research UK. *Bowel cancer mortality statistics*. 2021 [Accessed 2 July 2021]; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-One.

[3] Araghi, M., Arnold, M., Rutherford, M.J., Guren, M.G., Cabasag, C.J., Bardot, A., Ferlay, J., Tervonen, H., Shack, L., Woods, R.R., et al., *Colon and rectal cancer survival in seven high-income countries 2010-2014: variation by age and stage at diagnosis (the ICBP SURVMARK-2 project).* Gut, 2021. **70**(1): p. 114-126.

[4] Joachim, C., Macni, J., Drame, M., Pomier, A., Escarmant, P., Veronique-Baudin, J., and Vinh-Hung, V., *Overall survival of colorectal cancer by stage at diagnosis: Data from the Martinique Cancer Registry.* Medicine (Baltimore), 2019. **98**(35): p. e16941.

[5] White, A., Ironmonger, L., Steele, R.J.C., Ormiston-Smith, N., Crawford, C., and Seims, A., *A review of sex-related differences in colorectal cancer incidence, screening uptake, routes to diagnosis, cancer stage and survival in the UK.* BMC Cancer, 2018. **18**(1): p. 906.

[6] Bowel Cancer UK. *Staging and grading*. [Accessed 7 December 2020]; Available from: https://www.bowelcanceruk.org.uk/about-bowel-cancer/diagnosis/staging-and-grading/.

[7] Cancer Research UK. *Bowel cancer survival by stage at diagnosis*. 2019 [Accessed 7 December 2020]; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/survival#heading-Three.

[8] Adelstein, B.A., Macaskill, P., Chan, S.F., Katelaris, P.H., and Irwig, L., *Most bowel cancer symptoms do not indicate colorectal cancer and polyps: a systematic review.* BMC Gastroenterol, 2011. **11**: p. 65.

[9] National Health Service. *Symptoms - Bowel cancer*. 2019 [Accessed 8 December 2020]; Available from: https://www.nhs.uk/conditions/bowel-cancer/symptoms/.

[10] Rasmussen, S., Larsen, P.V., Sondergaard, J., Elnegaard, S., Svendsen, R.P., and Jarbol, D.E., *Specific and non-specific symptoms of colorectal cancer and contact to general practice.* Fam Pract, 2015. **32**(4): p. 387-94.

[11] Robertson, R., Campbell, C., Weller, D.P., Elton, R., Mant, D., Primrose, J., Nugent, K., Macleod, U., and Sharma, R., *Predicting colorectal cancer risk in patients with rectal bleeding.* Br J Gen Pract, 2006. **56**(531): p. 763-7.

[12] Walter, F.M., Emery, J.D., Mendonca, S., Hall, N., Morris, H.C., Mills, K., Dobson, C., Bankhead, C., Johnson, M., Abel, G.A., et al., *Symptoms and patient factors associated with longer time to diagnosis for colorectal cancer: results from a prospective cohort study.* Br J Cancer, 2016. **115**(5): p. 533-41.

[13] Holt, T., Birks, J., Bankhead, C., Nicholson, B.D., Fuller, A., and Patnick, J. *Do changes in full blood count indices predate symptom reporting in people with undiagnosed bowel cancer? Retrospective analysis using cohort and case control designs*. in *Society for Academic Primary Care*. 2021. UK.

[14] GOV.UK. *Bowel cancer screening: programme overview*. [Accessed 7 December 2020]; Available from: https://www.gov.uk/guidance/bowel-cancer-screening-programme-overview#screening-tests.

[15] National Health Service. *Bowel cancer screening overview*. 2018 [Accessed 7 December 2020]; Available from: https://www.nhs.uk/conditions/bowel-cancer-screening/.

[16] Alpha Laboratories. *Faecal-Immunochemical-Test*. [Accessed 7 December 2020]; Available from: https://www.faecal-immunochemical-test.co.uk/.

[17] Lee, J.K., Liles, E.G., Bent, S., Levin, T.R., and Corley, D.A., *Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis.* Ann Intern Med, 2014. **160**(3): p. 171.

[18] Niedermaier, T., Tikk, K., Gies, A., Bieck, S., and Brenner, H., *Sensitivity of Fecal Immunochemical Test for Colorectal Cancer Detection Differs According to Stage and Location.* Clin Gastroenterol Hepatol, 2020. **18**(13): p. 2920-2928 e6.

[19] Araghi, M., Soerjomataram, I., Bardot, A., Ferlay, J., Cabasag, C.J., Morrison, D.S., De, P., Tervonen, H., Walsh, P.M., Bucher, O., et al., *Changes in colorectal cancer incidence in seven high-income countries: a population-based study.* Lancet Gastroenterol Hepatol, 2019. **4**(7): p. 511-518.

[20] Cancer Research UK. *Bowel cancer incidence trends over time*. 2020 [Accessed 7 December 2020]; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence#heading-Two.

[21] Sandhu, G.S., Anders, R., Walde, A., Leal, A.D., King, G.T., Leong, S., S. Lindsey Davis, Purcell, W.T., Goodman, K.A., Herter, W., et al. *High incidence of advanced stage cancer and prolonged rectal bleeding history before diagnosis in young-onset patients with colorectal cancer*. 2019. Journal of Clinical Oncology.

[22] Vuik, F.E., Nieuwenburg, S.A., Bardou, M., Lansdorp-Vogelaar, I., Dinis-Ribeiro, M., Bento, M.J., Zadnik, V., Pellise, M., Esteban, L., Kaminski, M.F., et al., *Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years.* Gut, 2019. **68**(10): p. 1820-1826.

[23] Moss, S.M., Campbell, C., Melia, J., Coleman, D., Smith, S., Parker, R., Ramsell, P., Patnick, J., and Weller, D.P., *Performance measures in three rounds of the English bowel cancer screening pilot.* Gut, 2012. **61**(1): p. 101-7.

[24] Public Health Scotland Information Services Division. *Scottish Bowel Screening Programme Statistics For the two-year period of invitations between 1 May 2017 and 30 April 2019*. 2020 [Accessed 8 December 2020]; Available from: https://beta.isdscotland.org/find-publications-and-data/conditions-and-diseases/cancer/scottish-bowel-screening-programme-statistics/.

[25] Bowel Cancer UK. *Uptake rates for bowel cancer screening far too low - don't ignore the test!* 2018 [Accessed 8 December 2020]; Available from: https://www.bowelcanceruk.org.uk/news-and-blogs/news/uptake-rates-for-bowel-cancer-screening-far-too-low-dont-ignore-the-test!/.

[26] Bailey, S.E.R., Abel, G.A., Atkins, A., Byford, R., Davies, S.J., Mays, J., McDonald, T.J., Miller, J., Neck, C., Renninson, J., et al., *Diagnostic performance of a faecal immunochemical test for patients with low-risk symptoms of colorectal cancer in primary care: an evaluation in the South West of England.* Br J Cancer, 2021. **124**(7): p. 1231-1236.

[27] GOV.UK. *NHS bowel cancer screening: helping you decide (FIT kit)*. 2021 [Accessed 1 August 2021]; Available from: https://www.gov.uk/government/publications/bowel-cancer-screening-benefits-and-risks/nhs-bowel-cancer-screening-helping-you-decide.

[28] GOV.UK. *NHS Bowel Cancer Screening: bowel scope screening*. 2019 [Accessed 7 December 2020]; Available from:

https://www.gov.uk/government/publications/bowel-scope-screening-test-what-it-is/nhs-bowel-cancer-screening-bowel-scope-screening.

[29] National Health Service. *Bowel scope screening*. 2018 [Accessed 7 December 2020]; Available from: https://www.nhs.uk/conditions/bowel-cancer-screening/bowel-scope-screening/.

[30] American Cancer Society. *Colorectal Cancer Screening Tests*. 2020 [Accessed 7 December 2020]; Available from: https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/screening-tests-used.html.

[31] McGregor, L.M., Bonello, B., Kerrison, R.S., Nickerson, C., Baio, G., Berkman, L., Rees, C.J., Atkin, W., Wardle, J., and von Wagner, C., *Uptake of Bowel Scope (Flexible Sigmoidoscopy) Screening in the English National Programme: the first 14 months.* J Med Screen, 2016. **23**(2): p. 77-82.

[32] Bowel Cancer UK. *Hospital tests*. 2018 [Accessed 8 December 2020]; Available from: https://www.bowelcanceruk.org.uk/about-bowel-cancer/diagnosis/hospital-tests/.

[33] Lo, S.H., Halloran, S., Snowball, J., Seaman, H., Wardle, J., and von Wagner, C., *Colorectal cancer screening uptake over three biennial invitation rounds in the English bowel cancer screening programme.* Gut, 2015. **64**(2): p. 282-91.

[34] Gavin, D.R., Valori, R.M., Anderson, J.T., Donnelly, M.T., Williams, J.G., and Swarbrick, E.T., *The national colonoscopy audit: a nationwide assessment of the quality and safety of colonoscopy in the UK.* Gut, 2013. **62**(2): p. 242-9.

[35] NHS. *What is a colonoscopy?* 2019 [Accessed 1 August 2021]; Available from: https://www.nhs.uk/conditions/colonoscopy/.

[36] Hardcastle, J.D., Chamberlain, J.O., Robinson, M.H., Moss, S.M., Amar, S.S., Balfour, T.W., James, P.D., and Mangham, C.M., *Randomised controlled trial of faecal-occult-blood screening for colorectal cancer.* Lancet, 1996. **348**(9040): p. 1472-7.

[37] Hewitson, P., Glasziou, P., Watson, E., Towler, B., and Irwig, L., *Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update.* Am J Gastroenterol, 2008. **103**(6): p. 1541-9.

[38] Towler, B., Irwig, L., Glasziou, P., Kewenter, J., Weller, D., and Silagy, C., *A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, hemoccult.* BMJ, 1998. **317**(7158): p. 559-65.

[39] Ekberg, M., Callender, M., Hamer, H., and Rogers, S., *Exploring the decision to participate in the National Health Service Bowel Cancer Screening Programme.* Eur J Cancer Prev, 2014. **23**(5): p. 391-7.

[40] Gale, C.R., Deary, I.J., Wardle, J., Zaninotto, P., and Batty, G.D., *Cognitive ability and personality as predictors of participation in a national colorectal cancer screening programme: the English Longitudinal Study of Ageing.* J Epidemiol Community Health, 2015. **69**(6): p. 530-5.

[41] Hall, N.J., Rubin, G.P., Dobson, C., Weller, D., Wardle, J., Ritchie, M., and Rees, C.J., *Attitudes and beliefs of non-participants in a population-based screening programme for colorectal cancer.* Health Expect, 2015. **18**(5): p. 1645-57.

[42] Palmer, C.K., Thomas, M.C., von Wagner, C., and Raine, R., *Reasons for non-uptake and subsequent participation in the NHS Bowel Cancer Screening Programme: a qualitative study.* Br J Cancer, 2014. **110**(7): p. 1705-11.

[43] von Wagner, C., Good, A., Smith, S.G., and Wardle, J., *Responses to procedural information about colorectal cancer screening using faecal occult blood testing: the role of consideration of future consequences.* Health Expect, 2012. **15**(2): p. 176-86.

[44] Hofmann, B., *Ethical issues with colorectal cancer screening-a systematic review.* J Eval Clin Pract, 2017. **23**(3): p. 631-641.

[45] Benton, S.C., Butler, P., Allen, K., Chesters, M., Rickard, S., Stanley, S., Roope, R., Vulkan, D., and Duffy, S.W., *GP participation in increasing uptake in a national bowel cancer screening programme: the PEARL project.* Br J Cancer, 2017. **116**(12): p. 1551-1557.

[46] Hewitson, P., Ward, A.M., Heneghan, C., Halloran, S.P., and Mant, D., *Primary care endorsement letter and a patient leaflet to improve participation in colorectal cancer screening: results of a factorial randomised trial.* Br J Cancer, 2011. **105**(4): p. 475-80.

[47] Raine, R., Duffy, S.W., Wardle, J., Solmi, F., Morris, S., Howe, R., Kralj-Hans, I., Snowball, J., Counsell, N., Moss, S., et al., *Impact of general practice endorsement on the social gradient in uptake in bowel cancer screening.* Br J Cancer, 2016. **114**(3): p. 321-6.

[48] Shankleman, J., Massat, N.J., Khagram, L., Ariyanayagam, S., Garner, A., Khatoon, S., Rainbow, S., Rangrez, S., Colorado, Z., Hu, W., et al., *Evaluation of a service intervention to improve awareness and uptake of bowel cancer screening in ethnically-diverse areas.* Br J Cancer, 2014. **111**(7): p. 1440-7.

[49] Bain, B.J., *A Beginner's Guide to Blood Cells*. 2nd ed. 2004, Malden, Massachusetts, USA: Blackwell Pub.

[50] Bain, B.J., *Blood Cells: A Practical Guide*. 5th ed. 2015, The Atrium, Southern Gate, Chichester, West Sussex, UK: John Wiley & Sons Ltd.

[51] Virdee, P.S., Marian, I.R., Mansouri, A., Elhussein, L., Kirtley, S., Holt, T., and Birks, J., *The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal.* Cancers (Basel), 2020. **12**(9).

[52] Hamilton, W., Lancashire, R., Sharp, D., Peters, T.J., Cheng, K.K., and Marshall, T., *The importance of anaemia in diagnosing colorectal cancer: a case-control study using electronic primary care records.* Br J Cancer, 2008. **98**(2): p. 323-7.

[53] National Institute for Health and Care Excellence. *Suspected cancer recognition and referral: site or type of cancer*. 2020 [Accessed 20 October 2021]; Available from: http://pathways.nice.org.uk/pathways/suspected-cancer-recognition-and-referral.

[54] World Health Organisation. *Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity*. 2011 [Accessed 27 January 2021]; Available from: https://apps.who.int/iris/bitstream/handle/10665/85839/WHO_NMH_NHD_MNM_11.1_eng.pdf?ua=1.

[55] Ankus, E., Price, S.J., Ukoumunne, O.C., Hamilton, W., and Bailey, S.E.R., *Cancer incidence in patients with a high normal platelet count: a cohort study using primary care data.* Fam Pract, 2018. **35**(6): p. 671-675.

[56] Bailey, S.E., Ukoumunne, O.C., Shephard, E.A., and Hamilton, W., *Clinical relevance of thrombocytosis in primary care: a prospective cohort study of cancer incidence using English electronic medical records and cancer registry data.* Br J Gen Pract, 2017. **67**(659): p. e405-e413.

[57] Boursi, B., Mamtani, R., Hwang, W.T., Haynes, K., and Yang, Y.X., *A Risk Prediction Model for Sporadic CRC Based on Routine Lab Results.* Dig Dis Sci, 2016. **61**(7): p. 2076-86.

[58] Shi, C., Xie, M., Li, L., Li, K., and Hu, B.L., *The association and diagnostic value of red blood cell distribution width in colorectal cancer.* Medicine (Baltimore), 2019. **98**(19): p. e15560.

[59] Spell, D.W., Jones, D.V., Jr., Harper, W.F., and David Bessman, J., *The value of a complete blood count in predicting cancer of the colon.* Cancer Detect Prev, 2004. **28**(1): p. 37-42.

[60] Lee, Y.J., Lee, H.R., Nam, C.M., Hwang, U.K., and Jee, S.H., *White blood cell count and the risk of colon cancer.* Yonsei Med J, 2006. **47**(5): p. 646-56.

[61] Margolis, K.L., Rodabough, R.J., Thomson, C.A., Lopez, A.M., McTiernan, A., and Women's Health Initiative Research, G., *Prospective study of leukocyte count as a predictor of incident breast, colorectal, endometrial, and lung cancer and mortality in postmenopausal women.* Arch Intern Med, 2007. **167**(17): p. 1837-44.

[62] Pilling, L.C., Atkins, J.L., Kuchel, G.A., Ferrucci, L., and Melzer, D., *Red cell distribution width and common disease onsets in 240,477 healthy volunteers followed for up to 9 years.* PLoS One, 2018. **13**(9): p. e0203504.

[63] Medial EarlySign. *Who is EarlySign?* 2020 [Accessed 9 December 2020]; Available from: https://earlysign.com/about-us/.

[64] Kinar, Y., Kalkstein, N., Akiva, P., Levin, B., Half, E.E., Goldshtein, I., Chodick, G., and Shalev, V., *Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study.* J Am Med Inform Assoc, 2016. **23**(5): p. 879-90.

[65] Birks, J., Bankhead, C., Holt, T.A., Fuller, A., and Patnick, J., *Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records.* Cancer Med, 2017. **6**(10): p. 2453-2460.

[66] National Institute for Health and Care Excellence. *ColonFlag for identifying people at risk of colorectal cancer*. 2018 [Accessed 27 January 2021]; Available from: https://www.nice.org.uk/advice/mib142/resources/colonflag-for-identifying-people-at-risk-of-colorectal-cancer-pdf-2285963452805317#:~:text=The%20reported%20ORs%20showed%20that,consequences%20or%20utility%20of%20ColonFlag.

[67] Goshen, R., Mizrahi, B., Akiva, P., Kinar, Y., Choman, E., Shalev, V., Sopik, V., Kariv, R., and Narod, S.A., *Predicting the presence of colon cancer in members of a health maintenance organisation by evaluating analytes from standard laboratory records.* Br J Cancer, 2017. **116**(7): p. 944-950.

[68] Kilincalp, S., Coban, S., Akinci, H., Hamamci, M., Karaahmet, F., Coskun, Y., Ustun, Y., Simsek, Z., Erarslan, E., and Yuksel, I., *Neutrophil/lymphocyte ratio, platelet/lymphocyte ratio, and mean platelet volume as potential biomarkers for early detection and monitoring of colorectal adenocarcinoma.* Eur J Cancer Prev, 2015. **24**(4): p. 328-33.

[69] Hippisley-Cox, J. and Coupland, C., *Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm.* Br J Gen Pract, 2012. **62**(594): p. e29-37.

[70] Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., and Group, P., *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* PLoS Med, 2009. **6**(7): p. e1000097.

[71] Centre for Reviews and Dissemination University of York. *PROSPERO - International prospective register of systematic reviews*. [Accessed 9 December 2020]; Available from: https://www.crd.york.ac.uk/prospero/.

[72] Virdee, P.S., Kirtley, S., Elhussein, L., Watkinson, P.J., Holt, T.A., and Birks, J., *Components of the full blood count as risk factors for colorectal cancer detection: a systematic review protocol.* BMJ Open, 2019. **9**(12): p. e032759.

[73] Bramer, W.M., Rethlefsen, M.L., Kleijnen, J., and Franco, O.H., *Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study.* Syst Rev, 2017. **6**(1): p. 245.

[74] EBSCO. *CINAHL Database*. [Accessed 10 December 2020]; Available from: https://www.ebsco.com/products/research-databases/cinahl-database.

[75] Gusenbauer, M. and Haddaway, N.R., *Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources.* Res Synth Methods, 2020. **11**(2): p. 181-217.

[76] Haddaway, N.R., Collins, A.M., Coughlin, D., and Kirk, S., *The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching.* PLoS One, 2015. **10**(9): p. e0138237.

[77] Piasecki, J., Waligora, M., and Dranseika, V., *Google Search as an Additional Source in Systematic Reviews.* Sci Eng Ethics, 2018. **24**(2): p. 809-810.

[78] Clarivate. *EndNote*. [Accessed 27 January 2021]; Available from: https://endnote.com/.
[79] Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A., *Rayyan-a web and mobile app for systematic reviews.* Syst Rev, 2016. **5**(1): p. 210.
[80] Deeks, J.J., Higgins, J.P.T., and Altman, D.G. *Cochrane Handbook for Systematic Reviews of Interventions version 6.1. Chapter 10: Analysing data and undertaking meta-analyses.* 2020 [Accessed 3 March 2021]; Available from: https://training.cochrane.org/handbook/current/chapter-10#section-10-10-2.
[81] Etikan, I., Alkassim, R., and Abubakar, S., *Meta-analysis of using both cohort and case control study.* Biometrics & Biostatistics International Journal, 2016. **3**(3): p. 96-98.
[82] Higgins, J.P.T., Li, K., and Deeks, J.J. *Cochrane Handbook for Systematic Reviews of Interventions version 6.1. Chapter 6: Choosing effect measures and computing estimates of effect.* 2020 [Accessed 3 March 2021]; Available from: https://training.cochrane.org/handbook/current/chapter-06#section-6-3.
[83] Hayden, J.A., van der Windt, D.A., Cartwright, J.L., Cote, P., and Bombardier, C., *Assessing bias in studies of prognostic factors.* Ann Intern Med, 2013. **158**(4): p. 280-6.
[84] Wolff, R.F., Moons, K.G.M., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S., and Groupdagger, P., *PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies.* Ann Intern Med, 2019. **170**(1): p. 51-58.
[85] Acher, P.L., Al-Mishlab, T., Rahman, M., and Bates, T., *Iron-deficiency anaemia and delay in the diagnosis of colorectal cancer.* Colorectal Dis, 2003. **5**(2): p. 145-8.
[86] Ay, S., Eryilmaz, M.A., Aksoy, N., Okus, A., Unlu, Y., and Sevinc, B., *Is early detection of colon cancer possible with red blood cell distribution width?* Asian Pacific journal of cancer prevention : APJCP, 2015. **16**(2): p. 753-756.
[87] Ayling, R.M., Lewis, S.J., and Cotter, F., *Potential roles of artificial intelligence learning and faecal immunochemical testing for prioritisation of colonoscopy in anaemia.* British Journal of Haematology, 2019. **185**(2): p. 311-316.
[88] Bafandeh, Y., Khoshbaten, M., Sadat, A.T.E., and Farhang, S., *Clinical predictors of colorectal polyps and carcinoma in a low prevalence region: Results of a colonoscopy based study.* World Journal of Gastroenterology, 2008. **14**(10): p. 1534-1538.
[89] Cakmak, E., Soylu, S., Yonem, O., and Yilmaz, A., *Neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, and red blood cell distribution width as new biomarkers in patients with colorectal cancer.* Erciyes Medical Journal, 2017. **39**(3): p. 131-136.
[90] Collins, G.S. and Altman, D.G., *Identifying patients with undetected colorectal cancer: An independent validation of QCancer (Colorectal).* British Journal of Cancer, 2012. **107**(2): p. 260-265.
[91] Cross, A.J., Wooldrage, K., Robbins, E.C., Pack, K., Brown, J.P., Hamilton, W., Thompson, M.R., Flashman, K.G., Halligan, S., Thomas-Gibson, S., et al., *Whole-colon investigation vs. flexible sigmoidoscopy for suspected colorectal cancer based on presenting symptoms and signs: a multicentre cohort study.* British Journal of Cancer, 2019. **120**(2): p. 154-164.
[92] Cubiella, J., Vega, P., Salve, M., Diaz-Ondina, M., Alves, M.T., Quintero, E., Alvarez-Sanchez, V., Fernandez-Banares, F., Boadas, J., Campo, R., et al., *Development and external validation of a faecal immunochemical test-based prediction model for colorectal cancer detection in symptomatic patients.* BMC Medicine, 2016. **14**(1).
[93] Fijten, G.H., Starmans, R., Muris, J.W., Schouten, H.J., Blijham, G.H., and Knottnerus, J.A., *Predictive value of signs and symptoms for colorectal cancer in patients with rectal bleeding in general practice.* Family Practice, 1995. **12**(3): p. 279-86.
[94] Firat, F., Arslan, A.K., Colak, C., and Harputluoglu, H., *Estimation of risk factors associated with colorectal cancer: an application of knowledge discovery in databases.* Kuwait Journal of Science, 2016. **43**(2): p. 151-161.

[95] Goldshtein, I., Neeman, U., Chodick, G., and Shalev, V., *Variations in hemoglobin before colorectal cancer diagnosis.* Eur J Cancer Prev, 2010. **19**(5): p. 342-4.

[96] Hamilton, W., Round, A., Sharp, D., and Peters, T.J., *Clinical features of colorectal cancer before diagnosis: A population-based case-control study.* British Journal of Cancer, 2005. **93**(4): p. 399-405.

[97] Hamilton, W., Lancashire, R., Sharp, D., Peters, T.J., Cheng, K.K., and Marshall, T., *The risk of colorectal cancer with symptoms at different ages and between the sexes: A case-control study.* BMC Medicine, 2009. **7**(17).

[98] Hilsden, R.J., Heitman, S.J., Mizrahi, B., Narod, S.A., and Goshen, R., *Prediction of findings at screening colonoscopy using a machine learning algorithm based on complete blood counts (ColonFlag).* PLoS ONE, 2018. **13**(11).

[99] Hippisley-Cox, J. and Coupland, C., *Symptoms and risk factors to identify men with suspected cancer in primary care: Derivation and validation of an algorithm.* British Journal of General Practice, 2013. **63**(606): p. e1-e10.

[100] Hippisley-Cox, J. and Coupland, C., *Symptoms and risk factors to identify women with suspected cancer in primary care: Derivation and validation of an algorithm.* British Journal of General Practice, 2013. **63**(606): p. e11-e21.

[101] Hornbrook, M.C., Goshen, R., Choman, E., O'Keeffe-Rosetti, M., Kinar, Y., Liles, E.G., and Rust, K.C., *Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data.* Digestive Diseases and Sciences, 2017. **62**(10): p. 2719-2727.

[102] Huang, J., Zhao, Y., Liao, L., Liu, S., Lu, S., Wu, C., Wei, C., Xu, S., Zhong, H., Liu, J., et al., *Evaluation of Red Cell Distribution Width to Lymphocyte Ratio as Potential Biomarker for Detection of Colorectal Cancer.* BioMed Research International, 2019: p. 1-7.

[103] Hung, N., Shen, C.C., Hu, Y.W., Hu, L.Y., Yeh, C.M., Teng, C.J., Kuan, A.S., Chen, S.C., Chen, T.J., and Liu, C.J., *Risk of cancer in patients with iron deficiency anemia: A nationwide population-based study.* PLoS ONE, 2015. **10**(3).

[104] Joosten, E., Meeuwissen, J., Vandewinckele, H., and Hiele, M., *Iron status and colorectal cancer in symptomatic elderly patients.* American Journal of Medicine, 2008. **121**(12): p. 1072-7.

[105] Kinar, Y., Akiva, P., Choman, E., Kariv, R., Shalev, V., Levin, B., Narod, S.A., and Goshen, R., *Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer.* PLoS ONE, 2017. **12**(2).

[106] Lawrenson, R., Logie, J., and Marks, C., *Risk of colorectal cancer in general practice patients presenting with rectal bleeding, change in bowel habit or anaemia.* European Journal of Cancer Care, 2006. **15**(3): p. 267-271.

[107] Marshall, T., Lancashire, R., Sharp, D., Peters, T.J., Cheng, K.K., and Hamilton, W., *The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance.* Gut, 2011. **60**(9): p. 1242-8.

[108] Mashlab, S., Large, P., Laing, W., Ng, O., D'Auria, M., Thurston, D., Thomson, S., Acheson, A.G., Humes, D.J., Banerjea, A., et al., *Anaemia as a risk stratification tool for symptomatic patients referred via the two-week wait pathway for colorectal cancer.* Annals of the Royal College of Surgeons of England, 2018. **100**(5): p. 350-356.

[109] Naef, M., Buhlmann, M., and Baer, H.U., *Small bowel tumors: Diagnosis, therapy and prognostic factors.* Langenbeck's Archives of Surgery, 1999. **384**(2): p. 176-180.

[110] Nakama, H., Zhang, B., Fattah, A.S., and Zhang, X., *Colorectal cancer in iron deficiency anemia with a positive result on immunochemical fecal occult blood.* International Journal of Colorectal Disease, 2000. **15**(5-6): p. 271-4.

[111] Panagiotopoulou, I.G., Fitzrol, D., Parker, R.A., Kuzhively, J., Luscombe, N., Wells, A.D., Menon, M., Bajwa, F.M., and Watson, M.A., *The yield of colorectal cancer among fast*

track patients with normocytic and microcytic anaemia. Annals of the Royal College of Surgeons of England, 2014. **96**(4): p. 289-293.

[112] Panzuto, F., Chiriatti, A., Bevilacqua, S., Giovannetti, P., Russo, G., Impinna, S., Pistilli, F., Capurso, G., Annibale, B., Delle Fave, G., et al., *Symptom-based approach to colorectal cancer: Survey of primary care physicians in Italy.* Digestive and Liver Disease, 2003. **35**(12): p. 869-875.

[113] Prizment, A.E., Anderson, K.E., Visvanathan, K., and Folsom, A.R., *Association of inflammatory markers with colorectal cancer incidence in the atherosclerosis risk in communities study.* Cancer Epidemiol Biomarkers Prev, 2011. **20**(2): p. 297-307.

[114] Raje, D., Mukhtar, H., Oshowo, A., and Ingham Clark, C., *What proportion of patients referred to secondary care with iron deficiency anemia have colon cancer?* Diseases of the Colon and Rectum, 2007. **50**(8): p. 1211-1214.

[115] Schneider, C., Bodmer, M., Jick, S.S., and Meier, C.R., *Colorectal cancer and markers of anemia.* European Journal of Cancer Prevention, 2018. **27**(6): p. 530-538.

[116] Song, Y., Huang, Z., Kang, Y., Lin, Z., Lu, P., Lin, Q., Cai, Z., Cao, Y., and Zhu, X., *Clinical Usefulness and Prognostic Value of Red Cell Distribution Width in Colorectal Cancer.* BioMed Research International, 2018: p. 1-7.

[117] Stapley, S., Peters, T.J., Sharp, D., and Hamilton, W., *The mortality of colorectal cancer in relation to the initial symptom at presentation to primary care and to the duration of symptoms: A cohort study using medical records.* British Journal of Cancer, 2006. **95**(10): p. 1321-1325.

[118] Thompson, M.R., O'Leary, D.P., Flashman, K., Asiimwe, A., Ellis, B.G., and Senapati, A., *Clinical assessment to determine the risk of bowel cancer using Symptoms, Age, Mass and Iron deficiency anaemia (SAMI).* British Journal of Surgery, 2017. **104**(10): p. 1393-1404.

[119] van Boxtel-Wilms, S.J.M., van Boven, K., Bor, J.H.H., Bakx, J.C., Lucassen, P., Oskam, S., and van Weel, C., *The value of reasons for encounter in early detection of colorectal cancer.* European Journal of General Practice, 2016. **22**(2): p. 91-95.

[120] Wu, Y.Y., Zhang, X., Qin, Y.Y., Qin, J.Q., and Lin, F.Q., *Mean platelet volume/platelet count ratio in colorectal cancer: a retrospective clinical study.* Bmc Cancer, 2019. **19**: p. 7.

[121] Yang, D., Quan, W., Wu, J., Ji, X., Dai, Y., Xiao, W., Chew, H., Sun, Z., and Li, D., *The value of red blood cell distribution width in diagnosis of patients with colorectal cancer.* Clinica Chimica Acta, 2018. **479**: p. 98-102.

[122] Zhou, W.W., Chu, Y.P., and An, G.Y., *Significant difference of neutrophil-lymphocyte ratio between colorectal cancer, adenomatous polyp and healthy people.* European Review for Medical and Pharmacological Sciences, 2017. **21**(23): p. 5386-5391.

[123] Zhu, X.J., Cao, Y.P., Lu, P.X., Kang, Y.L., Lin, Z., Hao, T.S., and Song, Y.F., *Evaluation of platelet indices as diagnostic biomarkers for colorectal cancer.* Scientific Reports, 2018. **8**: p. 7.

[124] Hamilton, W. *Derivation of a score for identifying colorectal cancer in primary care (Abstract: 148)*. 2008 [Accessed 15 April 2021]; Available from: https://gut.bmj.com/content/56/suppl_2/a1.

[125] Steyerberg, E.W., *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 1st ed. 2009, New York, USA: Springer.

[126] Biau, D.J., Kerneis, S., and Porcher, R., *Statistics in brief: the importance of sample size in the planning and interpretation of medical research.* Clin Orthop Relat Res, 2008. **466**(9): p. 2282-8.

[127] Faber, J. and Fonseca, L.M., *How sample size influences research outcomes.* Dental Press J Orthod, 2014. **19**(4): p. 27-9.

[128] Virdee, P.S., Fuller, A., Jacobs, M., Holt, T., and Birks, J., *Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test.* Journal of Big Data, 2020. **7**(96): p. 1-18.

[129] Del Giudice, M.E., Vella, E.T., Hey, A., Simunovic, M., Harris, W., and Levitt, C., *Systematic review of clinical features of suspected colorectal cancer in primary care.* Can Fam Physician, 2014. **60**(8): p. e405-15.

[130] Montagnana, M. and Danese, E., *Red cell distribution width and cancer.* Ann Transl Med, 2016. **4**(20): p. 399.

[131] National Institute for Health and Care Excellence. *Anaemia - iron deficiency*. 2021 [Accessed 21 June 2021]; Available from: https://cks.nice.org.uk/topics/anaemia-iron-deficiency/.

[132] Gallagher, A.M., Dedman, D., Padmanabhan, S., Leufkens, H.G.M., and de Vries, F., *The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations.* Pharmacoepidemiol Drug Saf, 2019. **28**(5): p. 563-569.

[133] Lo Re, V., 3rd, Haynes, K., Forde, K.A., Localio, A.R., Schinnar, R., and Lewis, J.D., *Validity of The Health Improvement Network (THIN) for epidemiologic studies of hepatitis C virus infection.* Pharmacoepidemiol Drug Saf, 2009. **18**(9): p. 807-14.

[134] Mathur, R., Bhaskaran, K., Chaturvedi, N., Leon, D.A., vanStaa, T., Grundy, E., and Smeeth, L., *Completeness and usability of ethnicity data in UK-based primary care and hospital databases.* J Public Health (Oxf), 2014. **36**(4): p. 684-92.

[135] Muthee, V., Bochner, A.F., Osterman, A., Liku, N., Akhwale, W., Kwach, J., Prachi, M., Wamicwe, J., Odhiambo, J., Onyango, F., et al., *The impact of routine data quality assessments on electronic medical record data quality in Kenya.* PLoS One, 2018. **13**(4): p. e0195362.

[136] Usher-Smith, J.A., Walter, F.M., Emery, J.D., Win, A.K., and Griffin, S.J., *Risk Prediction Models for Colorectal Cancer: A Systematic Review.* Cancer Prev Res (Phila), 2016. **9**(1): p. 13-26.

[137] Altman, D.G. and Royston, P., *The cost of dichotomising continuous variables.* BMJ, 2006. **332**(7549): p. 1080.

[138] Riley, R.D., Windt, D.A.v.d., Croft, P., and Moon, K.G.M., *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. 2019, Oxford, UK: Oxford University Press.

[139] Royston, P. and Sauerbrei, W., *A new measure of prognostic separation in survival data.* Stat Med, 2004. **23**(5): p. 723-48.

[140] Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J., and Kattan, M.W., *Assessing the performance of prediction models: a framework for traditional and novel measures.* Epidemiology, 2010. **21**(1): p. 128-38.

[141] von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gotzsche, P.C., Vandenbroucke, J.P., and Initiative, S., *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies.* PLoS Med, 2007. **4**(10): p. e296.

[142] Benchimol, E.I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., Sorensen, H.T., von Elm, E., Langan, S.M., and Committee, R.W., *The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement.* PLoS Med, 2015. **12**(10): p. e1001885.

[143] Collins, G.S., Reitsma, J.B., Altman, D.G., and Moons, K.G., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.* BMJ, 2015. **350**: p. g7594.

[144] Speights, V.O., Johnson, M.W., Stoltenberg, P.H., Rappaport, E.S., Helbert, B., and Riggs, M.W., *Complete blood count indices in colorectal carcinoma.* Arch Pathol Lab Med, 1992. **116**(3): p. 258-60.

[145] Ni, K., Chu, H., Zeng, L., Li, N., and Zhao, Y., *Barriers and facilitators to data quality of electronic health records used for clinical research in China: a qualitative study.* BMJ Open, 2019. **9**(7): p. e029314.
[146] Charnock, V., *Electronic healthcare records and data quality.* Health Info Libr J, 2019. **36**(1): p. 91-95.
[147] Weiskopf, N.G. and Weng, C., *Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.* J Am Med Inform Assoc, 2013. **20**(1): p. 144-51.
[148] NHS Digital. *GP Systems of Choice*. 2021 [Accessed 22 June 2021]; Available from: https://digital.nhs.uk/services/gp-systems-of-choice.
[149] CPRD. *Clinical Practice Research Datalink*. 2021 [Accessed 22 June 2021]; Available from: https://www.cprd.com/.
[150] CPRD. *Data*. 2021 [Accessed 22 June 2021]; Available from: https://www.cprd.com/Data.
[151] CPRD. *Primary care data for public health research*. 2021 [Accessed 22 June 2021]; Available from: https://www.cprd.com/primary-care.
[152] Padmanabhan, S. *CPRD GOLD Data Specification*. 2017 [Accessed 23 June 2021]; Available from: https://cprdcw.cprd.com/_docs/CPRD_GOLD_Full_Data_Specification_v2.0.pdf.
[153] Benson, T., *The history of the Read Codes: the inaugural James Read Memorial Lecture 2011.* Inform Prim Care, 2011. **19**(3): p. 173-82.
[154] Watson, J., Nicholson, B.D., Hamilton, W., and Price, S., *Identifying clinical features in primary care electronic health record studies: methods for codelist development.* BMJ Open, 2017. **7**(11): p. e019637.
[155] NHS Digital. *Read Codes*. 2020 [Accessed 4 April 2022]; Available from: https://digital.nhs.uk/services/terminology-and-classifications/read-codes.
[156] NHS Digital. *SNOMED CT*. 2021 [Accessed 4 April 2022]; Available from: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct.
[157] NHS Digital. *Read v2 to SNOMED CT Mapping Lookup (April 2020)*. 2020 [Accessed 4 April 2022]; Available from: https://hscic.kahootz.com/t_c_home/viewdatastore?datviewmode=list&pagemode=&sortdir1=asc&nexturl=&showsingleitem=N&sourcedsid=&dir=asc&dsid=407588&showallcolumns=N&sortcol1=col_0&ajaxmode=&sortcol2=&cardcolno=&sort=col_0&shownum=100&sortdir3=&_=1648914747512&sortcol3=&sortdir2=&showadvancedsearch=N&adv=&rowid=&action=&sel=&search=&btn_search=.
[158] NHS Digital. *Read CTV3 to SNOMED CT Mapping Lookup (April 2020)*. 2020 [Accessed 4 April 2022]; Available from: https://hscic.kahootz.com/t_c_home/viewdatastore?datviewmode=list&pagemode=&sortdir1=asc&nexturl=&showsingleitem=N&sourcedsid=&dir=asc&dsid=407556&showallcolumns=N&sortcol1=col_0&ajaxmode=&sortcol2=&cardcolno=&sort=col_0&shownum=100&sortdir3=&_=1649067937722&sortcol3=&sortdir2=&showadvancedsearch=N&adv=&rowid=&sel=&action=&search=&btn_search=.
[159] World Health Organisation. *ICD-10 Version:2016*. 2016 [Accessed 29 June 2021]; Available from: https://icd.who.int/browse10/2016/en.
[160] Platt, A. *The Department of Laboratory Haematology - OUH Haematology Laboratory Handbook*. 2020 [Accessed 2 August 2021]; Available from: https://www.ouh.nhs.uk/services/departments/laboratory-medicine/haematology-laboratories/documents/haematology-handbook.pdf.
[161] POZ.com. *Understanding Your Lab Work (Blood Tests)*. 2021 [Accessed 12 November 2021]; Available from: https://www.poz.com/basics/hiv-basics/understanding-lab-work-blood-tests.

[162] O'Sullivan, J.W., Stevens, S., Hobbs, F.D.R., Salisbury, C., Little, P., Goldacre, B., Bankhead, C., Aronson, J.K., Perera, R., and Heneghan, C., *Temporal trends in use of tests in UK primary care, 2000-15: retrospective analysis of 250 million tests.* BMJ, 2018. **363**: p. k4666.

[163] Cancer Research UK. *Bowel cancer incidence by age*. 2020 [Accessed 2 July 2021]; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence#heading-One.

[164] CALIBER. *Diseases of the blood*. 2020 [Accessed 2 July 2021]; Available from: https://www.caliberresearch.org/portal/chapter/5#Diseases%20of%20the%20blood.

[165] Stocks, S.J., Kontopantelis, E., Akbarov, A., Rodger, S., Avery, A.T., and Ashcroft, D.M. *Examining variations in prescribing safety in UK general practice: a cross-sectional study using the Clinical Practice Research Datalink*. 2015 [Accessed 2 July 2021]; Available from: https://clinicalcodes.rss.mhs.man.ac.uk/medcodes/article/25/codelist/res25-m3_lft_fbc/.

[166] Iwagami, M., Caplin, B., Smeeth, L., Tomlinson, L.A., and Nitsch, D. *Clinical Codelist - Read codes for severe mental illness*. 2018 [Accessed 2 July 2021]; Available from: https://datacompass.lshtm.ac.uk/id/eprint/868/.

[167] Riley, R.D., Ensor, J., Snell, K.I., Debray, T.P., Altman, D.G., Moons, K.G., and Collins, G.S., *External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges.* BMJ, 2016. **353**: p. i3140.

[168] Abdala, O.T. and Saeed, M., *Estimation of Missing Values in Clinical Laboratory Measurements of ICU Patients Using a Weighted K-Nearest Neighbors Algorithm.* Computers in Cardiology, 2004: p. 693-696.

[169] Hughes, R.A., Heron, J., Sterne, J.A.C., and Tilling, K., *Accounting for missing data in statistical analyses: multiple imputation is not always the answer.* Int J Epidemiol, 2019. **48**(4): p. 1294-1304.

[170] Smuk, M., Carpenter, J.R., and Morris, T.P., *What impact do assumptions about missing data have on conclusions? A practical sensitivity analysis for a cancer survival registry.* BMC Med Res Methodol, 2017. **17**(1): p. 21.

[171] White, I.R. and Carlin, J.B., *Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values.* Stat Med, 2010. **29**(28): p. 2920-31.

[172] Grace-Martin, K. *Two Recommended Solutions for Missing Data: Multiple Imputation and Maximum Likelihood*. [Accessed 31 July 2021]; Available from: https://www.theanalysisfactor.com/missing-data-two-recommended-solutions/.

[173] Allison, P.D. *Handling Missing Data by Maximum Likelihood*. 2012 [Accessed 31 July 2021]; Available from: https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf.

[174] Allison, P.D. *Maximum Likelihood is Better than Multiple Imputation: Part II*. 2015 [Accessed 31 July 2021]; Available from: https://statisticalhorizons.com/ml-is-better-than-mi.

[175] Shin, T., Davison, M.L., and Long, J.D., *Maximum likelihood versus multiple imputation for missing data in small longitudinal samples with nonnormality.* Psychol Methods, 2017. **22**(3): p. 426-449.

[176] von Hippel, P.T., *New Confidence Intervals and Bias Comparisons Show That Maximum Likelihood Can Beat Multiple Imputation in Small Samples.* Structural Equation Modeling: A Multidisciplinary Journal, 2016. **23**(3): p. 422-437.

[177] MayoClinic. *Complete blood count (CBC)*. 2020 [Accessed 20 October 2021]; Available from: https://www.mayoclinic.org/tests-procedures/complete-blood-count/about/pac-20384919.

[178] The Royal Wolverhaptom NHS Trust. *Haematology Normal Adult Reference Range*. 2017 [Accessed 20 October 2021]; Available from:

https://www.royalwolverhampton.nhs.uk/services/service-directory-a-z/pathology-services/departments/haematology/haematology-normal-adult-reference-ranges/.

[179] York Teaching Hospital NHS Foundation Trust. *Full Blood Count (FBC) Reference Ranges*. 2021 [Accessed 20 October 2021]; Available from: https://www.yorkhospitals.nhs.uk/our-services/a-z-of-services/laboratory-medicine1/haematology-/.

[180] Maidstone and Tunbridge Wells NHS Trust. *Haematology Reference Ranges*. 2020 [Accessed 20 October 2021]; Available from: https://www.mtw.nhs.uk/wp-content/uploads/2020/11/Haematology-reference-ranges.pdf.

[181] Gloucestershire Hospitals NHS Foundation Trust. *Full Blood Count (FBC)*. 2013 [Accessed 20 October 2021]; Available from: https://www.gloshospitals.nhs.uk/our-services/services-we-offer/pathology/tests-and-investigations/full-blood-count-fbc/.

[182] Maner, B.S. and Moosavi, L. *Mean Corpuscular Volume*. 2021 [Accessed 20 October 2021]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK545275/.

[183] National Institute for Health and Care Excellence. *What investigations should I arrange to confirm iron deficiency anaemia?* 2021 [Accessed 20 October 2021]; Available from: https://cks.nice.org.uk/topics/anaemia-iron-deficiency/diagnosis/investigations/.

[184] StatisticsHowTo. *Lowess Smoothing in Statistics: What is it?* 2013 [Accessed 20 October 2021]; Available from: https://www.statisticshowto.com/lowess-smoothing/.

[185] Fitzmaurice, G.M., Laird, N.M., and Ware, J.H., *Applied Longitudinal Analysis*. 2nd ed. 2011: John Wiley & Sons.

[186] Diggle, P.J., Heagery, P., Lian, K., and Zeger, S.L., *Analysis of Longitudinal Data*. 2nd ed. 2002, New York, USA: Oxford University Press Inc.

[187] Edgren, G., Bagnardi, V., Bellocco, R., Hjalgrim, H., Rostgaard, K., Melbye, M., Reilly, M., Adami, H.O., Hall, P., and Nyren, O., *Pattern of declining hemoglobin concentration before cancer diagnosis.* Int J Cancer, 2010. **127**(6): p. 1429-36.

[188] EarlySign, M. *Patient Care Prioritized: Where it Matters Most*. 2020 [Accessed 1 Novemebr 2021]; Available from: https://earlysign.com/.

[189] Vayrynen, J.P., Tuomisto, A., Vayrynen, S.A., Klintrup, K., Karhu, T., Makela, J., Herzig, K.H., Karttunen, T.J., and Makinen, M.J., *Preoperative anemia in colorectal cancer: relationships with tumor characteristics, systemic inflammation, and survival.* Sci Rep, 2018. **8**(1): p. 1126.

[190] Whyte, M.B. and Kelly, P., *The normal range: it is not normal and it is not a range.* Postgrad Med J, 2018. **94**(1117): p. 613-616.

[191] MedlinePlus. *RDW (Red Cell Distribution Width)*. 2020 [Accessed 1 November 2021]; Available from: https://medlineplus.gov/lab-tests/rdw-red-cell-distribution-width/.

[192] Cox, D.R., *Regression Models and Life-Tables.* Journal of the Royal Statistical Society. Series B (Methodological), 1972. **34**(2): p. 187–220.

[193] Basha, L. and Hoxha, F., *Kernel Estimation of the Baseline Function in the Cox Model.* European Scientific Journal, 2019. **15**(6): p. 105-118.

[194] Cox, D.R. and Oakes, D., *Analysis of Survival Data*. 1984: Chapman and Hall/CRC.

[195] Kalbfleisch, J.D. and Prentice, R.L., *The Statistical Analysis of Failure Time Data*. 2nd ed. 2002: John Wiley & Sons.

[196] Lin, D.Y., *On the Breslow estimator.* Lifetime Data Anal, 2007. **13**(4): p. 471-80.

[197] Therneau, T.M. and Grambsch, P.M., *Modeling Survival Data: Extending the Cox Model*. 1st ed. 2000: Springer-Verlag, New York.

[198] Zhang, Z., Reinikainen, J., Adeleke, K.A., Pieterse, M.E., and Groothuis-Oudshoorn, C.G.M., *Time-varying covariates and coefficients in Cox regression models.* Ann Transl Med, 2018. **6**(7): p. 121.

[199] Sweeting, M.J., Barrett, J.K., Thompson, S.G., and Wood, A.M., *The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study.* Stat Med, 2017. **36**(28): p. 4514-4528.

[200] Wong, J., Taljaard, M., Forster, A.J., Escobar, G.J., and van Walraven, C., *Addition of time-dependent covariates to a survival model significantly improved predictions for daily risk of hospital death.* J Eval Clin Pract, 2013. **19**(2): p. 351-7.

[201] Sweeting, M.J. and Thompson, S.G., *Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture.* Biom J, 2011. **53**(5): p. 750-63.

[202] Maziarz, M., Heagerty, P., Cai, T., and Zheng, Y., *On longitudinal prediction with time-to-event outcome: Comparison of modeling options.* Biometrics, 2017. **73**(1): p. 83-93.

[203] Laird, N.M. and Ware, J.H., *Random-effects models for longitudinal data.* Biometrics, 1982. **38**(4): p. 963-74.

[204] Wulfsohn, M.S. and Tsiatis, A.A., *A joint model for survival and longitudinal data measured with error.* Biometrics, 1997. **53**(1): p. 330-9.

[205] Henderson, R., Diggle, P., and Dobson, A., *Joint modelling of longitudinal measurements and event time data.* Biostatistics, 2000. **1**(4): p. 465-80.

[206] Crowther, M.J., *merlin—A unified modeling framework for data analysis and methods development in Stata.* The Stata Journal, 2020. **20**: p. 763–784.

[207] Elashoff R.M., Li G., and N., L., *Joint Modeling of Longitudinal and Time-to-Event Data* 2016, Boca Raton, Florida, USA: CRC Press.

[208] Hickey, G.L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R., *joineRML: a joint model and software package for time-to-event and multivariate longitudinal outcomes.* BMC Med Res Methodol, 2018. **18**(1): p. 50.

[209] Rizopoulos, D., *Joint Models for Longitudinal and Time-to-Event Data With Applications in R*. 2012: Chapman and Hall/CRC.

[210] Asar, O., Ritchie, J., Kalra, P.A., and Diggle, P.J., *Joint modelling of repeated measurement and time-to-event data: an introductory tutorial.* Int J Epidemiol, 2015. **44**(1): p. 334-44.

[211] Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S., *Gaussian processes for time-series modelling.* Philos Trans A Math Phys Eng Sci, 2013. **371**(1984): p. 20110550.

[212] Liu, X., Rong, J., and Liu, X., *Best linear unbiased prediction for linear combinationsin general mixed linear models.* Journal of Multivariate Analysis 2008. **99**(8): p. 1503–1517.

[213] Andrinopoulou, E.R., Harhay, M.O., Ratcliffe, S.J., and Rizopoulos, D., *Reflections on modern methods: Dynamic prediction using joint models of longitudinal and time-to-event data.* Int J Epidemiol, 2021.

[214] Cekic, S., Aichele, S., Brandmaier, A.M., Köhncke, Y., and Ghisletta, P., *A Tutorial for Joint Modeling of Longitudinal and Time-to-Event Data in R.* Quantitative and Computational Methods in Behavioral Sciences, 2021(e2979): p. 1-40.

[215] Papageorgiou, G., Mauff, K., Tomer, A., and Rizopoulos, D., *An Overview of Joint Modeling of Time-to-Event and Longitudinal Outcomes.* Annual Review of Statistics and Its Application, 2019. **6**: p. 223–240.

[216] Bull, L.M., Lunt, M., Martin, G.P., Hyrich, K., and Sergeant, J.C., *Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods.* Diagn Progn Res, 2020. **4**: p. 9.

[217] Paige, E., Barrett, J., Stevens, D., Keogh, R.H., Sweeting, M.J., Nazareth, I., Petersen, I., and Wood, A.M., *Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk.* Am J Epidemiol, 2018. **187**(7): p. 1530-1538.

[218] Mann, C.J., *Observational research methods. Research design II: cohort, cross sectional, and case-control studies.* Emerg Med J, 2003. **20**(1): p. 54-60.

[219] Philipson, P., Hickey, G.L., Crowther, M.J., and Kolamunnage-Dona, R., *Faster Monte Carlo estimation of joint models for time-to-event and multivariate longitudinal data.* Computational Statistics and Data Analysis, 2020. **151**: p. 1-14.

[220] Benoit, J.S., Chan, W., and Doody, R.S., *Joint coverage probability in a simulation study on Continuous-Time Markov Chain parameter estimation.* J Appl Stat, 2015. **42**(12): p. 2531-2538.

[221] StatisticsHowTo. *Brier Score: Definition, Examples*. 2016 [Accessed 10 October 2021]; Available from: https://www.statisticshowto.com/brier-score/.

[222] Royston, P., *Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities.* The Stata Journal, 2014. **14**(4).

[223] Royston, P., *Tools for checking calibration of a Cox model in external validation: Prediction of population-averaged survival curves based on risk groups.* The Stata Journal, 2015. **15**(1): p. 275–291.

[224] Steyerberg, E.W. and Vergouwe, Y., *Towards better clinical prediction models: seven steps for development and an ABCD for validation.* Eur Heart J, 2014. **35**(29): p. 1925-31.

[225] Van Calster, B., van Smeden, M., De Cock, B., and Steyerberg, E.W., *Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study.* Stat Methods Med Res, 2020. **29**(11): p. 3166-3178.

[226] Moons, K.G., Altman, D.G., Vergouwe, Y., and Royston, P., *Prognosis and prognostic research: application and impact of prognostic models in clinical practice.* BMJ, 2009. **338**: p. b606.

[227] Royston, P. and Altman, D.G., *External validation of a Cox prognostic model: principles and methods.* BMC Med Res Methodol, 2013. **13**: p. 33.

[228] Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W., Topic Group 'Evaluating diagnostic, t., and prediction models' of the, S.i., *Calibration: the Achilles heel of predictive analytics.* BMC Med, 2019. **17**(1): p. 230.

[229] van Houwelingen, H., *Validation, calibration, revision and combination of prognostic survival models.* Statistics in Medicine, 2000. **19**(24): p. 3401-3415.

[230] Efron, B. and Tibshirani, R.J., *An Introduction to the Bootstrap*. 2000: CHAPMAN & HALL/CRC.

[231] Harrell, F.E., Jr., *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. 2015: Springer, Cham.

[232] Harrell, F.E., Jr., Lee, K.L., and Mark, D.B., *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat Med, 1996. **15**(4): p. 361-87.

[233] Luijken, K., Wynants, L., van Smeden, M., Van Calster, B., Steyerberg, E.W., Groenwold, R.H.H., and Collaborators, *Changing predictor measurement procedures affected the performance of prediction models in clinical examples.* J Clin Epidemiol, 2020. **119**: p. 7-18.

[234] Thangaratinam, S., Allotey, J., Marlin, N., Mol, B.W., Von Dadelszen, P., Ganzevoort, W., Akkermans, J., Ahmed, A., Daniels, J., Deeks, J., et al., *Development and validation of Prediction models for Risks of complications in Early-onset Pre-eclampsia (PREP): a prospective cohort study.* Health Technol Assess, 2017. **21**(18): p. 1-100.

[235] Thoya , D., Waititu, A., Magheto, T., and Ngunyi, A., *Evaluating methods of assessing "Optimism" in regression models.* American Journal of Applied Mathematics and Statistics, 2018. **6**(4): p. 126-134.

[236] Riley, R.D., Ensor, J., Snell, K.I.E., Harrell, F.E., Jr., Martin, G.P., Reitsma, J.B., Moons, K.G.M., Collins, G., and van Smeden, M., *Calculating the sample size required for developing a clinical prediction model.* BMJ, 2020. **368**: p. m441.

[237] van Smeden, M., de Groot, J.A., Moons, K.G., Collins, G.S., Altman, D.G., Eijkemans, M.J., and Reitsma, J.B., *No rationale for 1 variable per 10 events criterion for binary logistic regression analysis.* BMC Med Res Methodol, 2016. **16**(1): p. 163.

[238] Riley, R.D., Snell, K.I., Ensor, J., Burke, D.L., Harrell, F.E., Jr., Moons, K.G., and Collins, G.S., *Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes.* Stat Med, 2019. **38**(7): p. 1276-1296.

[239] Bujang, M.A. and Adnan, T.H., *Requirements for Minimum Sample Size for Sensitivity and Specificity Analysis.* J Clin Diagn Res, 2016. **10**(10): p. YE01-YE06.

[240] Negida, A., Fahim, N.K., and Negida, Y., *Sample Size Calculation Guide - Part 4: How to Calculate the Sample Size for a Diagnostic Test Accuracy Study based on Sensitivity, Specificity, and the Area Under the ROC Curve.* Adv J Emerg Med, 2019. **3**(3): p. e33.

[241] UK, C.R. *Bowel cancer incidence by sex and UK country*. 2021 [Accessed 7 November 2021]; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence#heading-Zero.

[242] Johnson, C.M., Wei, C., Ensor, J.E., Smolenski, D.J., Amos, C.I., Levin, B., and Berry, D.A., *Meta-analyses of colorectal cancer risk factors.* Cancer Causes Control, 2013. **24**(6): p. 1207-22.

[243] The Health Improvement Network. *The Health Improvement Network (THIN)*. [Accessed 11 November 2021]; Available from: https://www.the-health-improvement-network.com/.

[244] Royal College of general Practitioners. *RCGP Research and Surveillance Centre (RSC)*. 2021 [Accessed 16 November 2021]; Available from: https://www.rcgp.org.uk/clinical-and-research/our-programmes/research-and-surveillance-centre.aspx.

[245] Care and Health Information Exchange. *The Care and Health Information Exchange (CHIE)* [Accessed 16 November 2021]; Available from: https://careandhealthinformationexchange.org.uk/.

[246] QResearch. *QResearch*. [Accessed 11 November 2021]; Available from: https://www.qresearch.org/.

[247] Fitzgerald, M., Saville, B.R., and Lewis, R.J., *Decision curve analysis.* JAMA, 2015. **313**(4): p. 409-10.

[248] Holmberg, L. and Vickers, A., *Evaluation of prediction models for decision-making: beyond calibration and discrimination.* PLoS Med, 2013. **10**(7): p. e1001491.

[249] Localio, A.R. and Goodman, S., *Beyond the usual prediction accuracy metrics: reporting results for clinical decision making.* Ann Intern Med, 2012. **157**(4): p. 294-5.

[250] Van Calster, B., Wynants, L., Verbeek, J.F.M., Verbakel, J.Y., Christodoulou, E., Vickers, A.J., Roobol, M.J., and Steyerberg, E.W., *Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators.* Eur Urol, 2018. **74**(6): p. 796-804.

[251] Vickers, A.J. and Elkin, E.B., *Decision curve analysis: a novel method for evaluating prediction models.* Med Decis Making, 2006. **26**(6): p. 565-74.

[252] Vickers, A.J., Van Calster, B., and Steyerberg, E.W., *Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests.* BMJ, 2016. **352**: p. i6.

[253] Vickers, A.J., van Calster, B., and Steyerberg, E.W., *A simple, step-by-step guide to interpreting decision curve analysis.* Diagn Progn Res, 2019. **3**: p. 18.

[254] Society, A.C. *Colorectal Cancer Risk Factors*. 2020 [Accessed 11 November 2021]; Available from: https://www.cancer.org/cancer/colon-rectal-cancer/causes-risks-prevention/risk-factors.html#written_by.

[255] Ramspek, C.L., Jager, K.J., Dekker, F.W., Zoccali, C., and van Diepen, M., *External validation of prognostic models: what, why, how, when and where?* Clin Kidney J, 2021. **14**(1): p. 49-58.

[256] Hickey, G.L. and Blackstone, E.H., *External model validation of binary clinical risk prediction models in cardiovascular and thoracic surgery.* J Thorac Cardiovasc Surg, 2016. **152**(2): p. 351-5.
[257] UK Biobank. *UK Biobank*. [Accessed 11 November 2021]; Available from: https://www.ukbiobank.ac.uk/.
[258] SAIL. *SAIL Overview*. [Accessed 6 December 2021]; Available from: https://saildatabank.com/about-us/overview/.
[259] Nuffield Department of Clinical Neurosciences. *Hospital Alerting Via Electronic Noticeboard (HAVEN)*. 2021 [Accessed 11 November 2021]; Available from: https://www.ndcn.ox.ac.uk/research/critical-care-research-group-kadoorie-centre/research-studies/hospital-alerting-via-electronic-noticeboard-haven.
[260] Kho, A.N., Hui, S., Kesterson, J.G., and McDonald, C.J., *Which observations from the complete blood cell count predict mortality for hospitalized patients?* J Hosp Med, 2007. **2**(1): p. 5-12.
[261] Warny, M., Helby, J., Nordestgaard, B.G., Birgens, H., and Bojesen, S.E., *Incidental lymphopenia and mortality: a prospective cohort study.* CMAJ, 2020. **192**(2): p. E25-E33.
[262] Lab Tests Online. *Full Blood Count (FBC)*. 2020 [Accessed 4 August 2021]; Available from: https://labtestsonline.org.uk/tests/full-blood-count-fbc.
[263] myDr. *Full blood count (FBC)*. 2017 [Accessed 4 August 2021]; Available from: https://www.mydr.com.au/tests-investigations/full-blood-count-fbc/.
[264] MedlinePlus. *Complete Blood Count (CBC)*. 2020 [Accessed 4 August 2021]; Available from: https://medlineplus.gov/lab-tests/complete-blood-count-cbc/.
[265] Gaskell, H., Derry, S., and Moore, R.A., *Is there an association between low dose aspirin and anemia (without overt bleeding)? Narrative review.* BMC Geriatr, 2010. **10**: p. 71.
[266] Lassale, C., Curtis, A., Abete, I., van der Schouw, Y.T., Verschuren, W.M.M., Lu, Y., and Bueno-de-Mesquita, H.B.A., *Elements of the complete blood count associated with cardiovascular disease incidence: Findings from the EPIC-NL cohort study.* Sci Rep, 2018. **8**(1): p. 3290.
[267] Better Health Channel. *Blood count*. 2019 [Accessed 4 August 2021]; Available from: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/blood-count.
[268] Arakawa, K., Hosono, A., Shibata, K., Ghadimi, R., Fuku, M., Goto, C., Imaeda, N., Tokudome, Y., Hoshino, H., Marumoto, M., et al., *Changes in blood biochemical markers before, during, and after a 2-day ultramarathon.* Open Access J Sports Med, 2016. **7**: p. 43-50.
[269] Boudjeltia, K.Z., Faraut, B., Stenuit, P., Esposito, M.J., Dyzma, M., Brohee, D., Ducobu, J., Vanhaeverbeek, M., and Kerkhofs, M., *Sleep restriction increases white blood cells, mainly neutrophil count, in young healthy men: a pilot study.* Vasc Health Risk Manag, 2008. **4**(6): p. 1467-70.
[270] Liu, H., Wang, G., Luan, G., and Liu, Q., *Effects of sleep and sleep deprivation on blood cell count and hemostasis parameters in healthy humans.* J Thromb Thrombolysis, 2009. **28**(1): p. 46-9.
[271] Akunov, A., Sydykov, A., Toktash, T., Doolotova, A., and Sarybaev, A., *Hemoglobin Changes After Long-Term Intermittent Work at High Altitude.* Front Physiol, 2018. **9**: p. 1552.
[272] Rowles, P.M. and Williams, E.S., *Abnormal red cell morphology in venous blood of men climbing at high altitude.* Br Med J (Clin Res Ed), 1983. **286**(6375): p. 1396.
[273] Word Health Organisation. *Guide to cancer early diagnosis*. 2020 [Accessed 5 August 2021]; Available from: https://www.who.int/cancer/publications/cancer_early_diagnosis/en/.

[274] van der Ploeg, T., Austin, P.C., and Steyerberg, E.W., *Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints.* BMC Med Res Methodol, 2014. **14**: p. 137.

Appendix A

Appendices for Chapter 2 (systematic review)

Most of the contents in this appendix have previously been published (in supplementary material)[51]:

Virdee P.S. *et al*. The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37* (2020). DOI: 10.3390/cancers12092348. Published under a CC-BY licence – permissions not required.

A1. Search strategy used in the systematic review

MEDLINE

Database and platform: Medline (Ovid MEDLINE® Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE® Daily and Ovid MEDLINE®) 1946 to present. Search date: 3 September 2019.

1. Colonic Neoplasms/bl [Blood]

2. Colonic Neoplasms/di [Diagnosis]

3. Colonic Neoplasms/ep [Epidemiology]

4. Colorectal Neoplasms/di [Diagnosis]

5. Colorectal Neoplasms/ep [Epidemiology]

6. Colorectal Neoplasms/bl [Blood]

7. Rectal Neoplasms/bl [Blood]

8. Rectal Neoplasms/di [Diagnosis]

9. Rectal Neoplasms/ep [Epidemiology]

10. Adenomatous Polyposis Coli/

11. Sigmoid Neoplasms/

12. Colorectal Neoplasms, Hereditary Nonpolyposis/

13. ((colorectal or bowel or colon or colonic or rectal or rectum) adj3 (cancer$ or carcinoma$ or adenoma$ or neoplas$ or metasta$ or carcinogen$ or tumour$ or tumor$ or malignan$ or adenocarcinoma$)).ti,ab,kw.

14. or/1-13

15. exp Blood Cell Count/

16. exp Hemoglobins/

17. Blood Platelets/

18. Neutrophils/

19. Basophils/

20. Eosinophils/

21. Lymphocytes/

22. Monocytes/

23. Occult Blood/

24. Thrombocytosis/

25. Leukocytosis/

26. Lymphocytosis/

27. Eosinophilia/

28. Anemia/

29. Leukopenia/

30. Neutropenia/

31. Lymphopenia/

32. Thrombocytopenia/

33. Polycythemia/

34. Erythrocytes/

35. Leukocytes/

36. Pancytopenia/

37. ((blood or platelet) adj2 count$).ti,ab,kw.

38. (CBC or FBC).ti,ab,kw.

39. (blood adj2 exam$).ti,ab,kw.

40. (haematolog$ or hematolog$ or haemoglobin or hemoglobin or haematocrit or hematocrit).ti,ab,kw.

41. ((red or white) adj1 blood adj1 cell$).ti,ab,kw.

42. (mean adj1 (platelet or corpuscular) adj1 volume$).ti,ab,kw.

43. (mean adj1 corpuscular adj1 (haemoglobin or hemoglobin)).ti,ab,kw.

44. (platelet$ or basophil or basophils or eosinophil or eosinophils or lymphocyte$ or monocyte$ or neutrophil or neutrophils or erythrocyte$ or leukocyte$).ti,ab,kw.

45. (blood adj1 (test$ or draw$)).ti,ab,kw.

46. (neutrophili$ or monocytosis or basophili$ or anemi$ or anaemi$ or monocytopenia or eosinopenia or basopenia or basocytopenia or thrombocytopeni$ or leucocytosis or lymphocytosis or eosinophili$ or leucopenia or leukopenia or neutropenia or lymphopenia or lymphocytopenia or pancytopenia or polycythemia or bicytopenia).ti,ab,kw.

47. or/15-46

48. (abnormalit$ or diagnos$ or "pre-diagnos$" or prediagnos$ or change$ or detect$ or elevat$ or distribut$ or deficien$ or identif$ or presence or indicati$ or determin$ or undiagnosed or definition$ or altered or alteration$).ti,ab,kw.

49. 47 and 48

50. (predict$ or prognos$ or suspected).ti,ab,kw.

51. (risk adj1 (predict$ or marker$ or scor$)).ti,ab,kw.

52. Predictive Value of Tests/

53. Probability/

54. Prognosis/

55. Risk Factors/

56. Risk Assessment/

57. Incidence/

58. or/50-57

59. 14 and 49 and 58

EMBASE

Database and platform: Embase 1974 to present. Search date: 3 September 2019.

1. exp Colon Cancer/

2. exp Colon Tumor/

3. exp Rectum Tumor/

4. Colon Polyposis/

5. Hereditary Nonpolyposis Colorectal Cancer/

6. ((colorectal or bowel or colon or colonic or rectal or rectum) adj3 (cancer$ or carcinoma$ or adenoma$ or neoplas$ or metasta$ or carcinogen$ or tumour$ or tumor$ or malignan$ or adenocarcinoma$)).ti,ab,kw.

7. or/1-6

8. exp Blood Cell Count/

9. Hemoglobin/

10. Hemoglobin Blood Level/

11. Thrombocyte/

12. Neutrophil/

13. Basophil/

14. Eosinophil/

15. Lymphocyte/

16. exp Monocyte/

17. Occult Blood/

18. Thrombocytosis/

19. Leukocytosis/

20. Basophilia/

21. exp Lymphocytosis/

22. Eosinophilia/

23. Monocytosis/

24. Neutrophilia/

25. Anemia/

26. Leukopenia/

27. Monocytopenia/

28. exp Neutropenia/

29. Eosinopenia/

30. Lymphocytopenia/

31. Thrombocytopenia/

32. Polycythemia/

33. Erythrocyte/

34. Leukocyte/

35. Pancytopenia/

36. Mean Corpuscular Volume/

37. ((blood or platelet) adj2 count$).ti,ab,kw.

38. (CBC or FBC).ti,ab,kw.

39. (blood adj2 exam$).ti,ab,kw.

40. Hematocrit/

41. (haematolog$ or hematolog$ or haemoglobin or hemoglobin or haematocrit or hematocrit).ti,ab,kw.

42. ((red or white) adj1 blood adj1 cell$).ti,ab,kw.

43. (mean adj1 (platelet or corpuscular) adj1 volume$).ti,ab,kw.

44. (mean adj1 corpuscular adj1 (haemoglobin or hemoglobin)).ti,ab,kw.

45. (platelet$ or basophil or basophils or eosinophil or eosinophils or lymphocyte$ or monocyte$ or neutrophil or neutrophils or erythrocyte$ or leukocyte$).ti,ab,kw.

46. (blood adj1 (test$ or draw$)).ti,ab,kw.

47. (neutrophili$ or monocytosis or basophili$ or anemi$ or anaemi$ or monocytopenia or eosinopenia or basopenia or basocytopenia or thrombocytopeni$ or leucocytosis or lymphocytosis or eosinophili$ or leucopenia or leukopenia or neutropenia or lymphopenia or lymphocytopenia or pancytopenia or polycythemia or bicytopenia).ti,ab,kw.

48. or/8-47

49. (abnormalit$ or diagnos$ or "pre-diagnos$" or prediagnos$ or change$ or detect$ or elevat$ or distribut$ or deficien$ or identif$ or presence or indicati$ or determin$ or undiagnosed or definition$ or altered or alteration$).ti,ab,kw.

50. 48 and 49

51. (predict$ or prognos$ or suspected).ti,ab,kw.

52. (risk adj1 (predict$ or marker$ or scor$)).ti,ab,kw.

53. Predictive Value/

54. exp Prediction/

55. Probability/

56. exp Prognosis/

57. "Sensitivity and Specificity"/

58. Risk Factor/

59. Risk Assessment/

60. or/51-59

61. 7 and 50 and 60

CINAHL

Database and platform: CINAHL (via EBSCOhost). Search date: 3 September 2019.

1. (MH "Colonic Neoplasms")

2. (MH "Colorectal Neoplasms")

3. (MH "Rectal Neoplasms")

4. (MH "Adenomatous Polyposis Coli")

5. (MH "Sigmoid Neoplasms")

6. (MH "Colorectal Neoplasms, Hereditary Nonpolyposis")

7. TI ((colorectal or bowel or colon or colonic or rectal or rectum) N3 (cancer* or carcinoma* or adenoma* or neoplas* or metasta* or carcinogen* or tumour* or tumor* or malignan* or adenocarcinoma*)) OR AB ((colorectal or bowel or colon or colonic or rectal or rectum) N3 (cancer* or carcinoma* or adenoma* or neoplas* or metasta* or carcinogen* or tumour* or tumor* or malignan* or adenocarcinoma*))

8. S1 OR S2 OR S3 OR S4 OR S5 OR S6 OR S7

9. (MH "Blood Cells+")

10. (MH "Hemoglobins+")

11. (MH "Occult Blood")

12. (MH "Thrombocytosis")

13. (MH "Eosinophilia")

14. (MH "Anemia")

15. (MH "Leukopenia")

16. (MH "Neutropenia")

17. (MH "Lymphopenia")

18. (MH "Thrombocytopenia")

19. (MH "Polycythemia")

20. (MH "Pancytopenia")

21. TI ((blood or platelet) N2 count*) OR AB ((blood or platelet) N2 count*)

22. TI (CBC or FBC) OR AB (CBC or FBC)

23. TI (blood N2 exam*) OR AB (blood N2 exam*)

24. TI (haematolog* or hematolog* or haemoglobin or hemoglobin or haematocrit or hematocrit) OR AB (haematolog* or hematolog* or haemoglobin or hemoglobin or haematocrit or hematocrit)

25. TI ((red or white) N1 blood N1 cell*) OR AB ((red or white) N1 blood N1 cell*)

26. TI (mean N1 (platelet or corpuscular) N1 volume*) OR AB (mean N1 (platelet or corpuscular) N1 volume*)

27. TI (mean N1 corpuscular N1 (haemoglobin or hemoglobin)) OR AB (mean N1 corpuscular N1 (haemoglobin or hemoglobin))

28. TI (platelet* or basophil or basophils or eosinophil or eosinophils or lymphocyt* or monocyt* or neutrophil or neutrophils or erythrocyt* or leukocyt*) OR AB (platelet* or basophil or basophils or eosinophil or eosinophils or lymphocyt* or monocyt* or neutrophil or neutrophils or erythrocyt* or leukocyt*)

29. TI (blood N1 (test* or draw*)) OR AB (blood N1 (test* or draw*))

30. TI (neutrophili* or monocytosis or basophili* or anemi* or anaemi* or monocytopenia or eosinopenia or basopenia or basocytopenia or thrombocytopeni* or leucocytosis or lymphocytosis or eosinophili* or leucopenia or leukopenia or neutropenia or lymphopenia or lymphocytopenia or pancytopenia or polycythemia or bicytopenia) OR AB (neutrophili* or monocytosis or basophili* or anemi* or anaemi* or monocytopenia or eosinopenia or basopenia or basocytopenia or thrombocytopeni* or leucocytosis or lymphocytosis or eosinophili* or leucopenia or leukopenia or neutropenia or lymphopenia or lymphocytopenia or pancytopenia or polycythemia or bicytopenia)

31. S9 OR S10 OR S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25 OR S26 OR S27 OR S28 OR S29 OR S30

32. TI (abnormalit* or diagnos* or "pre-diagnos*" or prediagnos* or change* or detect* or elevat* or distribut* or deficien* or identif* or presence or indicati* or determin* or undiagnosed or definition* or altered or alteration*) OR AB (abnormalit* or diagnos* or "pre-diagnos*" or prediagnos* or change* or detect* or elevat* or distribut* or deficien* or identif* or presence or indicati* or determin* or undiagnosed or definition* or altered or alteration*)

33. S31 AND S32

34. TI (predict* or prognos* or suspected) OR AB (predict* or prognos* or suspected)

35. TI (risk N1 (predict* or marker* or scor*)) OR AB (risk N1 (predict* or marker* or scor*))

36. (MH "Predictive Value of Tests")

37. (MH "Probability")

38. (MH "Prognosis")

39. (MH "Risk Factors")

40. (MH "Risk Assessment")

41. (MH "Incidence")

42. S34 OR S35 OR S36 OR S37 OR S38 OR S39 OR S40 OR S41

43. S8 and S33 and S42

Web of Science

Database and platform: Web of Science (Web of Science Core Collection: Science Citation Index Expanded (SCI-EXPANDED) --1945-present; Social Sciences Citation Index (SSCI) --1956-present; Conference Proceedings Citation Index- Science (CPCI-S) --1990-present) (via Clarivate). Search date: 3 September 2019.

1. TS=((colorectal or bowel or colon or colonic or rectal or rectum or sigmoid) NEAR/3 (cancer* or carcinoma* or adenoma* or neoplas* or metasta* or carcinogen* or tumour* or tumor* or malignan* or adenocarcinoma*))

2. SU=Hematology

3. TS="blood cell count"

4. TS="occult blood"

5. TS=((blood or platelet) NEAR/2 count*)

6. TS=(CBC or FBC)

7. TS=(blood NEAR/2 exam*)

8. TS=(haematolog* or hematolog* or haemoglobin or hemoglobin or haematocrit or hematocrit)

9. TS=((red or white) NEAR/1 blood NEAR/1 cell*)

10. TS=(mean NEAR/1 (platelet or corpuscular) NEAR/1 volume*)

11. TS=(mean NEAR/1 corpuscular NEAR/1 (haemoglobin or hemoglobin))

12. TS=(platelet* or basophil or basophils or eosinophil or eosinophils or lymphocyte* or monocyte* or neutrophil or neutrophils or erythrocyte* or leukocyte*)

13. TS=(blood NEAR/1 (test* or draw*))

14. TS=(neutrophili* or monocytosis or basophili* or anemi* or anaemi* or monocytopenia or eosinopenia or eosinophilia or basopenia or basocytopenia or thrombocytopeni* or thrombocytosis or leucocytosis or lymphocytosis or eosinophili* or leucopenia or leukopenia or

neutropenia or lymphopenia or lymphocytopenia or pancytopenia or polycythemia or bicytopenia)

15. #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14

16. TS=(abnormalit* or diagnos* or "pre-diagnos*" or prediagnos* or change* or detect* or elevat* or distribut* or deficien* or identif* or presence or indicati* or determin* or undiagnosed or definition* or altered or alteration*)

17. #15 and #16

18. TS=(predict* or prognos* or probabilit* or suspected)

19. TS=(risk NEAR/1 (predict* or marker* or scor*))

20. #18 or #19

21. #1 and #17 and #20

A2. Data extraction forms in the systematic review

Form 1 (Aim 1) – Study design, analysis, and participant details overview:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**Publication Details**					**Study Design**								**Study Setting**			
2	Who is the first author?	What is the title of the paper?	What is the year of publication?	If multiple rows are used for a single paper, please give a unique code to each	If multiple rows for the same study, please indicate what each row is for	Is the data collected prospectively or retrospectively?	If other, please state	What is the type of study?	If other, please state	When did recruitment start? Specify month and year (if reported)	When did recruitment stop? Specify month and year (if reported)	What year did recruitment start?	What year did recruitment stop?	What is the geographical location of the data (i.e. the country)?	Where was the data obtained from?	If other, please state	If electronic record, please specify which database. List all if multiple are used. If a local centre(s), specify the centre(s)
3																	
4																	
5																	

	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	**Participants**															**Blood Test**	
2	Please copy and paste the patient population into this field, i.e. inclusion or exclusion criteria	How many patients were recruited, I.e. before exclusions?	How many patients were included in the study, i.e. after exclusions?	What is the patient setting?	If other, please state	Were patients symptomatic or asymptomatic?	If other, please state	What statistic is used to describe the age at study entry?	If other, please state	What is the value for the age at study entry?	What statistic is used to describe the age at diagnosis/censor date?	If other, please state	What is the value for the age at diagnosis/censor date?	What proportion of patients were male?	What proportion of patients were female?	What is the total number of blood tests available, if reported?	Were multiple blood tests per patient used in the analysis?
3																	
4																	
5																	

	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT
1				**Analysis Overview**								
2	If single, which blood test was chosen for analysis (assuming multiple were availale over time)?	If other, please specify	If multiple, describe how they were used and the analysis methods	Is the analysis stratified, and using what stratification variables?	If other, please specify stratification variables here	Is the analysis matched, and what are the matching variables?	If other, please specify matching variables here	How many controls were matched to each case?	How was missing data dealt with for the analysis of blood components?	If other, please specify	Were assumptions assessed for any of the (relevant) tests performed?	What significance level was used?
3												
4												
5												

Form 1 (Aim 1) – Red blood cell count details:

Note: this same form is replicated for the remaining 19 FBC parameters

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**Publication Details**				**Data Cleaning & Outcome Time Frame**							**Descriptive Analysis**					
2	Who is the first author?	What is the title of the paper?	What is the year of publication?	If multiple rows are used for a single paper, please give a unique code to each	Was red blood cell count analysed?	What is the range (min, max) of the blood component analysed?	If certain blood component values were deleted before analysis, please state which and why (copy and paste)?	Was data cleaning of the blood component described?	What is the outcome time frame, e.g. two-year risk of diagnosis, one-year incidence, etc.?	If a reason has been provided for the choice of the outcome time frame, please specify this here	What is the unit of measurement used for the blood component?	Was a descriptive analysis (only) performed (i.e. no statistical tests used)?	Specify the outcome and it's levels	Was the component analysed as a continuous or categorical variable?	If continuous, please indicate what transformation was applied, if any	What statistic was used to describe the continous variable?	If other, please specify
3																	

	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1																	
2	Specify the mean/median or 'other' statistic for each outcome level	If categorical, what is blood component group 1?	If categorical, what is blood component group 2?	If categorical, what is blood component group 3?	If categorical, what is blood component group 4?	If categorical, what is blood component group 5?	If categorical, what is blood component group 6?	If categorical, what is blood component group 7?	If categorical, what is blood component group 8?	If categorical, what is blood component group 9?	If categorical, what is blood component group 10?	How many patients are in group 1?	How many patients are in group 2?	How many patients are in group 3?	How many patients are in group 4?	How many patients are in group 5?	How many patients are in group 6?
3																	

	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY
1																	
2	How many patients are in group 7?	How many patients are in group 8?	How many patients are in group 9?	How many patients are in group 10?	How many experienced the outcome in group 1. Specify for each outcome level if multiple	How many experienced the outcome in group 2. Specify for each outcome level if multiple	How many experienced the outcome in group 3. Specify for each outcome level if multiple	How many experienced the outcome in group 4. Specify for each outcome level if multiple	How many experienced the outcome in group 5. Specify for each outcome level if multiple	How many experienced the outcome in group 6. Specify for each outcome level if multiple	How many experienced the outcome in group 7. Specify for each outcome level if multiple	How many experienced the outcome in group 8. Specify for each outcome level if multiple	How many experienced the outcome in group 9. Specify for each outcome level if multiple	How many experienced the outcome in group 10. Specify for each outcome level if multiple	If a +/- likelihood ratio (and confidence interval) has been reported, state this here for group 1. State for each outcome level	If a +/- likelihood ratio (and confidence interval) has been reported, state this here for group 2. State for each outcome level	If a +/- likelihood ratio (and confidence interval) has been reported, state this here for group 3. State for each outcome level
3																	

	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP
1								Chi-Squared Test									
2	If a +/- likelihood ratio (and confidence interval) has been reported, state this here for group 4. State for each outcome level	If a +/- likelihood ratio (and confidence interval) has been reported, state this here for group 5. State for each outcome level	If a +/- likelihood ratio (and confidence interval) has been reported, state this here for group 6. State for each outcome level	If a +/- likelihood ratio (and confidence interval) has been reported, state this here for group 7. State for each outcome level	If a +/- likelihood ratio (and confidence interval) has been reported, state this here for group 8. State for each outcome level	If a +/- likelihood ratio (and confidence interval) has been reported, state this here for group 9. State for each outcome level	If a +/- likelihood ratio (and confidence interval) has been reported, state this here for group 10. State for each outcome level	Was a chi-squared performed?	Specify the outcome and it's levels	What is blood component group 1?	What is blood component group 2?	What is blood component group 3?	What is blood component group 4?	What is blood component group 5?	What is blood component group 6?	What is blood component group 7?	What is blood component group 8?
3																	

	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG
1																	
2	What is blood component group 9?	What is blood component group 10?	How many patients are in group 1?	How many patients are in group 2?	How many patients are in group 3?	How many patients are in group 4?	How many patients are in group 5?	How many patients are in group 6?	How many patients are in group 7?	How many patients are in group 8?	How many patients are in group 9?	How many patients are in group 10?	Please specify the number who experienced the outcome in group 1. Specify for each outcome level if multiple	Please specify the number who experienced the outcome in group 2. Specify for each outcome level if multiple	Please specify the number who experienced the outcome in group 3. Specify for each outcome level if multiple	Please specify the number who experienced the outcome in group 4. Specify for each outcome level if multiple	Please specify the number who experienced the outcome in group 5. Specify for each outcome level if multiple
3																	

	CH	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX
1							T-Test or Mann-Whitney U test										
2	Please specify the number who experienced the outcome in group 6. Specify for each outcome level if multiple	Please specify the number who experienced the outcome in group 7. Specify for each outcome level if multiple	Please specify the number who experienced the outcome in group 8. Specify for each outcome level if multiple	Please specify the number who experienced the outcome in group 9. Specify for each outcome level if multiple	Please specify the number who experienced the outcome in group 10. Specify for each outcome level if multiple	What was the p-value from the chi-squared test	Was a t-test or Mann-Whitney U test performed?	How many t-tests/Mann-Whitney U tests were performed for this component?	Test 1. What is group 1?	Test 1. What is group 2?	Test 1. How many patients are in group 1	Test 1. How many patients are in group 2	Test 1. What is the mean in group 1	Test 1. What is the mean in group 2	Test 1. What is the mean difference, confidence interval and p-value from the t-test	Test 1. If Mann-Whitney U, was the median or rank presented?	Test 1. What is the median or rank in group 1. Provide median if both are reported.
3																	

	CY	CZ	DA	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK	DL	DM	DN	DO
1														Odds Ratio or Risk Ratio, with or without Logistic Regression			
2	Test 1. What is the median or rank in group 2. Provide median if both are reported.	Test 1. What is the p-value from the Mann-Whitney U test	Test 2. What is group 1?	Test 2. What is group 2?	Test 2. How many patients are in group 1	Test 2. How many patients are in group 2	Test 2. What is the mean in group 1	Test 2. What is the mean in group 2	Test 2. What is the mean difference, confidence interval and p-value from the t-test	Test 2. If Mann-Whitney U, was the median or rank presented?	Test 2. What is the median or rank in group 1. Provide median if both are reported.	Test 2. What is the median or rank in group 2. Provide median if both are reported.	Test 2. What is the p-value from the Mann-Whitney U test	Was an odds/risk ratio derived, either with or without the use of logistic regression?	If logistic regression was used, what type?	Specify the outcome and it's levels	Was a multivariable model developed?
3																	

	DP	DQ	DR	DS	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC	ED	EE	EF
1																	
2	Was the blood component included as a continuous or categorical variable in the logistic regression?	If continuous, state the number of patients	If continuous, state the number with the outcome. State for each outcome event if multiple	If continuous, state the coefficient, confidence interval and p-value. State for each outcome event if multiple	If categorical, what is blood component group 1 and how many patients are in this group?	If categorical, what is blood component group 2 and how many patients are in this group?	If categorical, what is blood component group 3 and how many patients are in this group?	If categorical, what is blood component group 4 and how many patients are in this group?	If categorical, what is blood component group 5 and how many patients are in this group?	If categorical, what is blood component group 6 and how many patients are in this group?	If categorical, what is blood component group 7 and how many patients are in this group?	If categorical, what is blood component group 8 and how many patients are in this group?	If categorical, what is blood component group 9 and how many patients are in this group?	If categorical, what is blood component group 10 and how many patients are in this group?	Please specify the number with the outcome in group 1. Specify for each outcome event if multiple	Please specify the number with the outcome in group 2. Specify for each outcome event if multiple	Please specify the number with the outcome in group 3. Specify for each outcome event if multiple
3																	

	EG	EH	EI	EJ	EK	EL	EM	EN	EO	EP	EQ	ER	ES	ET	EU	EV	EW
1																	
2	Please specify the number with the outcome in group 4. Specify for each outcome event if multiple	Please specify the number with the outcome in group 5. Specify for each outcome event if multiple	Please specify the number with the outcome in group 6. Specify for each outcome event if multiple	Please specify the number with the outcome in group 7. Specify for each outcome event if multiple	Please specify the number with the outcome in group 8. Specify for each outcome event if multiple	Please specify the number with the outcome in group 9. Specify for each outcome event if multiple	Please specify the number with the outcome in group 10. Specify for each outcome event if multiple	Which blood component group is the reference group for odds ratios?	Specify the odds/risk ratio, confidence interval and p-value for group 1. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 2. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 3. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 4. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 5. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 6. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 7. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 8. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 9. Specify for each outcome event if multiple
3																	

	EX	EY	EZ	FA	FB	FC	FD	FE	FF	FG	FH	FI	FJ	FK	FL	FM	FN
1																	
2	Specify the odds/risk ratio, confidence interval and p-value for group 10. Specify for each outcome event if multiple	If multivariable developed, please indicate here	If categorical, what is blood component group 1 and how many patients are in this group?	If categorical, what is blood component group 2 and how many patients are in this group?	If categorical, what is blood component group 3 and how many patients are in this group?	If categorical, what is blood component group 4 and how many patients are in this group?	If categorical, what is blood component group 5 and how many patients are in this group?	If categorical, what is blood component group 6 and how many patients are in this group?	If categorical, what is blood component group 7 and how many patients are in this group?	If categorical, what is blood component group 8 and how many patients are in this group?	If categorical, what is blood component group 9 and how many patients are in this group?	If categorical, what is blood component group 10 and how many patients are in this group?	Please specify the number with the outcome in group 1. Specify for each outcome event if multiple	Please specify the number with the outcome in group 2. Specify for each outcome event if multiple	Please specify the number with the outcome in group 3. Specify for each outcome event if multiple	Please specify the number with the outcome in group 4. Specify for each outcome event if multiple	Please specify the number with the outcome in group 5. Specify for each outcome event if multiple
3																	

	FO	FP	FQ	FR	FS	FT	FU	FV	FW	FX	FY	FZ	GA	GB	GC	GD	GE
1																	
2	Please specify the number with the outcome in group 6. Specify for each outcome event if multiple	Please specify the number with the outcome in group 7. Specify for each outcome event if multiple	Please specify the number with the outcome in group 8. Specify for each outcome event if multiple	Please specify the number with the outcome in group 9. Specify for each outcome event if multiple	Please specify the number with the outcome in group 10. Specify for each outcome event if multiple	Which blood component group is the reference group for odds ratios?	Specify the odds/risk ratio, confidence interval and p-value for group 1. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 2. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 3. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 4. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 5. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 6. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 7. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 8. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 9. Specify for each outcome event if multiple	Specify the odds/risk ratio, confidence interval and p-value for group 10. Specify for each outcome event if multiple	If multivariable, list all variables in the model. There is no need to write individual categories, just write the overall variable
3																	

	GF	GG	GH	GI	GJ	GK	GL	GM	GN	GO	GP	GQ	GR	GS	GT	GU	GV
1		Risk Difference								Hazard Ratio & Time-to-Event, with or without Cox/Fine-Gray Regression							
2	If multivariable provided, please also provide univariable here	Was the risk difference calculated?	What is group 1?	What is group 2?	How many patients are in group 1	How many patients are in group 2	What is the risk in group 1	What is the risk in group 2	What is the risk difference and confidence interval	Was a hazard ratio derived, either with or without the use of Cox or Fine-Gray (competing risks) regression?	What is the outcome and it's levels	Was a multivariable model developed?	Was the blood component included as a continuous or categorical variable in the logistic regression?	If continuous, state the number of patients	If continuous, state the number with the outcome. State for each outcome event if multiple	If continuous, state the coefficient, confidence interval and p-value. State for each outcome event if multiple	What is blood component group 1 and how many patients are in this group?
3																	

	GW	GX	GY	GZ	HA	HB	HC	HD	HE	HF	HG	HH	HI	HJ	HK	HL	HM
1																	
2	What is blood component group 2 and how many patients are in this group?	What is blood component group 3 and how many patients are in this group?	What is blood component group 4 and how many patients are in this group?	What is blood component group 5 and how many patients are in this group?	What is blood component group 6 and how many patients are in this group?	What is blood component group 7 and how many patients are in this group?	What is blood component group 8 and how many patients are in this group?	What is blood component group 9 and how many patients are in this group?	What is blood component group 10 and how many patients are in this group?	How many patients experienced the outcome in group 1, if reported? Specify for each outcome event	How many patients experienced the outcome in group 2, if reported? Specify for each outcome event	How many patients experienced the outcome in group 3, if reported? Specify for each outcome event	How many patients experienced the outcome in group 4, if reported? Specify for each outcome event	How many patients experienced the outcome in group 5, if reported? Specify for each outcome event	How many patients experienced the outcome in group 6, if reported? Specify for each outcome event	How many patients experienced the outcome in group 7, if reported? Specify for each outcome event	How many patients experienced the outcome in group 8, if reported? Specify for each outcome event
3																	

	HN	HO	HP	HQ	HR	HS	HT	HU	HV	HW	HX	HY	HZ	IA	IB	IC	ID
1																	
2	How many patients experienced the outcome in group 9, if reported? Specify for each outcome event	How many patients experienced the outcome in group 10, if reported? Specify for each outcome event	What is the median time-to-event value in group 1, if reported? Specify for each outcome level	What is the median time-to-event value in group 2, if reported? Specify for each outcome level	What is the median time-to-event value in group 3, if reported? Specify for each outcome level	What is the median time-to-event value in group 4, if reported? Specify for each outcome level	What is the median time-to-event value in group 5, if reported? Specify for each outcome level	What is the median time-to-event value in group 6, if reported? Specify for each outcome level	What is the median time-to-event value in group 7, if reported? Specify for each outcome level	What is the median time-to-event value in group 8, if reported? Specify for each outcome level	What is the median time-to-event value in group 9, if reported? Specify for each outcome level	What is the median time-to-event value in group 10, if reported? Specify for each outcome level	If performed, what is the log-rank test p-value for differences in time-to-event between groups?	Which group is the reference group for hazard ratios?	What is the hazard ratio, confidence interval and p-value in group 1. If Fine-Gray, specify for each event	What is the hazard ratio, confidence interval and p-value in group 2. If Fine-Gray, specify for each event	What is the hazard ratio, confidence interval and p-value in group 3. If Fine-Gray, specify for each event
3																	

	IE	IF	IG	IH	II	IJ	IK	IL	IM	IN	IO	IP	IQ	IR	IS	IT	IU
1											Receiver Operating Characteristic						
2	What is the hazard ratio, confidence interval and p-value in group 4. If Fine-Gray, specify for each event	What is the hazard ratio, confidence interval and p-value in group 5. If Fine-Gray, specify for each event	What is the hazard ratio, confidence interval and p-value in group 6. If Fine-Gray, specify for each event	What is the hazard ratio, confidence interval and p-value in group 7. If Fine-Gray, specify for each event	What is the hazard ratio, confidence interval and p-value in group 8. If Fine-Gray, specify for each event	What is the hazard ratio, confidence interval and p-value in group 9. If Fine-Gray, specify for each event	What is the hazard ratio, confidence interval and p-value in group 10. If Fine-Gray, specify for each event	If Cox, is a Kaplan-Meier curve presented?	If Fine-Gray, are cumulative incidence curves presented?	If multivariable, list the other variables were in the model?	Was a ROC analysis performed?	Specify the outcome and it's levels	How many patients were included in the analysis?	How many patients experienced the outcome?	Was a ROC curve reported?	If a cut-off for the blood component was derived, please state the cut-off here	If a cut-off for the blood component was derived, what method was used to derive it?
3																	

	IV	IW	IX	IY	IZ	JA	JB	JC	JD	JE	JF	JG	JH	JI	JJ	JK	JL
1												ANOVA or Kruskal-Wallis					
2	If other, please specify	What is the sensitivity at this cut-off?	What is the specificity at this cut-off?	What is the positive predictive value at this cut-off?	What is the negative predictive value at this cut-off?	What is the area under the curve and confidence interval at this cut-off?	Were specific sensitivity, specificity, positive or negative predictive values assessed?	If a specific sensitivity was assessed, what was it and corresponding blood component value/specificity ?	If a specific specificity was assessed, what was it and corresponding blood component value/sensitivity ?	If a specific PPV was assessed, what was it and corresponding blood component value/negative NVV?	If a specific NPV was assessed, what was it and corresponding blood component value/PPV?	Was ANOVA/Kruskal-Wallis performed?	Is ANOVA one-way or two-way?	What is group 1?	What is group 2?	What is group 3?	What is group 4?
3																	

	JM	JN	JO	JP	JQ	JR	JS	JT	JU	JV	JW	JX	JY	JZ	KA	KB	KC
1																	
2	What is group 5?	What is group 6?	What is group 7?	What is group 8?	What is group 9?	What is group 10?	How many patients are in group 1	How many patients are in group 2	How many patients are in group 3	How many patients are in group 4	How many patients are in group 5	How many patients are in group 6	How many patients are in group 7?	How many patients are in group 8?	How many patients are in group 9?	How many patients are in group 10?	What is the mean in group 1
3																	

	KD	KE	KF	KG	KH	KI	KJ	KK	KL	KM	KN	KO	KP	KQ	KR	KS	KT
1																	
2	What is the mean in group 2	What is the mean in group 3	What is the mean in group 4	What is the mean in group 5	What is the mean in group 6	What is the mean in group 7	What is the mean in group 8	What is the mean in group 9	What is the mean in group 10	If Kruskal-Wallis, are medians or ranks provided for each group?	What is the median/rank for group 1. Provide median if both are presented	What is the median/rank for group 2. Provide median if both are presented	What is the median/rank for group 3. Provide median if both are presented	What is the median/rank for group 4. Provide median if both are presented	What is the median/rank for group 5. Provide median if both are presented	What is the median/rank for group 6. Provide median if both are presented	What is the median/rank for group 7. Provide median if both are presented
3																	

	KU	KV	KW	KX	KY	KZ	LA	LB	LC	LD	LE	LF	LG	LH	LI
1								Correlation					Other Analysis Method		
2	What is the median/rank for group 8. Provide median if both are presented	What is the median/rank for group 9. Provide median if both are presented	What is the median/rank for group 10. Provide median if both are presented	What is the p-value from the test?	Were pairwise comparisons performed?	If pairwise comparisons performed, what type of adjustment was made for multiple testing?	If other, please specify	Was Pearson or Spearman correlation performed?	If yes, what other variable(s) were assessed with this blood component? List each if there were multiple	For each assessment, how many patients were included in this analysis?	For each assessment, specify the correlation coefficient and confidence interval, if reported	For each assessment, what was the p-value, if reported	Were other analyses performed to assess associations or differences of this blood component?	If so, what analysis was performed?	What were the main results from this test (give numbers and estimates if possible)?
3															

Form 1 (Aim 1) – Ratios between parameters

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	**Publication Details**				**Ratio Details**																
2	Who is the first author?	What is the title of the paper?	What is the year of publication?	If multiple rows are used for a single paper, please give a unique code to each	Was at least one ratio between multiple blood components analysed, e.g. platelet-to-lymphocyte count?	How many blood component ratios were analysed?	What is ratio 1	Ratio 1. What analysis is performed on this blood component?	Ratio 1. Please give the main result and estimates	What is ratio 2	Ratio 2. What analysis is performed on this blood component?	Ratio 2. Please give the main result and estimates	What is ratio 3	Ratio 3. What analysis is performed on this blood component?	Ratio 3. Please give the main result and estimates	What is ratio 4	Ratio 4. What analysis is performed on this blood component?	Ratio 4. Please give the main result and estimates	What is ratio 5	Ratio 5. What analysis is performed on this blood component?	Ratio 5. Please give the main result and estimates
3																					

Form 1 (Aim 1) – QUIPS assessment for risk of bias

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**Publication Details**				**Study Participation**							**Study Attrition**					**Prognostic Factor**
2	Who is the first author?	What is the year of publication?	What is the title of the paper?	If multiple rows are used for a single paper, please give a unique code to each	The population is adequately described for key characteristics	Sampling or recruitment are adequately described	Period of recruitment is adequately described	Place of recruitment (setting and geographic location) are adequately described	Inclusion and exclusion criteria are adequately described	There is adequate participation in the study by eligible individuals	The baseline study sample is adequately described for key characteristics	Response rate (i.e. proportion of study sample completing the study and providing outcome data) is adequate	Attempts to collect information on participants who dropped out of the study are described	Reasons for loss to follow-up are provided	Participants lost to follow-up are adequately described for key characteristics	There are no big differences in key characteristics or outcomes between those completing and not completing the study	A clear definition or description of predictor variables is provided
3																	

	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
1	**(i.e. Variable) Measurement**					**Outcome Measurement**			**Study Confounding**							**Statistical Analysis & Reporting**			
2	Method of predictor variable measurement is adequately valid and reliable to limit misclassification bias	Continuous predictor variables are reported or appropriate cut-points (i.e. not data-dependent) are used	The method and setting of measurement of predictor variables is the same for all study participants	Adequate proportion of the study sample has complete data for predictor variables	Appropriate methods of imputation are used for missing predictor variable data	A clear definition of the outcome is provided (e.g. including duration of follow-up and outcome levels)	The method of outcome measurement used is adequately valid and reliable to limit misclassification bias	The method and setting of outcome measurement is the same for all study participants	All important confounders are measured (e.g. treatments)	Clear definitions of the important confounders measured are provided	Measurement of all important confounders is adequately valid and reliable	The method and setting of confounding measurement are the same for all study participants	Appropriate methods are used if imputation is used for missing confounder data	Important potential confounders are accounted for in the study design (e.g. matching or stratification)	Important potential confounders are accounted for in the analysis	There is sufficient presentation of data to assess the adequacy of the analysis	The strategy for model building is appropriate, if relevant	The selected statistical model is adequate for the design of the study	There is no selective reporting of results
3																			

Form 2 (Aim 2) – Study design, analysis, and participant details overview:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**Publication Details**					**Study Design**											
2	Who is the first author?	What is the title of the paper?	What is the year of publication?	If multiple rows are used for a single paper, please give a unique code to each	If multiple rows for the same study, please indicate what each row is for	Is this a development/ internal validation study or external validation study?	If model development, what name has been given to the model, if any?	If external validation, what is the name of the model being validated?	If a model name is not reported, specify the first author, year of publication, and title of the paper developing the model	Is the data collected prospectively or retrospectively?	If other, please state	What is the type of study?	If other, please state	When did recruitment start? Specify month and year (if reported)	When did recruitment stop? Specify month and year (if reported)	What year did recruitment start?	What year did recruitment stop?
3																	

	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	**Study Setting**				**Participants**												
2	What is the geographical location of the data (i.e. the country)?	Where was the data obtained from?	If other, please state	If electronic record, please specify which database. List all if multiple are used. If a local centre(s), specify the centre(s)	Please copy and paste the patient population into this field, i.e. inclusion or exclusion criteria	How many patients were recruited, I.e. before exclusions?	How many patients were included in the study, i.e. after exclusions?	What is the patient setting?	If other, please state	Were patients symptomatic or asymptomatic?	If other, please state	What statistic is used to describe the age at study entry?	If other, please state	What is the value for the age at study entry? Provide for development and internal validation separately	What statistic is used to describe the age at diagnosis/censor date?	If other, please state	What is the value for the age at diagnosis/censor date? Provide for development and internal validation separately
3																	

	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY
1			**Blood Test**					**Analysis Overview**									
2	What proportion of patients were male? Provide for development and internal validation separately	What proportion of patients were female? Provide for development and internal validation separately	What is the total number of blood tests available, if reported?	Were multiple blood tests per patient used in the analysis?	If single, which blood test was chosen for analysis (assuming multuple were availale over time)?	If other, please specify	If multiple, describe how they were used and the analysis methods	Is the analysis stratified, and using what stratification variables?	If other, please specify stratification variables here	Is the analysis matched, and what are the matching variables?	If other, please specify matching variables here	How many controls were matched to each case?	How was missing data dealt with for the analysis of blood components?	If other, please specify	Were assumptions assessed for any of the (relevant) tests performed?	What significance level was used?	
3																	

Form 2 (Aim 2) – Development of prediction models:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**Publication Details**				**Methods Applied**												
2	Who is the first author?	What is the title of the paper?	What is the year of publication?	If multiple rows are used for a single paper, please give a unique code to each	How was the risk score developed?	If other, please specify	If statistical methods, please specify the method	If other, please specify	If logistic regression, are odds ratios or risk ratios reported for the final model?	If statistical methods, are coefficients presented for all predictors in the final model?	Was the full regression equation shown, including the intercept or baseline risk/hazard?	If machine learning, is the derived algorithm available to the public (e.g. reference to website)?	If machine learning, copy and paste the model building approach (i.e. classification method)?	What is the outcome?	If the outcome is categorical, what is outcome group 1?	If the outcome is categorical, what is outcome group 2?	If the outcome is categorical, what is outcome group 3?
3																	

	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1							**Model Summary Results**										
2	If the outcome is categorical, what is outcome group 4?	If the outcome is categorical, what is outcome group 5?	What is the outcome time frame, e.g. two-year risk of diagnosis, one-year incidence, etc.?	If a reason has been provided for the choice of the outcome time frame, please specify this here	How was variable selection performed?	If other, please specify	How many patients are included in the final model?	How many patients experienced outcome group 1 in the final model?	How many patients experienced outcome group 2 in the final model?	How many patients experienced outcome group 3 in the final model?	How many patients experienced outcome group 4 in the final model?	How many patients experienced outcome group 5 in the final model?	How many candidate predictors were there?	How many predictors were included in the final model?	List the predictors in the final model	How many full blood count components were in the final model?	Was Pearson or Spearman correlation performed, where at least one variable was a blood component?
3																	

	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW
1				**Model Apparent Performance**											
2	List the pair and the number of patients included. List for each combination if there were multiple pairs	List the pair, coefficient and confidence interval. List each for combination if there were multiple pairs	List the pair and p-value. List each for combination if there were multiple pairs	Was apparent (i.e. in the same sample used to derive the model) performance assessed?	Was overall performance assessed. If so, what method was used?	If other, please specify	What is the value and confidence interval, if reported?	Was apparent discrimination assessed? If so, what method was used?	If other, please specify	What is the value and confidence interval, if reported?	Was apparent calibration assessed? If so, what method was used?	If other, please specify	What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?	Was a ROC curve reported?	
3															

Form 2 (Aim 2) – Red blood cell count details in the model:

Note: this same form is replicated for the remaining 19 FBC parameters

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**Publication Details**				**Data Cleaning**				**Regression - Blood Component Results**								
2	Who is the first author?	What is the title of the paper?	What is the year of publication?	If multiple rows are used for a single paper, please give a unique code to each	Was red blood cell count analysed?	What is the range (min, max) of the blood component analysed?	If certain blood component values were deleted, please state which and why (copy and paste)?	Was data cleaning of the blood component described?	What is the unit of measurement used for the blood component?	How was this blood component included in the model?	If as-is, what is the coefficient (or weight)?	If as-is, what is the confidence interval?	If as-is, what is the p-value?	If categorised, what is the reference (blood component) group for model estimates?	If categorised, what is (blood component) group 1	If categorised, what is the coefficient (or weight) in (blood component) group 1?	If categorised, what is the confidence interval in (blood component) group 1?
3																	

	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1																	
2	If categorised, what is the p-value in (blood component) group 1?	If categorised, what is group 2	If categorised, what is the coefficient (or weight) in (blood component) group 2?	If categorised, what is the confidence interval in (blood component) group 2?	If categorised, what is the p-value in (blood component) group 2?	If categorised, what is group 3	If categorised, what is the coefficient (or weight) in (blood component) group 3?	If categorised, what is the confidence interval in (blood component) group 3?	If categorised, what is the p-value in (blood component) group 3?	If categorised, what is group 4	If categorised, what is the coefficient (or weight) in (blood component) group 4?	If categorised, what is the confidence interval in (blood component) group 4?	If categorised, what is the p-value in (blood component) group 4?	If categorised, what is group 5	If categorised, what is the coefficient (or weight) in (blood component) group 5?	If categorised, what is the confidence interval in (blood component) group 5?	If categorised, what is the p-value in (blood component) group 5?
3																	

	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY
1																	
2	If categorised, what is group 6	If categorised, what is the coefficient (or weight) in (blood component) group 6?	If categorised, what is the confidence interval in (blood component) group 6?	If categorised, what is the p-value in (blood component) group 6?	If categorised, what is group 7	If categorised, what is the coefficient (or weight) in (blood component) group 7?	If categorised, what is the confidence interval in (blood component) group 7?	If categorised, what is the p-value in (blood component) group 7?	If categorised, what is group 8	If categorised, what is the coefficient (or weight) in (blood component) group 8?	If categorised, what is the confidence interval in (blood component) group 8?	If categorised, what is the p-value in (blood component) group 8?	If categorised, what is group 9	If categorised, what is the coefficient (or weight) in (blood component) group 9?	If categorised, what is the confidence interval in (blood component) group 9?	If categorised, what is the p-value in (blood component) group 9?	If categorised, what is group 10
3																	

	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN
1															
2	If categorised, what is the coefficient (or weight) in (blood component) group 10?	If categorised, what is the confidence interval in (blood component) group 10?	If categorised, what is the p-value in (blood component) group 10?	If transformed, what was the transformation	If transformed, what is the coefficient (or weight) of the blood component?	If transformed, what is the confidence interval of the blood component?	If transformed, what is the p-value of the blood component?	If fractional polynomials, list all of the powers used to model the blood component	If fractional polynomials, list the coefficient, confidence interval and p-value of the blood component for power 1	If fractional polynomials, list the coefficient, confidence interval and p-value of the blood component for power 2	If fractional polynomials, list the coefficient, confidence interval and p-value of the blood component for power 3	If fractional polynomials, list the coefficient, confidence interval and p-value of the blood component for power 4	If fractional polynomials, list the coefficient, confidence interval and p-value of the blood component for power 5	If splines, specify each knot	If other, please specify
3															

Form 2 (Aim 2) – Ratios between parameters in the model

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**Publication Details**				**Ratio Details**												
2	Who is the first author?	What is the title of the paper?	What is the year of publication?	If multiple rows are used for a single paper, please give a unique code to each	Was at least one ratio between multiple blood components analysed, e.g. platelet-to-lymphocyte count?	How many blood component ratios were analysed?	What is ratio 1	Ratio 1. What analysis is performed on this blood component?	Ratio 1. Please give the main result and estimates	What is ratio 2	Ratio 1. What analysis is performed on this blood component?	Ratio 1. Please give the main result and estimates	What is ratio 3	Ratio 1. What analysis is performed on this blood component?	Ratio 1. Please give the main result and estimates	What is ratio 4	Ratio 1. What analysis is performed on this blood component?
3																	

R	S	T	U
Ratio 1. Please give the main result and estimates	What is ratio 5	Ratio 1. What analysis is performed on this blood component?	Ratio 1. Please give the main result and estimates

Form 2 (Aim 2) – Model validation (internal and external)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Publication Details				**Internal Validation**												
Who is the first author?	What is the title of the paper?	What is the year of publication?	If multiple rows are used for a single paper, please give a unique code to each	Was internal validation (i.e. in the same population but not the sample used to derive the model) performed?	How was the sample for internal validation obtained?	If non-random, please specify how the sample was chosen	If other, please specify	Was the model adjusted for optimism/shrinkage?	How many patients were included in the model validation analysis (i.e. after exclusions)?	Is the outcome used in the internal validation the same as that in the development stage?	If no, what is the outcome?	Is the outcome continuous or categorical?	If the outcome is categorical, what is outcome group 1	If the outcome is categorical, what is outcome group 2	If the outcome is categorical, what is outcome group 3	If the outcome is categorical, what is outcome group 4

R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
If the outcome is categorical, what is outcome group 5	How many outcome time frames are assessed during internal validation?	Time frame 1. What is the outcome time frame?	If a reason has been provided for the choice of the outcome time frame, please specify this here	Time frame 1. How many patients were there with the outcome?	Time frame 1. How many patients were there in outcome group 1?	Time frame 1. How many patients were there in outcome group 2?	Time frame 1. How many patients were there in outcome group 3?	Time frame 1. How many patients were there in outcome group 4?	Time frame 1. How many patients were there in outcome group 5?	Time frame 1. Was overall performance assessed. If so, what method was used?	Time frame 1. If other, please specify	Time frame 1. What is the value and confidence interval, if reported?	Time frame 1. Was discrimination assessed? If so, what method was used?	Time frame 1. If other, please specify	Time frame 1. What is the value and confidence interval, if reported?	Time frame 1. Was calibration assessed? If so, what method was used?

AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY
Time frame 1. If other, please specify	Time frame 1. What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?	Time frame 1. Was a ROC curve reported?	Time frame 2. What is the outcome time frame?	If a reason has been provided for the choice of the outcome time frame, please specify this here	Time frame 2. How many patients were there with the outcome?	Time frame 2. How many patients were there in outcome group 1?	Time frame 2. How many patients were there in outcome group 2?	Time frame 2. How many patients were there in outcome group 3?	Time frame 2. How many patients were there in outcome group 4?	Time frame 2. How many patients were there in outcome group 5?	Time frame 2. Was overall performance assessed. If so, what method was used?	Time frame 2. If other, please specify	Time frame 2. What is the value and confidence interval, if reported?	Time frame 2. Was discrimination assessed? If so, what method was used?	Time frame 2. If other, please specify	Time frame 2. What is the value and confidence interval, if reported?

AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP
Time frame 2. Was calibration assessed? If so, what method was used?	Time frame 2. If other, please specify	Time frame 2. What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?	Time frame 2. Was a ROC curve reported?	Time frame 3. What is the outcome time frame?	If a reason has been provided for the choice of the outcome time frame, please specify this here	Time frame 3. How many patients were there with the outcome?	Time frame 3. How many patients were there in outcome group 1?	Time frame 3. How many patients were there in outcome group 2?	Time frame 3. How many patients were there in outcome group 3?	Time frame 3. How many patients were there in outcome group 4?	Time frame 3. How many patients were there in outcome group 5?	Time frame 3. Was overall performance assessed. If so, what method was used?	Time frame 3. If other, please specify	Time frame 3. What is the value and confidence interval, if reported?	Time frame 3. Was discrimination assessed? If so, what method was used?	Time frame 3. If other, please specify

BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG
Time frame 3. What is the value and confidence interval, if reported?	Time frame 3. Was calibration assessed? If so, what method was used?	Time frame 3. If other, please specify	Time frame 3. What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?	Time frame 3. Was a ROC curve reported?	Time frame 4. What is the outcome time frame?	If a reason has been provided for the choice of the outcome time frame, please specify this here	Time frame 4. How many patients were there with the outcome?	Time frame 4. How many patients were there in outcome group 1?	Time frame 4. How many patients were there in outcome group 2?	Time frame 4. How many patients were there in outcome group 3?	Time frame 4. How many patients were there in outcome group 4?	Time frame 4. How many patients were there in outcome group 5?	Time frame 4. Was overall performance assessed. If so, what method was used?	Time frame 4. If other, please specify	Time frame 4. What is the value and confidence interval, if reported?	Time frame 4. Was discrimination assessed? If so, what method was used?

CH	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX
Time frame 4. If other, please specify	Time frame 4. What is the value and confidence interval, if reported?	Time frame 4. Was calibration assessed? If so, what method was used?	Time frame 4. If other, please specify	Time frame 4. What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?	Time frame 4. Was a ROC curve reported?	Time frame 5. What is the outcome time frame?	If a reason has been provided for the choice of the outcome time frame, please specify this here	Time frame 5. How many patients were there with the outcome?	Time frame 5. How many patients were there in outcome group 1?	Time frame 5. How many patients were there in outcome group 2?	Time frame 5. How many patients were there in outcome group 3?	Time frame 5. How many patients were there in outcome group 4?	Time frame 5. How many patients were there in outcome group 5?	Time frame 5. Was overall performance assessed. If so, what method was used?	Time frame 5. If other, please specify	Time frame 5. What is the value and confidence interval, if reported?

	CY	CZ	DA	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK	DL	DM	DN	DO
1								External validation									
2	Time frame 5. Was discrimination assessed? If so, what method was used?	Time frame 5. If other, please specify	Time frame 5. What is the value and confidence interval, if reported?	Time frame 5. Was calibration assessed? If so, what method was used?	Time frame 5. If other, please specify	Time frame 5. What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?	Time frame 5. Was a ROC curve reported?	Was external validation (i.e. in a different population to the one used to derive the model) performed?	How many patients are included in the model validation analysis (i.e. after exclusions)?	What is the outcome?	Is the outcome continuous or categorical?	If the outcome is categorical, what is outcome group 1	If the outcome is categorical, what is outcome group 2	If the outcome is categorical, what is outcome group 3	If the outcome is categorical, what is outcome group 4	If the outcome is categorical, what is outcome group 5	How many outcome time frames are assessed during external validation?
3																	

	DP	DQ	DR	DS	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC	ED	EE	EF
1																	
2	Time frame 1. What is the outcome time frame?	If a reason has been provided for the choice of the outcome time frame, please specify this here	Time frame 1. How many patients were there with the outcome?	Time frame 1. How many patients were there in outcome group 1?	Time frame 1. How many patients were there in outcome group 2?	Time frame 1. How many patients were there in outcome group 3?	Time frame 1. How many patients were there in outcome group 4?	Time frame 1. How many patients were there in outcome group 5?	Time frame 1. Was overall performance assessed. If so, what method was used?	Time frame 1. If other, please specify	Time frame 1. What is the value and confidence interval, if reported?	Time frame 1. Was discrimination assessed? If so, what method was used?	Time frame 1. If other, please specify	Time frame 1. What is the value and confidence interval, if reported?	Time frame 1. Was calibration assessed? If so, what method was used?	Time frame 1. If other, please specify	Time frame 1. What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?
3																	

	EG	EH	EI	EJ	EK	EL	EM	EN	EO	EP	EQ	ER	ES	ET	EU	EV	EW
1																	
2	Time frame 1. Was a ROC curve reported?	Time frame 2. What is the outcome time frame?	If a reason has been provided for the choice of the outcome time frame, please specify this here	Time frame 2. How many patients were there with the outcome?	Time frame 2. How many patients were there in outcome group 1?	Time frame 2. How many patients were there in outcome group 2?	Time frame 2. How many patients were there in outcome group 3?	Time frame 2. How many patients were there in outcome group 4?	Time frame 2. How many patients were there in outcome group 5?	Time frame 2. Was overall performance assessed. If so, what method was used?	Time frame 2. If other, please specify	Time frame 2. What is the value and confidence interval, if reported?	Time frame 2. Was discrimination assessed? If so, what method was used?	Time frame 2. If other, please specify	Time frame 2. What is the value and confidence interval, if reported?	Time frame 2. Was calibration assessed? If so, what method was used?	Time frame 2. If other, please specify
3																	

	EX	EY	EZ	FA	FB	FC	FD	FE	FF	FG	FH	FI	FJ	FK	FL	FM	FN
1																	
2	Time frame 2. What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?	Time frame 2. Was a ROC curve reported?	Time frame 3. What is the outcome time frame?	If a reason has been provided for the choice of the outcome time frame, please specify this here	Time frame 3. How many patients were there with the outcome?	Time frame 3. How many patients were there in outcome group 1?	Time frame 3. How many patients were there in outcome group 2?	Time frame 3. How many patients were there in outcome group 3?	Time frame 3. How many patients were there in outcome group 4?	Time frame 3. How many patients were there in outcome group 5?	Time frame 3. Was overall performance assessed. If so, what method was used?	Time frame 3. If other, please specify	Time frame 3. What is the value and confidence interval, if reported?	Time frame 3. Was discrimination assessed? If so, what method was used?	Time frame 3. If other, please specify	Time frame 3. What is the value and confidence interval, if reported?	Time frame 3. Was calibration assessed? If so, what method was used?
3																	

	FO	FP	FQ	FR	FS	FT	FU	FV	FW	FX	FY	FZ	GA	GB	GC	GD	GE
1																	
2	Time frame 3. If other, please specify	Time frame 3. What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?	Time frame 3. Was a ROC curve reported?	Time frame 4. What is the outcome time frame?	If a reason has been provided for the choice of the outcome time frame, please specify this here	Time frame 4. How many patients were there with the outcome?	Time frame 4. How many patients were there in outcome group 1?	Time frame 4. How many patients were there in outcome group 2?	Time frame 4. How many patients were there in outcome group 3?	Time frame 4. How many patients were there in outcome group 4?	Time frame 4. How many patients were there in outcome group 5?	Time frame 4. Was overall performance assessed. If so, what method was used?	Time frame 4. If other, please specify	Time frame 4. What is the value and confidence interval, if reported?	Time frame 4. Was discrimination assessed? If so, what method was used?	Time frame 4. If other, please specify	Time frame 4. What is the value and confidence interval, if reported?
3																	

	GF	GG	GH	GI	GJ	GK	GL	GM	GN	GO	GP	GQ	GR	GS	GT	GU	GV
1																	
2	Time frame 4. Was calibration assessed? If so, what method was used?	Time frame 4. If other, please specify	Time frame 4. What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?	Time frame 4. Was a ROC curve reported?	Time frame 5. What is the outcome time frame?	If a reason has been provided for the choice of the outcome time frame, please specify this here	Time frame 5. How many patients were there with the outcome?	Time frame 5. How many patients were there in outcome group 1?	Time frame 5. How many patients were there in outcome group 2?	Time frame 5. How many patients were there in outcome group 3?	Time frame 5. How many patients were there in outcome group 4?	Time frame 5. How many patients were there in outcome group 5?	Time frame 5. Was overall performance assessed. If so, what method was used?	Time frame 5. If other, please specify	Time frame 5. What is the value and confidence interval, if reported?	Time frame 5. Was discrimination assessed? If so, what method was used?	Time frame 5. If other, please specify
3																	

	GW	GX	GY	GZ	HA
1					
2	Time frame 5. What is the value and confidence interval, if reported?	Time frame 5. Was calibration assessed? If so, what method was used?	Time frame 5. If other, please specify	Time frame 5. What are the estimates (e.g. statistic, confidence interval, mean difference, p-value, etc.)?	Time frame 5. Was a ROC curve reported?
3					

Form 2 (Aim 2) – PROBAST assessment for risk of bias

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**Publication Details**				**Study Type**	**Participants**		**Predictors**			**Outcome**						**Analysis**
2	Who is the first author?	What is the title of the paper?	What is the year of publication?	If multiple rows are used for a single paper, please give a unique code to each	Is this a development, external validation, or development and external validation study?	Were appropriate data sources used, e.g. cohort, RCT or nested case-control study? data?	Were all inclusions and exclusions of participants appropriate?	Were predictors defined and assessed in a similar way for all participants?	Were predictor assessments made without knowledge of outcome data?	Are all predictors available at the time the model is intended to be used?	Was the outcome determined appropriately?	Was a pre-specified or standard outcome definition used?	Were predictors excluded from the outcome definition?	Was the outcome defined and determined in a similar way for all participants?	Was the outcome determined without knowledge of predictor information?	Was the time interval between predictor assessment and outcome determination appropriate?	Were there a reasonable number of participants with the outcome?
3																	

	R	S	T	U	V	W	X	Y
1								
2	Were continuous and categorical predictors handled appropriately?	Were all enrolled participants included in the analysis?	Were participants with missing data handled appropriately?	Was selection of predictors based on univariable analysis avoided?	Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?	Were relevant model performance measures evaluated appropriately?	Was model underfitting, overfitting and optimism in model performance accounted for?	Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?
3								

A3. Red blood cell count (Aim 1)

Table A1: Red blood cell count for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
China	Everyone	Wu 2019 (At diagnosis)	T-test	Yes, n=186 No, n=108		Mean=4.42 10^{12}/L (SD=0.63) Mean=4.73 10^{12}/L (SD=0.42)	<0.05
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 Polyp, n=132		Mean=4.42 10^{12}/L (SD=0.63) Mean=4.78 10^{12}/L (SD=0.72)	<0.05
		Wu 2019 (At diagnosis)	ANOVA	Yes, n=186 Polyp, n=132 Healthy, n=108		Mean=4.42 10^{12}/L Mean=4.78 10^{12}/L Mean=4.73 10^{12}/L	<0.001
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
	Males	Goshen 2017	T-test	Yes, n=936		Mean=4.76 10^{12}/L	<0.0001

		(1-6 months)		No, n=28491		Mean=4.87 10^{12}/L	
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=1.75 (95% CI=1.45, 2.24) Reference	
	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=4.48 10^{12}/L Mean=4.39 10^{12}/L	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=1.97 (95% CI=1.51, 2.61) Reference	
UK	Everyone	Schneider 2018 (6 months)	Odds ratio	Yes, n=5004 No, n=2117	<3.5 10^{12}/L, n=433, events=373 3.5-4.2 10^{12}/L, n=2507, events=1897 4.3-4.9 10^{12}/L, n=3316, events=2189 5-5.8 10^{12}/L, n=845, events=531 ≥5.9 10^{12}/L, n=20, events=14	OR=3.20 (95% CI=2.41, 4.24) OR=1.60 (95% CI=1.43, 1.8) Reference OR=0.87 (95% CI=0.74, 1.02) OR=1.20 (95% CI=0.46, 3.13)	
		Schneider 2018[1] (6 months)	Odds ratio	Yes, n=5004 No, n=2117	<3.5 10^{12}/L, n=433, events=373 3.5-4.2 10^{12}/L, n=2507, events=1897 4.3-4.9 10^{12}/L, n=3316, events=2189 5-5.8 10^{12}/L, n=845, events=531 ≥5.9 10^{12}/L, n=20, events=14	OR=4.33 (95% CI=3.15, 5.96) OR=1.89 (95% CI=1.65, 2.15) Reference OR=0.84 (95% CI=0.70, 1.00) X	

Males	Schneider 2018 (6 months)	Odds ratio	Yes, n=2806 No, n=466	<3.5 10^{12}/L, n=191, events=162 3.5-4.2 10^{12}/L, n=951, events=721 4.3-4.9 10^{12}/L, n=1608, events=1603 5-5.8 10^{12}/L, n=516, events=314 ≥5.9 10^{12}/L, n=6, events=6	OR=2.86 (95% CI=1.90, 4.31) OR=1.61 (95% CI=1.34, 1.93) Reference OR=0.80 (95% CI=0.65, 0.98) X	
	Schneider 2018[1] (6 months)	Odds ratio	Yes, n=2806 No, n=466	<3.5 10^{12}/L, n=191, events=162 3.5-4.2 10^{12}/L, n=951, events=721 4.3-4.9 10^{12}/L, n=1608, events=1603 5-5.8 10^{12}/L, n=516, events=314 ≥5.9 10^{12}/L, n=6, events=6	OR=3.72 (95% CI=2.36, 5.88) OR=1.93 (95% CI=1.57, 2.37) Reference OR=0.83 (95% CI=0.66, 1.04) X	
Females	Schneider 2018 (6 months)	Odds ratio	Yes, n=1897 No, n=865	<3.5 10^{12}/L, n=213, events=184 3.5-4.2 10^{12}/L, n=1302, events=960 4.3-4.9 10^{12}/L, n=1119, events=680 5-5.8 10^{12}/L, n=122, events=67 ≥5.9 10^{12}/L, n=6, events=6	OR=4.10 (95% CI=2.72, 6.17) OR=1.81 (95% CI=1.53, 2.15) Reference OR=0.79 (95% CI=0.21, 1.15) X	
	Schneider 2018[1] (6 months)	Odds ratio	Yes, n=1897 No, n=865	<3.5 10^{12}/L, n=213, events=184 3.5-4.2 10^{12}/L, n=1302, events=960	OR=5.68 (95% CI=3.55, 9.09) OR=1.94 (95% CI=1.60, 2.36)	

					4.3-4.9 10^{12}/L, n=1119, events=680 5-5.8 10^{12}/L, n=122, events=67 ≥5.9 10^{12}/L, n=6, events=6	Reference OR=0.75 (95% CI=0.49, 1.14) X	
6 < outcome time window ≤ 12 months:							
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as continuous*	OR=0.62 (95% CI=0.57, 0.67)	<0.001

Abbreviations: OR=odds ratio, RR=risk ratio.

[1]Multivariable effect estimate, adjusted for: BMI, smoking status, history of hypertension, diabetes, aspirin or NSAIDS use, vitamin K antagonists, platelet inhibitors.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A4. Haemoglobin (Aim 1)

Table A2: Haemoglobin levels for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
China	Everyone	Huang 2019 (At admission)	T-test	Yes, n=162 No, n=78		Mean=119.62 g/dL (SD=23.8) Mean=146.25 g/dL (SD=15.1)	<0.05
		Huang 2019 (At admission)	T-test	Yes, n=162 Polyp, n=92		Mean=119.62 g/dL (SD=23.8) Mean=134.1 g/dL (SD=16.1)	<0.05
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 No, n=108		Mean=121.27 g/L (SD=23.07) Mean=142.47 g/L (SD=11.80)	<0.05
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 Polyp, n=132		Mean=121.27 g/L (SD=23.07) Mean=132.12 g/L (SD=20.03)	<0.05
		Wu 2019 (At diagnosis)	ANOVA	Yes=186 Polyp=132		Mean=121.27 g/L Mean=132.12 g/L	<0.001

				Healthy=108		Mean=142.47 g/L	
		Yang 2018 (At admission)	Mann-Whitney U	Yes, n=85 Polyp, n=54		Median=122 g/L Median=131.5 g/L	0.004
Belgium	Everyone	Joosten 2008 (8 weeks)	Chi-squared	Yes, n=55 No, n=304	Men<13 g/dL, Women<12 g/dL, n=251, events=42 Men≥13 g/dL, Women≥12 g/dL, n=108, events=13		0.26
		Joosten 2008 (8 weeks)	T-test	Yes, n=55 No, n=304		Mean=10.2 g/dL (SD=2.9) Mean=10.8 g/dL (SD=2.7)	0.14
		Joosten 2008 (8 weeks)	Odds ratio	Yes, n=55 No, n=304	Men<13 g/dL, Women<12 g/dL, n=251, events=42 Men≥13 g/dL, Women≥12 g/dL, n=108, events=13	OR=1.47 (95% CI=0.75, 2.86) Reference	
		Joosten 2008[1] (8 weeks)	Odds ratio	Yes, n=55 No, n=304			≥0.05
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			

	Males	Goshen 2017 (1-6 months)	T-test	Yes, n=936 No, n=28491		Mean=13.30 g/dL Mean=14.43 g/dL	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=3.77 (95% CI=2.66, 5.08) Reference	
		Goshen 2017[2] (1-6 months)	Risk ratio	Yes, n=936 No, n=28491	Highest-risk quintile Lowest-risk quintile	RR=3.06 (95% CI=2.76, 3.52) Reference	
	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=11.80 g/dL Mean=13.02 g/dL	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=5.69 (95% CI=4.31, 7.97) Reference	
		Goshen 2017[3] (1-6 months)	Risk ratio	Yes, n=819 No, n=26239	Highest-risk quintile Lowest-risk quintile	RR=3.83 (95% CI=3.38, 4.46) Reference	
Netherlands	Everyone	van Boxtel-Wilms 2016 (3 months)	Descriptive	Yes No	Anaemia, n=5, events=5 No anaemia, n=545, events=0		
Spain	Everyone	Cubiella 2016[4] (1 week)	Odds ratio	Yes, n=214 No, n=1358	<10 g/dL 10-12 g/dL >12 g/dL	OR=4.8 (95% CI=2.2, 10.3) OR=1.8 (95% CI=1.1, 3.0) Reference	

Turkey	Everyone	Ay 2015 (1 week)	T-test	Yes, n=30 Polyp, n=110		Mean=13.5 g/dL (SD=1.1) Mean=13.9 g/dL (SD=1.1)	≥0.05
		Cakmak 2017 (6 months)	T-test	Yes, n=59 No, n=59		Mean=11.9 g/dL (SD=2.2) Mean=14.4 g/dL (SD=1.1)	<0.001
		Firat 2016 (At diagnosis)	Chi-squared	Yes No			0.002
		Firat 2016[5] (At diagnosis)	Machine-learning	Yes No			
		Kilincalp 2015 (At diagnosis)	T-test	Yes, n=144 No, n=143		Mean=11.6 g/dL (SD=2.20) Mean=14.2 g/dL (SD=1.17)	<0.001
UK	Everyone	Acher 2003[6] (6 months)	Descriptive	Yes No	<10.1 g/dl, n>5000, events=112 ≥10.1 g/dL, events=274		
		Mashlab 2018 (2 weeks)	Chi-squared	Yes, n=60 No, n=955	Men<130 g/L, Women<120 g/L, n=388, events=39 Men≥130 g/L, Women≥120 g/L, n=627, events=21		0.001

	Mashlab 2018 (2 weeks)	Odds ratio	Yes, n=60 No, n=955	Men<130 g/L, Women<120 g/L, n=388, events=39 Men≥130 g/L, Women≥120 g/L, n=627, events=21	OR=3.22 (95% CI=1.87, 5.57) Reference	
	Mashlab 2018[7] (2 weeks)	Odds ratio	Yes, n=60 No, n=955	Men<130 g/L, Women<120 g/L, n=388, events=39 Men≥130 g/L, Women≥120 g/L, n=627, events=21	OR=2.77 (95% CI=1.55, 4.95) Reference	
	Raje 2007[8] (1-2 months)	Descriptive	Yes No	Men<11 g/dL, Women <10 g/dL, n=142, events=9		
Centre A	Panagiotopoulou 2014 (3 months)	Chi-squared	Yes, n=30 No, n=199	Anaemia, n=105, events=16 No anaemia, n=124, events=14		0.434
	Panagiotopoulou 2014 (3 months)	Odds ratio	Yes, n=30 No, n=199	Anaemia, n=105, events=16 No anaemia, n=124, events=14	OR=1.4 (95% CI=0.7, 3.1) Reference	
Centre B	Panagiotopoulou 2014 (3 months)	Chi-squared	Yes, n=76 No, n=613	Anaemia, n=257, events=35 No anaemia, n=432, events=41		0.103

	Panagiotopoulou 2014 (3 months)	Odds ratio	Yes, n=76 No, n=613	Anaemia, n=257, events=35 No anaemia, n=432, events=41	OR=1.5 (95% CI=0.9, 2.4) Reference	
	Panagiotopoulou 2014[9] (3 months)	Odds ratio	Yes, n=76 No, n=613	Anaemia, n=257, events=35 No anaemia, n=432, events=41	OR=1.5 (95% CI=0.9, 2.5) Reference	
Everyone	Schneider 2018 (6 months)	Odds ratio	Yes, n=5528 No, n=2342	≤9 g/dL, n=668, events=660 9-9.9 g/dL, n=525, events=493 10-10.9 g/dL, n=742, events=663 11-11.9 g/dL, n=930, events=720 12-12.9 g/dL, n=1295, events=879 13-15.9 g/dL, n=3459, events=1982 ≥16 g/dL, n=251, events=131	OR=61.5 (95% CI=30.5, 123.9) OR=11.5 (95% CI=8.0, 16.5) OR=6.3 (95% CI=4.9, 8.0) OR=2.6 (95% CI=2.2, 3.0) OR=1.6 (95% CI=1.4, 1.8) Reference OR=0.8 (95% CI=0.6, 1.1)	
	Schneider 2018[10] (6 months)	Odds ratio	Yes, n=5528 No, n=2342	≤9 g/dL, n=668, events=660 9-9.9 g/dL, n=525, events=493 10-10.9 g/dL, n=742, events=663 11-11.9 g/dL, n=930, events=720 12-12.9 g/dL, n=1295, events=879 13-15.9 g/dL, n=3459, events=1982	OR=74.3 (95% CI=34.7, 158.9) OR=16.6 (95% CI=10.9, 25.2) OR=8.9 (95% CI=6.6, 12.0) OR=3.3 (95% CI=2.7, 4.0) OR=1.8 (95% CI=1.5, 2.1) Reference	

				≥16 g/dL, n=251, events=131	OR=0.8 (95% CI=0.6, 1.0)	
Males	Schneider 2018 (6 months)	Odds ratio	Yes, n=2551 No, n=1113	≤9 g/dL, n=243, events=243 9-9.9 g/dL, n=207, events=193 10-10.9 g/dL, n=284, events=255 11-11.9 g/dL, n=379, events=296 12-12.9 g/dL, n=497, events=384 13-15.9 g/dL, n=1834, events=1052 ≥16 g/dL, n=180, events=88	X OR=10.3 (95% CI=5.9, 17.8) OR=6.5 (95% CI=4.4, 9.7) OR=2.7 (95% CI=2.0, 3.4) OR=2.5 (95% CI=2.0, 3.2) Reference OR=0.71 (95% CI=0.52, 0.97)	
	Schneider 2018[10] (6 months)	Odds ratio	Yes, n=2551 No, n=1113	≤9 g/dL, n=243, events=243 9-9.9 g/dL, n=207, events=193 10-10.9 g/dL, n=284, events=255 11-11.9 g/dL, n=379, events=296 12-12.9 g/dL, n=497, events=384 13-15.9 g/dL, n=1834, events=1052 ≥16 g/dL, n=180, events=88	OR=95.9 (95% CI=23.5, 391.8) OR=12.2 (95% CI=6.8, 21.8) OR=8.6 (95% CI=5.3, 13.8) OR=3.1 (95% CI=2.3, 4.2) OR=2.9 (95% CI=2.2, 3.8) Reference OR=0.7 (95% CI=0.5, 1.04)	
Females	Schneider 2018 (6 months)	Odds ratio	Yes, n=2089 No, n=1086	≤9 g/dL, n=341, events=336 9-9.99 g/dL, n=368, events=252	OR=70.6 (95% CI=29, 172.2) OR=16.5 (95% CI=9.8, 27.8)	

					10-10.9 g/dL, n=379, events=333 11-11.9 g/dL, n=442, events=326 12-12.9 g/dL, n=667, events=365 13-15.9 g/dL, n=978, events=477 ≥16 g/dL, n=0	OR=7.6 (95% CI=5.5, 10.6) OR=3.0 (95% CI=2.3, 3.8) OR=1.5 (95% CI=1.2, 1.8) Reference X	
		Schneider 2018[10] (6 months)	Odds ratio	Yes, n=2089 No, n=1086	≤9 g/dL, n=341, events=336 9-9.99 g/dL, n=368, events=252 10-10.9 g/dL, n=379, events=333 11-11.9 g/dL, n=442, events=326 12-12.9 g/dL, n=667, events=365 13-15.9 g/dL, n=978, events=477 ≥16 g/dL, n=0	OR=84.6 (95% CI=30.4, 235.2) OR=23.3 (95% CI=12.4, 43.5) OR=10.6 (95% CI=6.9, 16.1) OR=3.7 (95% CI=2.7, 5.1) OR=1.5 (95% CI=1.2, 1.9) Reference X	
USA	Everyone	Spell 2004 (6 months)	Chi-squared	Yes, n=225 No, n=487	Men<13 g/dL, Women<11 g/dL, n=160, events=130 Men≥13 g/dL, Women≥11 g/dL, n=552, events=95		<0.001

		Spell 2004 (6 months)	Odds ratio	Yes, n=225 No, n=487	Men<13, Women<11 g/dL, n=160, events=130 Men≥13 g/dL, Women≥11 g/dL, n=552, events=95	OR=20.8 (95% CI=13.2, 32.8) Reference	
6 < outcome time window ≤ 12 months:							
UK	Everyone	Acher 2003[6] (6-12 months)	Descriptive	Yes No	<10.1 g/dl, n>5000, events=28 ≥10.1 g/dL, events=274		
		Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11491	*Modelled as continuous*	OR=0.67 (95% CI=0.66, 0.69)	<0.001
		Boursi 2016[11] (1 year)	Odds ratio	Yes, n=3375 No, n=8560	*Modelled as fractional polynomials (powers: 2, 2)*	OR=0.02*Hb^2 OR=32.17*Hb^2 × ln(Hb)	
		Hamilton 2009 (2 years)	Odds ratio	Yes, n=5477 No, n=38314	<12 g/dL, n=3227, events=1424 ≥12 g/dL,n=40564, events=4053	OR=7.11 (95% CI=6.59, 7.68) Reference	
		Hamilton 2009[12] (2 years)	Odds ratio	Yes No	<9 g/dL 9-9.9 g/dl 10-10.9 g/dl	OR=18 (95% CI=14, 25) OR=9.3 (95% CI=7.1, 12) OR=5.9 (95% CI=4.8, 7.2)	

					11-11.9 g/dl 12-12.9 g/dl ≥12 g/dL	OR=2.8 (95% CI=2.4, 3.2) OR=1.7 (95% CI=1.5, 1.9) Reference	
		Lawrenson 2006 (1 year)	Rate ratios	Yes No	Anaemia No anaemia		
		Marshall 2011 (2 years)	Odds ratio	Yes, n=5477 No, n=38314	<9 g/dL, n=487, events=385 9-9.999 g/dL, n=421, events=268 10-10.999 g/dL, n=771, events=354 11-11.999 g/dL, n=1548, events=417 12-12.999 g/dL, n=3001, events=517 13-13.999 g/dL, n=4284, events=573 ≥14 g/dL, n=33279, events=2963	OR=50.9 (95% CI=40.2, 64.5) OR=23.5 (95% CI=18.9, 29.1) OR=12.3 (95% CI=10.5, 14.4) OR=5.4 (95% CI=4.7, 6.1) OR=3.0 (95% CI=2.7, 3.3) OR=2.0 (95% CI=1.8, 2.2) Reference	
		Marshall 2011 (2 years)	Odds ratio	Yes, n=5477 No, n=38314	Men<12 g/dL, Women<11, n=2211, events=1181 Men≥12 g/dL, Women≥11, n=41580, events=4296	OR=11.4 (95% CI=10.3, 12.6) Reference	
		Marshall 2011[13]	Odds ratio	Yes, n=5477	<9 g/dL, n=487, events=385	OR=15.9 (95% CI=11.8, 21.6)	

		(2 years)		No, n=38314	9-9.999 g/dL, n=421, events=268 10-10.999 g/dL, n=771, events=354 11-11.999 g/dL, n=1548, events=417 12-12.999 g/dL, n=3001, events=517 13-13.999 g/dL, n=4284, events=573 ≥14 g/dL, n=33279, events=2963	OR=8.08 (95% CI=6.13, 10.65) OR=5.18 (95% CI=4.19, 6.39) OR=2.54 (95% CI=2.16, 2.99) OR=1.63 (95% CI=1.42, 1.87) OR=1.33 (95% CI=1.18, 1.50) Reference	
		Stapley 2006 (1 year)	Odds ratio	Yes No	<9 g/dL, n=73, events=39 ≥9 g/dL, n=276	 Reference	0.73
	Males	Hamilton 2008 (1 year)	Odds ratio	Yes, n=1604 No, n=5226	<9 g/dL, n=225, events=178 9-9.9 g/dL, n=167, events=118 10-10.9 g/dL, n=260, events=129 11-11.9 g/dL, n=464, events=171 12-12.9 g/dL, n=775, events=203 >12.9 g/dL, n=4939, events=805	OR=19.4 (95% CI=14.0, 27.0) OR=12.4 (95% CI=8.8, 17.4) OR=5.1 (95% CI=3.9, 6.5) OR=3.0 (95% CI=2.4, 3.7) OR=1.8 (95% CI=1.5, 2.2) Reference	
		Hippisley-Cox 2012[14] (2 years)	Hazard ratio	Yes No	<11 g/dL ≥11 g/dL	HR=3.33 (95% CI=2.86, 3.87) Reference	
		Hippisley-Cox 2013[15, 16]	Odds ratio	Yes, n=3250	<11 g/dL	OR=4.08 (95% CI=3.65, 4.57)	

		(2 years)		No, n=1240550	≥11 g/dL	Reference	
		Stapley 2006 (1 year)	Odds ratio	A B C D	10-12.9 g/dL, n=80, Stage A=3, B=3, C=11, D=10 ≥12.9 g/dL, n=269	OR=2.2 (95% CI=1.2, 4.3) Reference	0.021
	Females	Hamilton 2008 (1 year)	Odds ratio	Yes, n=1579 No, n=5226	<9 g/dL, n=257, events=221 9-9.9 g/dL, n=231, events=146 10-10.9 g/dL, n=451, events=226 11-11.9 g/dL, n=854, events=238 12-12.9 g/dL, n=1626, events=289 >12.9 g/dL, n=3451, events=459	OR= 40.0 (95% CI=27.8, 57.7) OR=14.0 (95% CI=3.3, 59.3) OR=6.6 (95% CI=5.3, 8.1) OR=2.5 (95% CI=2.1, 3.0) OR=1.4 (95% CI=1.2, 1.7) Reference	
		Hippisley-Cox 2012[17] (2 years)	Hazard ratio	Yes No	<11 g/dL ≥11 g/dL	HR=3.26 (95% CI=2.84, 3.74) Reference	
		Hippisley-Cox 2013[15, 16] (2 years)	Odds ratio	Yes, n=2607 No, n=1217648	<11 g/dL ≥11 g/dL	OR=4.37 (95% CI=3.94, 4.86) Reference	
12 < outcome time window ≤ 36 months:							

UK	Everyone	Cross 2019 (3 years)	Descriptive	Distal, n=240 Proximal, n=97 No, n=4405	Men<13 g/dL, Women<12 g/dL, n=1660, distal=106, proximal=78 Men≥13 g/dL, Women≥12 g/dL, n=3082, distal=134, proximal=19		
		Cross 2019 (3 years)	Odds ratio	Yes, n=337 No, n=4405	Men<13 g/dL, Women<12 g/dL, n=1660, events=184 Men≥13 g/dL, Women≥12 g/dL, n=3082, events=153	OR=2.39 (95% CI=1.91, 2.98) Reference	
		Cross 2019 (3 years)	Yield	Yes No	Anaemia with distal cancer Anaemia with proximal cancer No anaemia with distal cancer No anaemia with proximal cancer	Yield=6.4% Yield=4.7% Yield=4.3% Yield=0.6%	
		Hamilton 2005 (2 years)	Odds ratio	Yes, n=349 No, n=1744	<10 g/dl, n=61, events=40 10-11.9 g/dl, n=87, events=38 12-12.9 g/dl, n=37, events=17 ≥13 g/dL, n=1908, events=254	OR=12.4 (95% CI=7.2, 21.38) OR=5.05 (95% CI=3.24, 7.87) OR=5.5 (95% CI=2.7, 10.7) Reference	

		Hamilton 2005[18] (2 years)	Odds ratio	Yes, n=349 No, n=1744	<10 g/dl, n=61, events=40 10-11.9 g/dl, n=87, events=38 12-12.9 g/dl, n=37, events=17 ≥13 g/dL, n=1908, events=254	OR=13.0 (95% CI=6.2, 28.0) OR=4.3 (95% CI=2.1, 9.0) OR=2.5 (95% CI=0.95, 6.8) Reference	
		Thompson 2017 (3 years)	Odds ratio	Yes No	IDA No IDA	OR=6.09 (95% CI=5.04, 7.35) Reference	
		Thompson 2017[19] (3 years)	Odds ratio	Yes, n=990 No, n=16413	IDA No IDA	OR=8.38 (95% CI=5.10, 16.05) Reference	
Outcome time window > 36 months:							
Taiwan	Everyone	Hung 2015 (1-10 years)	Incidence ratio	Yes No	IDA, n=32390, events=171 No IDA	SIR=1.48 (95% CI=1.27, 1.72)	
		Hung 2015 (1-10 years)	Incidence ratio	Yes No	IDA, n=32390, events=54 No IDA	SIR=1.14 (95% CI=0.85, 1.48)	
UK	Everyone	Pilling 2018[20] (4.5 years)	Hazard ratio	Yes, n=914 No, n=237,302	*Modelled as continuous*	sHR=0.97 (95% CI=0.87, 1.08)	
		Pilling 2018[20] (4.5-9 years)	Hazard ratio	Yes, n=413 No, n=237,451	*Modelled as continuous*	sHR=1.01 (95% CI=0.87, 1.18)	

OUTCOME WINDOW NOT CATEGORISABLE: > 3-month risk of bowel cancer diagnosis:							
Iran	Everyone	Bafandeh 2008 (>3 months)	Odds ratio	Yes Polyp	Unexplained anaemia, n=35 No unexplained anaemia, n=445	 Reference	0.004
		Bafandeh 2008[21] (>3 months)	Odds ratio	Yes Polyp	Unexplained anaemia, n=35, events=5 No unexplained anaemia, n=445	 Reference	0.006
OUTCOME WINDOW NOT CATEGORISABLE: > 12-month risk of bowel cancer diagnosis:							
Netherlands	Everyone	Fijten 1995 (> 1 year)	Chi-squared	Yes, n=6 No, n=219	Men<8.5 mmol/L, Women<7.5 mmol/L, n=14, events=2 Men≥8.5 mmol/L, Women≥7.5 mmol/L, n=211, events=4		<0.01
		Fijten 1995 (> 1 year)	Odds ratio	Yes, n=6 No, n=219	Low, n=14, events=2 High, n=211, events=4	OR=8.6 (95% CI=1.4, 51.9) Reference	
UK	Everyone	Acher 2003[6] (> 1 year)	Descriptive	Yes No	<10.1 g/dl, n>5000, events=26 ≥10.1 g/dL, events=274		
Unspecified outcome time window:							
Israel	Males	Goldshtein 2010	Mixed model	Yes		Y_{cases} = 0.04ln(time)+13.11	

				No	*Haemoglobin was the outcome and time to diagnosis was modelled*	$Y_{controls}$ = 0.3ln(time)+12.72	
	Females	Goldshtein 2010	Mixed model	Yes No	*Haemoglobin was the outcome and time to diagnosis was modelled*	Y_{cases} = 0.09ln(time)+14.49 $Y_{controls}$ = 0.48ln(time)+14.12	
Italy	Everyone	Panzuto 2003[22]	Odds ratio	Yes, n=41 No, n=239	Men<14 g/dL, Women<12 g/dL, n=69, events=28 Men≥14 g/dL, Women≥12 g/dL, n=211, events=13	OR=10.4 (95% CI=4.9, 21.7) Reference	
		Panzuto 2003[23, 24]	Odds ratio	Yes, n=41 No, n=239	Men<14 g/dL, Women<12 g/dL, n=69, events=28 Men≥14 g/dL, Women≥12 g/dL, n=211, events=13	OR=8.8 (95% CI=3.9–19.8) Reference	
Japan	Everyone	Nakama 2000[25]	Chi-squared	Yes, n=96 No, n=17568	Men<12.5 g/dL, Women<11.5 g/dL, n=1132, events=31 Men≥12.5 g/dL, Women≥11.5 g/dL, n=16532, events=65		<0.05

		Nakama 2000[25]	Odds ratio	Yes, n=96 No, n=17568	Men<12.5 g/dL, Women<11.5 g/dL, n=1132, events=31 Men≥12.5 g/dL, Women≥11.5 g/dL, n=16532, events=65	OR=7.1 (95% CI=4.6, 11.0) Reference	
Switzerland	Everyone	Naef 1999	Descriptive	Yes Polyp	Anaemic, n=23, events=16 Non-anaemic, n=31		

Abbreviations: OR=odds ratio, RR=risk ratio, HR=hazard ratio, SIR=standardised incidence ratio, sHR=sub-distribution hazard ratio.

[1]Multivariable effect estimate, adjusted for age, sex, serum iron, transferrin, saturation index, and ferritin.

[2]Multivariable effect estimate, adjusted for mean corpuscular volume, neutrophil count, platelets, red blood cell distribution width, alanine aminotransferase, protein, iron, ferritin.

[3]Multivariable effect estimate, adjusted for mean corpuscular volume, monocyte count, platelets, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, iron, ferritin.

[4]Multivariable effect estimate, adjusted for change in bowl habit, rectal bleeding, benign anorrectal lesion, rectal mass, serum CEA, Faecal haemoglobin, previous colonoscopy, aspirin use, sex, age.

[5]Multivariable effect estimate, adjusted for platelets, sodium, total bilirubin, creatinine, calcium.

[6]In the presence of mean corpuscular volume<78 fL and/or mean corpuscular haemoglobin concentration <32 g/dL.

[7]Multivariable effect estimate, adjusted for age, sex.

[8]In the presence of serum ferritin<12 ng/ml and mean corpuscular volume<78 fL.

[9]Multivariable effect estimate, adjusted for sex, age, change in bowel habit, weight loss, bleeding per rectum, mucus per rectum, abdominal mass, abdominal fullness, lesion on digital rectal examination, anal lesion, abdominal distension, abdominal pain, family history, previous polyps, FOBt.

[10]Multivariable effect estimate, adjusted for BMI, smoking status, history of hypertension, diabetes, aspirin or NSAIDS use, vitamin K antagonists, platelet inhibitors.

[11]In the presence of mean corpuscular volume<78 fL and/or mean corpuscular haemoglobin concentration<32 g/dL.

[12]Multivariable effect estimate, adjusted for mean corpuscular volume, white blood cell count, neutrophil-lymphocyte ratio, platelets, sex, previous metformin prescriptions, previous prescriptions for oral hypoglycemic drugs other than metformin.

[13]Multivariable effect estimate, adjusted for rectal bleeding, change in bowel habit, abdominal pain, diarrhoea, constipation, weight loss, mean corpuscular volume.

[14]Multivariable effect estimate, adjusted for constipation, diarrhoea, change in bowel habit, flatulence, Irritable bowel syndrome, abdominal pain/antispasmodic, rectal bleeding, mean corpuscular volume, weight loss, deep venous thrombosis/pulmonary embolism, diabetes, obesity.

[15]Multivariable effect estimate, adjusted for alcohol status, family history of gastrointestinal cancer, current rectal bleeding, current abdominal pain, current appetite loss, current weight loss, change in bowel habit in previous year.
[16]Effect estimates are from multinomial logistic regression model, where the outcomes are different types of cancer. The estimates for the colorectal cancer vs no cancer are reported here.
[17]Multivariable effect estimate, adjusted for family history gastrointestinal cancer, alcohol status, abdominal distension, abdominal pain, appetite loss, rectal bleeding, weight loss, change in bowel habit, constipation.
[18]Multivariable effect estimate, adjusted for family history of gastrointestinal cancer, current rectal bleeding, current abdominal pain, current appetite loss, current weight loss.
[19]Multivariable effect estimate, adjusted for rectal bleeding, weight loss, number of episodes of abdominal pain, constipation, number of episodes of diarrhoea, rectal disease on rectal examination, tenderness on palpation of abdomen, positive faecal occult blood, blood sugar.
[20]Multivariable effect estimate, adjusted for age, sex, symptom combinations, physical signs, characteristics of rectal bleeding, characteristics of change in bowel habit, other characteristics of bowel cancer
[21]Multivariable effect estimate, adjusted for age, sex, smoking status, highest education level attained, mean corpuscular volume, red blood cell distribution width.
[22]Multivariable effect estimate, adjusted for age, sex, duration of symptoms.
[23]In the presence of ferritin<30 and mean corpuscular volume<80 fL
[24]Multivariable effect estimate, adjusted for age, weight loss.
[25]In the presence of serum ferritin<45.5 µg/L and serum iron<40 µg/L
* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A5. Haematocrit (Aim 1)

Table A3: Haematocrit for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
6 < outcome time window ≤ 12 months:							
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as continuous*	OR=0.97 (95% CI=0.95, 0.98)	<0.001
		Boursi 2016[1] (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as fractional polynomials (powers: -1, -1)*	OR=0.681*Hc^{-1} OR=0.894*Hc^{-1} × ln(Hc)	<0.001

Abbreviations: OR=odds ratio.

[1]Multivariable effect estimate, adjusted for: mean corpuscular volume, lymphocyte count, neutrophil-lymphocyte ratio.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A6. Mean corpuscular volume (Aim 1)

Table A4: Mean corpuscular volume for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
	Males	Goshen 2017 (1-6 months)	T-test	Yes, n=936 No, n=28491		Mean=85.7 fL Mean=88.9 fL	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=3.44 (95% CI=2.7, 4.87) Reference	
		Goshen 2017[1] (1-6 months)	Risk ratio	Yes, n=936 No, n=28491	Highest-risk quintile Lowest-risk quintile	RR=2.98 (95% CI=2.58, 3.42) Reference	<0.001
	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=84.5 fL Mean=88.6 fL	<0.0001

		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=3.52 (95% CI=2.84, 4.39) Reference	
		Goshen 2017[2] (1-6 months)	Risk ratio	Yes, n=819 No, n=26239	Highest-risk quintile Lowest-risk quintile	RR=3.04 (95% CI=2.7, 3.54) Reference	<0.001
Spain	Everyone	Cubiella 2016 (1 week)	Mann-Whitney U	Yes, n=214 No, n=1358		Median=89.1 fL Median=90.8 fL	<0.001
Turkey	Everyone	Ay 2015 (1 week)	T-test	Yes, n=30 Polyp, n=110		Mean=85.2 fL (SD=4.8) Mean=86.7 fL (SD=4.9)	≥0.05
UK	Everyone	Raje 2007[3] (1-2 months)	Descriptive	Yes No	<78 fL, n=142, events=9 ≥78 fL		
		Acher 2003[4] (6 months)	Descriptive	Yes No	<78 fL, n>5000, events-112 ≥78 fL, events-274		
		Schneider 2018 (6 months)	Odds ratio	Yes, n=544 No, n=3000	≤80 fL, n=1381, events=1322 81-85 fL, n=1140, events=926 86-95 fL, n=4064, events=2571 96-100 fL, n=100, events=468 >100 fL, n=241, events=127	OR=13.0 (95% CI=10.0, 17.0) OR=2.5 (95% CI=2.1, 3.0) Reference OR=0.7 (95% CI=0.6, 0.8) OR=0.7 (95% CI=0.5, 0.8)	

	Schneider 2018[5] (6 months)	Odds ratio	Yes, n=2457 No, n=1087	≤80 fL, n=1381, events=1322 81-85 fL, n=1140, events=926 86-95 fL, n=4064, events=2571 96-100 fL, n=100, events=468 >100 fL, n=241, events=127	OR=13.9 (95% CI=10.3, 18.9) OR=2.6 (95% CI=2.2, 3.1) Reference OR=0.6 (95% CI=0.5, 0.8) OR=0.6 (95% CI=0.5, 0.8)	
Males	Schneider 2018 (6 months)	Odds ratio	Yes, n=544 No, n=3000	≤80 fL, n=561, events=544 81-85 fL, n=440, events=364 86-95 fL, n=1944, events=1226 96-100 fL, n=475, events=260 >100 fL, n=124, events=63	OR=18.7 (95% CI=11.5, 30.6) OR=2.80 (95% CI=2.15, 3.65) Reference OR=0.7 (95% CI=0.6, 0.9) OR=0.6 (95% CI=0.4, 0.9)	
	Schneider 2018[5] (6 months)	Odds ratio	Yes, n=2457 No, n=1087	≤80 fL, n=561, events=544 81-85 fL, n=440, events=364 86-95 fL, n=1944, events=1226 96-100 fL, n=475, events=260 >100 fL, n=124, events=63	OR=25.5 (95% CI=13.9, 46.8) OR=2.8 (95% CI=2.1, 3.8) Reference OR=0.7 (95% CI=0.5, 0.8) OR=0.6 (95% CI=0.4, 0.9)	
Females	Schneider 2018 (6 months)	Odds ratio	Yes, n=2089 No, n=1086	≤80 fL, n=616, events=585 81-85 fL, n=512, events=409	OR=12.8 (95% CI=8.8, 18.7) OR=2.7 (95% CI=2.1, 3.4)	

					86-95 fL, n=1499, events=893 96-100 fL, n=280, events=127 >100 fL, n=82, events=42	Reference OR=0.6 (95% CI=0.4, 0.7) OR=0.7 (95% CI=0.5, 1.1)	
		Schneider 2018[6] (6 months)	Odds ratio	Yes, n=2056 No, n=933	≤80 fL, n=616, events=585 81-85 fL, n=512, events=409 86-95 fL, n=1499, events=893 96-100 fL, n=280, events=127 >100 fL, n=82, events=42	OR=11.4 (95% CI=7.6, 17.1) OR=2.8 (95% CI=2.1, 3.6) Reference OR=0.5 (95% CI=0.4, 0.6) OR=0.7 (95% CI=0.4, 1.1)	
	Centre B	Panagiotopoulou 2014 (3 months)	Odds ratio	Yes, n=17 No, n=672	<80 fL, n=106, events=17 ≥80 fL	OR=1.73 (95% CI=0.96, 3.1) Reference	
		Panagiotopoulou 2014[2] (3 months)	Odds ratio	Yes, n=76 No, n=613	<80 fL, n=106, events=17 ≥80 fL	OR=2.2 (95% CI=1.2, 4.1) Reference	
USA	Everyone	Spell 2004 (6 months)	Chi-squared	Yes, n=225 No, n=487	<80 fL, n=108, events=92 ≥80 fL, n=604, events=133		<0.001
		Spell 2004 (6 months)	Odds ratio	Yes, n=92 No, n=620	<80 fL, n=108, events=92 ≥80 fL, n=604, events=133	OR=20.4 (95% CI=11.6, 35.8) Reference	
6 < outcome time window ≤ 12 months:							

UK	Everyone	Acher 2003[4] (6-12 months)	Descriptive	Yes No	<78 fL, n>5000, events=28 ≥78 fL, events-274		
		Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as continuous*	OR=0.90 (95% CI=0.89, 0.91)	<0.001
		Boursi 2016[7] (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as fractional polynomials (powers: 3, 3)*	OR=0.933*MCV^3 OR=1.026*MCV^3 × ln(MCV)	
		Boursi 2016[8] (1 year)	Odds ratio	Yes, n=3375 No, n=8560	*Modelled as fractional polynomials (powers: 3, 3)*	OR=0.971*MCV^3 OR=1.010*MCV^3 × ln(MCV)	
		Hamilton 2008 (1 year)	Odds ratio	Yes, n=2951 No, n=9648	<80, n=974 >=80, n=11625	OR=15.7 (95% CI=13.4, 18.4) Reference	
12 < outcome time window ≤ 36 months:							
UK	Everyone	Marshall 2011 (2 years)	Odds ratio	Yes, n=5477 No, n=38314	<80, n=1045, events=761 80-84.999 fL, n=1306, events=444 ≥85 fL, n=41440, events=4272	OR=26.1 (95% CI=22.4, 30.4) OR=4.95 (95% CI=4.37, 5.61) Reference	
		Marshall 2011	Odds ratio	Yes, n=5477	<80 fL, n=1045, events=761	OR=23.3 (95% CI=20.0, 27.1)	

		(2 years)		No, n=38314	≥80 fL, n=42746, events=4716	Reference	
		Marshall 2011[9] (2 years)	Odds ratio	Yes, n=5477 No, n=38314	<80, n=1045, events=761 80-84.999 fL, n=1306, events=444 ≥85 fL, n=41440, events=4272	OR=7.67 (95% CI=6.23, 9.44) OR=2.71 (95% CI=2.30, 3.19) Reference	
		Hamilton 2009 (2 years)	Odds ratio	Yes, n=363 No, n=43428	<80, n=1286, events=363 ≥80 fL, n=42505, events=5114	OR=2.86 (95% CI=2.52, 3.24) Reference	
		Hamilton 2009[10] (2 years)	Odds ratio	Yes No	<80 fL ≥80 fL	OR=6.5 (95% CI=5.3, 7.9) Reference	
Outcome time window > 36 months:							
UK	Everyone	Pilling 2018[11] (4.5 years)	Hazard ratio	Yes, n=914 No, n=237,302	*Modelled as continuous*	sHR=0.98 (95% CI=0.96, 1.00)	
		Pilling 2018[11] (4.5-9 years)	Hazard ratio	Yes, n=413 No, n=237,451	*Modelled as continuous*	sHR=1.00 (95% CI=0.97, 1.04)	
OUTCOME WINDOW NOT CATEGORISABLE: > 12-month risk of bowel cancer diagnosis:							
UK	Everyone	Acher 2003[4] (> 1 year)	Descriptive	Yes No	<78 fL, n>5000, events=26 ≥78 fL, events=274		
Unspecified outcome time window:							

Italy	Everyone	Panzuto 2003[12, 13]	Odds ratio	Yes, n=41 No, n=170	<80 fL, n=69, events=28 ≥80 fL, n=211, events=13	OR=8.8 (95% CI=3.9–19.8) Reference	<0.001

Abbreviations: OR=odds ratio, RR=risk ratio, sHR=sub-distribution hazard ratio.

[1]Multivariable effect estimate, adjusted for: haemoglobin, neutrophil count, platelets, red blood cell distribution width, alanine aminotransferase, protein, iron, ferritin.

[2]Multivariable effect estimate, adjusted for: haemoglobin, monocyte count, platelets, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, iron, ferritin.

[3]In the presence of serum ferritin<12 ng/ml and haemoglobin<11 g/dL for males and <10 g/dL for females.

[4]In the presence of haemoglobin<10.1 g/dL and/or mean corpuscular haemoglobin concentration<32 g/dL.

[5]Multivariable effect estimate, adjusted for: BMI, smoking status, history of hypertension, diabetes, aspirin or NSAIDS use, vitamin K antagonists, platelet inhibitors.

[6]Multivariable effect estimate, adjusted for: sex, age, change in bowel habit, weight loss, bleeding per rectum, mucus per rectum, abdominal mass, abdominal fullness, lesion on digital rectal examination, anal lesion, abdominal distension, abdominal pain, family history, previous polyps, FOBt.

[7]Multivariable effect estimate, adjusted for: haematocrit, lymphocyte count, neutrophil-lymphocyte ratio.

[8]Multivariable effect estimate, adjusted for: haemoglobin, white blood cell count, neutrophil-lymphocyte ratio, platelets, sex, previous metformin prescriptions, previous prescriptions for oral hypoglycemic drugs other than metformin.

[9]Multivariable effect estimate, adjusted for: constipation, diarrhoea, change in bowel habit, flatulence, irritable bowel syndrome, abdominal pain/antispasmodic, rectal bleeding, haemoglobin, weight loss, deep venous thrombosis/pulmonary embolism, diabetes, obesity.

[10]Multivariable effect estimate, adjusted for: rectal bleeding, change in bowel habit, abdominal pain, diarrhoea, constipation, weight loss, haemoglobin.

[11]Multivariable effect estimate, adjusted for: age, sex, smoking status, highest education level attained, haemoglobin, red blood cell distribution width.

[12]In the presence of ferritin<30 and haemoglobin<14 g/dL for males and <12 g/dL for females.

[13]Multivariable effect estimate, adjusted for: age, weight loss.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A7. Mean corpuscular haemoglobin (Aim 1)

Table A5: Mean corpuscular haemoglobin for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
6 < outcome time window ≤ 12 months:							
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as continuous*	OR=0.75 (95% CI=0.74, 0.76)	<0.001

Abbreviations: OR=odds ratio.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A8. Mean corpuscular haemoglobin concentration (Aim 1)

Table A6: Mean corpuscular haemoglobin concentration for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
UK	Everyone	Acher 2003[1] (6 months)	Descriptive	Yes No	<32 g/dL, n>5000, events-112 ≥32 g/dL, events-274		
6 < outcome time window ≤ 12 months:							
UK	Everyone	Acher 2003[1] (6-12 months)	Descriptive	Yes No	<32 g/dL, n>5000, events=28 ≥32 g/dL, events-274		
OUTCOME WINDOW NOT CATEGORISABLE: > 12-month risk of bowel cancer diagnosis:							
UK	Everyone	Acher 2003[1] (> 1 year)	Descriptive	Yes No	<32 g/dL, n>5000, events=26 ≥32 g/dL, events=274		

[1]In the presence of haemoglobin<10.1 g/dL and/or mean corpuscular volume<78 g/dL.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A9. Red blood cell distribution width (Aim 1)

Table A7: Red blood cell distribution width for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
China	Everyone	Yang 2018 (At admission)	Mann-Whitney U	Yes, n=85 Polyp, n=54		Median=13.2% Median=12.6%	0.004
		Yang 2018 (At admission)	ROC	Yes, n=30 Polyp, n=110	13.25% (derived using Youden's index)	AUC=0.72 (95% CI=0.61, 0.83) Sensitivity=65.9% Specificity=75.6% PPV=81.2% NPV=58.6%	
		Shi 2019 (2 weeks)	T-test	Yes, n=211 Polyp, n=103		Mean=14.3% (SD=2.7) Mean=12.7% (SD=1.1)	<0.001
		Shi 2019	ROC	Yes, n=30	13.2% (derived using Youden's index)	AUC=0.72	

		(2 weeks)		Polyp, n=110		Sensitivity=53.1% Specificity=77.7% PPV=58.3% NPV=18.9%	
		Song 2018 (At diagnosis)	Mann-Whitney U	Yes, n=783 No, n=331		Median=13.3% Median=12.9%	<0.001
		Song 2018 (At diagnosis)	Mann-Whitney U	Yes, n=783 Polyp, n=463		Median=13.3% Median=13.0%	<0.05
		Song 2018 (At diagnosis)	ROC	Yes, n=783 No, n=331	13.95% (derived using Youden's index)	AUC=0.64 (95% CI=0.61, 0.67) Sensitivity=41% Specificity=94% PPV=94% NPV=40%	
		Song 2018 (At diagnosis)	ROC	Yes, n=30 Polyp, n=110	14.05% (derived using Youden's index)	AUC=0.50 (95% CI=0.47, 0.53) Sensitivity=29% Specificity=82% PPV=73%	

						NPV=41%	
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
	Males	Goshen 2017 (1-6 months)	T-test	Yes, n=936 No, n=28491		Mean=14.26% Mean=13.61%	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=2.87 (95% CI=2.23, 3.78) Reference	
	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=14.81% Mean=13.71%	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=4.54 (95% CI=3.58, 6.26) Reference	
		Goshen 2017[1] (1-6 months)	Risk ratio	Yes, n=819 No, n=26239	Highest-risk quintile Lowest-risk quintile	RR=3.14 (95% CI=2.81, 3.66) Reference	<0.0001
Turkey	Everyone	Ay 2015 (1 week)	T-test	Yes, n=30 Polyp, n=110		Mean=17.7% (SD=2.7) Mean=15.5% (SD=1.9)	0.02
		Ay 2015	ROC	Yes, n=30	17.5% (derived using unknown methods)	AUC=0.747	

		(1 week)		Polyp, n=110		Sensitivity=53.3% Specificity=91.4%	
		Cakmak 2017 (6 months)	T-test	Yes, n=59 No, n=59		Mean=16.1% (SD=3.4) Mean=13.6% (SD=0.6)	<0.001
		Cakmak 2017 (6 months)	ROC	Yes, n=59 No, n=59	14% (derived using unknown methods)	AUC=0.774 Sensitivity=68% Specificity=73%	
USA	Everyone	Spell 2004 (6 months)	Chi-squared	Yes, n=255 No, n=487	≥14.2%, n=213, events=156 <14.2%, n=499, events=69		<0.001
		Spell 2004 (6 months)	Odds ratio	Yes, n=156 No, n=556	≥14.2%, n=213, events=156 <14.2%, n=499, events=69	OR=17.1 (95% CI=11.5, 25.3) Reference	
Outcome time window > 36 months:							
UK	Everyone	Pilling 2018[2] (4.5 years)	Hazard ratio	Yes No	<12% ≥12.5-12.9% ≥13-13.4% ≥13.5-13.9% ≥14-14.4%	Reference sHR=1.25 (95% CI=0.90, 1.72) sHR=1.28 (95% CI=0.94, 1.75) sHR=1.55 (95% CI=1.33, 2.12) sHR=1.39 (95% CI=0.99, 1.97)	

					≥14-14.9%	sHR=1.88 (95% CI=1.26, 2.80)	
					≥15%	sHR=2.24 (95% CI=1.47, 3.40)	
		Pilling 2018[2]	Hazard ratio	Yes	<12%	Reference	
		(4.5-9 years)		No	≥12.5-12.9%	sHR=1.04 (95% CI=0.68, 1.59)	
					≥13-13.4%	sHR=1.23 (95% CI=0.82, 1.84)	
					≥13.5-13.9%	sHR=0.91 (95% CI=0.59, 1.40)	
					≥14-14.4%	sHR=1.13 (95% CI=0.70, 1.81)	
					≥14-14.9%	sHR=1.25 (95% CI=0.69, 2.24)	
					≥15%	sHR=1.46 (95% CI=0.76, 2.79)	

Abbreviations: OR=odds ratio, RR=risk ratio, sHR=sub-distribution hazard ratio, ROC=receiver operating characteristic, AUC=area under the curve, PPV=positive predictive value, NPV=negative predictive value.

[1]Multivariable effect estimate, adjusted for: haemoglobin, monocyte count, platelets, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, iron, ferritin.

[2]Multivariable effect estimate, adjusted for: age, sex, smoking status, highest education level attained, haemoglobin, red blood cell distribution width.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A10. Platelet count (Aim 1)

Table A8: Platelet count for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
China	Everyone	Wu 2019 (At diagnosis)	T-test	Yes, n=186 No, n=108		Mean=279.8 10^9/L (SD=80.56) Mean=207.83 10^9/L (SD=37.4)	<0.05
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 Polyp, n=132		Mean=279.8 10^9/L (SD=80.56) Mean=223.9 10^9/L (SD=42.59)	<0.05
		Wu 2019 (At diagnosis)	ANOVA	Yes=186 Polyp=132 Healthy=108		Mean=279.8 10^9/L (SD=80.56) Mean=223.9 10^9/L (SD=42.59) Mean=207.83 10^9/L (SD=37.4)	<0.001
		Yang 2018 (At admission)	Mann-Whitney U	Yes, n=85 Polyp, n=54		Median=219 10^9/L Median=201 10^9/L	0.021
		Zhu 2018[1]	T-test	Yes, n=783		Mean=272.4 10^9/L (SD=86.86)	<0.01

		(At diagnosis)		No, n=689		Mean=220 10^9/L	
		Zhu 2018[1] (At diagnosis)	T-test	Yes, n=783 Polyp, n=463		Mean=272.4 10^9/L (SD=86.86) Mean=216.67 10^9/L	<0.01
		Zhu 2018 (At diagnosis)	ROC	Yes, n=783 Polyp, n=689	242.5 10^9/L (derived using Youden's index)	AUC=0.71 (95% CI=0.68, 0.74) Sensitivity=62% Specificity=72% PPV=78.9% NPV=52.8%	
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
	Males	Goshen 2017 (1-6 months)	T-test	Yes, n=936 No, n=28491		Mean=261 10^9/L Mean=222 10^9/L	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=3.78 (95% CI=2.95, 4.88) Reference	
		Goshen 2017[2] (1-6 months)	Risk ratio	Yes, n=936 No, n=28491	Highest-risk quintile Lowest-risk quintile	RR=2.84 (95% CI=2.5, 3.27) Reference	
	Females	Goshen 2017	T-test	Yes, n=819		Mean=305 10^9/L	<0.0001

		(1-6 months)		No, n=26239		Mean=254 10^9/L	
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=3.87 (95% CI=3.09, 5.21) Reference	
		Goshen 2017[3] (1-6 months)	Risk ratio	Yes, n=819 No, n=26239	Highest-risk quintile Lowest-risk quintile	RR=2.95 (95% CI=2.56, 3.35) Reference	
Turkey	Everyone	Ay 2015 (1 week)	T-test	Yes, n=30 Polyp, n=110		Mean=287.7 /µL (SD=78.4) Mean=278.9 /µL (SD=59.6)	≥0.05
		Cakmak 2017 (6 months)	T-test	Yes, n=59 No, n=59		Mean=308.9 10^9/L (SD=99.1) Mean=243 10^9/L (SD=46.2)	<0.001
		Firat 2016 (At diagnosis)	Chi-squared	Yes No			0.001
		Firat 2016[4] (At diagnosis)	Machine-learning	Yes No			
		Kilincalp 2015 (At diagnosis)	T-test	Yes, n=144 No, n=143		Mean=280.8 10^9/L (SD=106.1) Mean=239.7 10^9/L (SD=50.7)	<0.001
		Kilincalp 2015	T-test	Yes (non-anaemic), n=44		Mean=262.4, 10^9/L (SD=79.3)	0.089

		(At diagnosis)		No, n=143		Mean=239.7 10^9/L (SD=50.7)	
6 < outcome time window ≤ 12 months:							
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11491	*Modelled as continuous*	OR=1.01 (95% CI=1.005, 1.01)	<0.001
		Boursi 2016[5] (1 year)	Odds ratio	Yes, n=3375 No, n=8560	*Modelled as fractional polynomials (powers: 2, 3)*	OR=605076.39*Plat^2 OR=0.00001*Plat^3	
		Ankus 2018 (1 year)	Descriptive	Yes, n=22 No, n=2697	325-349 10^9/L, n=1439, events=7 350-374 10^9/L, n=779, events=9 375-399, n=486, events=7		
		Ankus 2018 (1 year)	Odds ratio	Yes, n=22 No, n=2697	325-349 10^9/L, n=1439, events=7 350-374 10^9/L, n=779, events=9 375-399, n=486, events=7	Reference OR=2.39 (95% CI=0.89, 6.45)	
12 < outcome time window ≤ 36 months:							
UK	Everyone	Bailey 2017 (2 years)	Descriptive	Yes No	≤400 10^9/L, n=7969, events=627 >400 10^9/L, n=31261		

Abbreviations: OR=odds ratio, RR=risk ratio, ROC=receiver operating characteristic, AUC=area under the curve, PPV=positive predictive value, NPV=negative predictive value.

[1]Mean measured from graphs.
[2]Multivariable effect estimate, adjusted for: haemoglobin, mean corpuscular volume, neutrophil count, red blood cell distribution width, alanine aminotransferase, protein, iron, ferritin.
[3]Multivariable effect estimate, adjusted for: haemoglobin, mean corpuscular volume, monocyte count, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, iron, ferritin.
[4]Multivariable effect estimate, adjusted for: haemoglobin, sodium, total bilirubin, creatinine, calcium
[5]Multivariable effect estimate, adjusted for: haemoglobin, mean corpuscular volume, white blood cell count, neutrophil-lymphocyte ratio, sex, previous metformin prescriptions, previous prescriptions for oral hypoglycemic drugs other than metformin.
* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A11. Mean platelet volume (Aim 1)

Table A9: Mean platelet volume for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
China	Everyone	Wu 2019 (At diagnosis)	T-test	Yes, n=186 No, n=108		Mean=8.48 fL (SD=1.10) Mean=8.98 fL (SD=0.77)	<0.001
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 Polyp, n=132		Mean=8.48 fL (SD=1.10) Mean=8.83 fL (SD=0.90)	<0.05
		Wu 2019 (At diagnosis)	ANOVA	Yes, n=186 Polyp, n=132 Healthy, n=108		Mean=8.48 fL (SD=1.10) Mean=8.83 fL (SD=0.90) Mean=8.98 fL (SD=0.77)	<0.001
		Wu 2019 (At diagnosis)	ROC	Yes, n=186 Healthy, n=108		AUC=0.66 (95% CI=0.60, 0.71) Sensitivity=92.6% Specificity=44.6%	

						PPV=49.3% NPV=91.2%	
		Zhu 2018[1] (At diagnosis)	T-test	Yes, n=783 No, n=689		Mean=10 fL (SD=5.82) Mean=9.13 fL	<0.01
		Zhu 2018[1] (At diagnosis)	T-test	Yes, n=783 Polyp, n=463		Mean=10 fL (SD=5.82) Mean=9.2 fL	<0.01
		Zhu 2018 (At diagnosis)	ROC	Yes, n=783 Polyp, n=463	<9.25 fL optimal (calculated using Youden's index)	AUC=0.66 (95% CI=0.66, 0.69) Sensitivity=69% Specificity=59% PPV=74% NPV=52.9%	
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
	Males	Goshen 2017 (1-6 months)	T-test	Yes, n=936 No, n=28491		Mean=10.08 fL Mean=11.07 fL	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=2.33 (95% CI=1.8, 2.93) Reference	

	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=10.78 fL Mean=11.06 fL	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=2.33 (95% CI=1.72, 3.26) Reference	
Turkey	Everyone	Kilincalp 2015 (At diagnosis)	T-test	Yes (non-anaemic), n=44 No, n=143		Mean=8.23 fL (SD=0.98) Mean=7.82 fL (SD=0.49)	<0.001
		Kilincalp 2015 (At diagnosis)	T-test	Yes, n=144 No, n=143		Mean=8.41 fL (SD=0.98) Mean=7.82 fL (SD=0.49)	<0.001
		Kilincalp 2015 (At diagnosis)	ROC	Yes, n=144 No, n=143	<8.25 fL optimal (calculated using unknown method)	AUC=0.717 Sensitivity=54% Specificity=76%	

Abbreviations: RR=risk ratio, ROC=receiver operating characteristic, AUC=area under the curve, PPV=positive predictive value, NPV=negative predictive value.
[1]Mean measured from graphs.

A12. White blood cell count (Aim 1)

Table A10: White blood cell count for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
China	Everyone	Huang 2019 (At admission)	T-test	Yes, n=162 No, n=78		Mean=6.76 10^9/L (SD=1.68) Mean=6.42 10^9/L (SD=1.60)	≥0.05
		Huang 2019 (At admission)	T-test	Yes, n=162 Polyp, n=92		Mean=6.76 10^9/L (SD=1.68) Mean=6.25 10^9/L (SD=1.5)	<0.05
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 No, n=108		Mean=6.77 10^9/L (SD=1.64) Mean=6.23 10^9/L (SD=1.02)	<0.05
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 Polyp, n=132		Mean=6.77 10^9/L (SD=1.64) Mean=6.32 10^9/L (SD=1.61)	<0.05
		Wu 2019 (At diagnosis)	ANOVA	Yes=186 Polyp=132		Mean=6.77 10^9/L Mean=6.32 10^9/L	0.003

				Healthy=108		Mean=6.23 10^9/L	
		Zhou 2017 (At diagnosis)	Mann-Whitney U	Yes, n=242 No, n=248		Median=6.62 10^9/L Median=6.15 10^9/L	<0.001
		Zhou 2017 (At diagnosis)	Mann-Whitney U	Yes, n=242 Polyp, n=248		Median=6.62 10^9/L Median=6.22 10^9/L	<0.001
		Zhou 2017 (At diagnosis)	Kruskal-Wallis	Yes=242 Polyp=248 Healthy=262		Median=6.62 10^9/L Median=6.22 10^9/L Median=6.15 10^9/L	<0.001
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
	Males	Goshen 2017 (1-6 months)	T-test	Yes, n=936 No, n=28491		Mean=7.79 10^9/L Mean=7.20 10^9/L	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=2.31 (95% CI=1.87, 3.05) Reference	
	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=7.46 10^9/L Mean=6.65 10^9/L	<0.0001
		Goshen 2017	Risk ratio	Yes	Highest-risk quintile	RR=2.17 (95% CI=1.66, 3.02)	

		(1-6 months)		No	Lowest-risk quintile	Reference	
Turkey	Everyone	Firat 2016 (At diagnosis)	Chi-squared	Yes No			0.463
6 < outcome time window ≤ 12 months:							
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as continuous*	OR=1.11 (95% CI=1.09, 1.13)	<0.001
		Boursi 2016[1] (1 year)	Odds ratio	Yes, n=3375 No, n=8560	*Modelled as fractional polynomials (powers: 1, 1)*	OR=5.25*WBC[1] OR=0.30*WBC[1] × ln(WBC)	
Outcome time window ≥ 36 months:							
Korea	Males	Lee 2006 (10 years)	Odds ratio	Yes, n=1122 No, n=107785	≤5000 µL, n=18611, events=183 5501-6500 µL, n=24567, events=228 6501-7600 µL, n=28018, events=276 >7600 µL, n=37711, events=435	Reference OR=0.94 (95% CI=0.78, 1.15) OR=1.00 (95% CI=0.83, 1.21) OR=1.18 (95% CI=0.99, 1.4)	
	Males	Lee 2006[2] (10 years)	Hazard ratio	Yes, n=1122 No, n=107785	≤5000 µL, n=18611, events=183 5501-6500 µL, n=24567, events=228 6501-7600 µL, n=28018, events=276	Reference HR=0.95 (95% CI=0.78, 1.15) HR=1.02 (95% CI=0.84, 1.23)	

					>7600 µL, n=37711, events=435	HR=1.23 (95% CI=1.03, 1.47)	
	Females	Lee 2006 (10 years)	Odds ratio	Yes, n=1529 No, n=313983	≤5000, n=90790, events=405 5501-6500 µL, n=84260, events=400 6501-7600 µL, n=73364, events=353 >7600 µL, n=67098, events=371	Reference OR=1.06 (95% CI=0.93, 1.22) OR=1.08 (95% CI=0.94, 1.24) OR=1.24 (95% CI=1.08, 1.43)	
	Females	Lee 2006[2] (10 years)	Hazard ratio	Yes, n=1529 No, n=313983	≤5000, n=90790, events=405 5501-6500 µL, n=84260, events=400 6501-7600 µL, n=73364, events=353 >7600 µL, n=67098, events=371	Reference HR=1.03 (95% CI=0.90, 1.19) HR=1.03 (95% CI=0.89, 1.19) HR=1.15 (95% CI=0.99, 1.33)	
UK	Everyone	Margolis 2007[3] (11 years)	Hazard ratio	Yes, n=1341 No, n=142407	*Modelled as continuous*	HR=1.08 (95% CI=1.04, 1.12)	
USA	Everyone	Prizment 2011 (19 years)	Odds ratio	Yes, n=308 No, n=13106	≤4.8 10^9/L, n=3554, events=79 4.9-5.8 10^9/L, n=3413, events=65 5.9-7.0 10^9/L, n=3155, events=86 ≥7.1 10^9/L, n=3292, events=78	Reference OR=0.80 (95% CI=0.58, 1.12) OR=1.18 (95% CI=0.86, 1.60) OR=1.01 (95% CI=0.73, 1.38)	
		Prizment 2011[4] (19 years)	Hazard ratio	Yes, n=308 No, n=13106	≤4.8 10^9/L, n=3554, events=79 4.9-5.8 10^9/L, n=3413, events=65	Reference HR=0.86 (95% CI=0.61, 1.21)	

					5.9-7.0 10^9/L, n=3155, events=86 ≥7.1 10^9/L, n=3292, events=78	HR=1.26 (95% CI=0.91, 1.74) HR=1.13 (95% CI=0.79, 1.60)	
OUTCOME WINDOW NOT CATEGORISABLE: > 12-month risk of bowel cancer diagnosis:							
Netherlands	Everyone	Fijten 1995 (>12 months)	Chi-squared	Yes, n=4 No, n=215	Low, n=194, events=1 High, n=25, events=3		<0.01
		Fijten 1995 (>12 months)	Odds ratio	Yes, n=4 No, n=215	Low, n=194, events=1 High, n=25, events=3	Reference OR=26.3 (95% CI=2.6, 264.0)	

Abbreviations: OR=odds ratio, RR=risk ratio, HR=hazard ratio.

[1]Multivariable effect estimate, adjusted for: haemoglobin, mean corpuscular volume, neutrophil-lymphocyte ratio, platelet count, sex, previous metformin prescriptions, previous prescriptions for oral hypoglycemic drugs other than metformin.

[2]Multivariable effect estimate, adjusted for: age, BMI, total cholesterol, smoking status, regular exercise, alcohol consumption per day, frequency of meat intake per week, hypertension, diabetes.

[3]Multivariable effect estimate, adjusted for: age, ethnicity, smoking, alcohol use, physical activity, aspirin/nonsteroidal anti-inflammatory drug use, hormone therapy use, BMI, history of diabetes, family history of colorectal cancer.

[4]Multivariable effect estimate, adjusted for: age, race, centre, education, BMI, aspirin use, smoking status, pack-years of smoking, sex–HRT, diabetes.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A13. Basophil count (Aim 1)

Table A11: Basophil count for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
	Males	Goshen 2017 (1-6 months)	T-test	Yes, n=936 No, n=28491		Mean=0.03 10^9/L Mean=0.03 10^9/L	0.0017
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=1.4 (95% CI=1.14, 1.75) Reference	
	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=0.03 10^9/L Mean=0.03 10^9/L	0.0003
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=1.19 (95% CI=1.02, 1.48) Reference	

6 < outcome time window ≤ 12 months:							
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as continuous*	OR=1.34 (95% CI=0.93, 1.95)	0.12

Abbreviations: OR=odds ratio, RR=risk ratio.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A14. Basophil proportion (Aim 1)

Table A12: Basophil proportion for bowel cancer

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			

A15. Eosinophil count (Aim 1)

Table A13: Eosinophil count for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
	Males	Goshen 2017 (1-6 months)	T-test	Yes, n=936 No, n=28491		Mean=0.25 10^9/L Mean=0.22 10^9/L	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=1.62 (95% CI=1.29, 2.04) Reference	
	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=0.21 10^9/L Mean=0.18 10^9/L	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=2.03 (95% CI=1.58, 2.79) Reference	

6 < outcome time window ≤ 12 months:							
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as continuous*	OR=1.09 (95% CI=0.98, 1.2)	0.1

Abbreviations: OR=odds ratio, RR=risk ratio.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A16. Eosinophil proportion (Aim 1)

Table A14: Eosinophil proportion for bowel cancer

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			

A17. Lymphocyte count (Aim 1)

Table A15: Lymphocyte count for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
China	Everyone	Huang 2019 (At admission)	T-test	Yes, n=162 No, n=78		Mean=1.97 10^9/L (SD=0.57) Mean=2.03 10^9/L (SD=0.57)	≥0.05
		Huang 2019 (At admission)	T-test	Yes, n=162 Polyp, n=92		Mean=1.97 10^9/L (SD=0.57) Mean=1.98 10^9/L (SD=0.61)	≥0.05
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 No, n=108		Mean=1.99 10^9/L (SD=0.58) Mean=2.18 10^9/L (SD=0.51)	<0.05
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 Polyp, n=132		Mean=1.99 10^9/L (SD=0.58) Mean=1.99 10^9/L (SD=0.60)	≥0.05
		Wu 2019 (At diagnosis)	ANOVA	Yes=186 Polyp=132		Mean=1.99 10^9/L (SD=0.58) Mean=1.99 10^9/L (SD=0.60)	0.01

				Healthy=108		Mean=2.18 10^9/L (SD=0.51)	
		Yang 2018 (At admission)	Mann-Whitney U	Yes, n=85 Polyp, n=54		Median=1.6 10^9/L Median=1.7 10^9/L	0.526
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
	Males	Goshen 2017 (1-6 months)	T-test	Yes, n=936 No, n=28491		Mean=2.13 10^9/L Mean=2.21 10^9/L	0.026
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=1.17 (95% CI=1.01, 1.53) Reference	
	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=2.25 10^9/L Mean=2.21 10^9/L	0.43
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=1.37 (95% CI=1.06, 1.78) Reference	
6 < outcome time window ≤ 12 months:							
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as continuous*	OR=1.02 (95% CI=1.00, 1.05) Reference	0.03
		Boursi 2016[1]	Odds ratio	Yes, n=4929		OR=1.16*Lym^0	

		(1 year)		No, n=11311	*Modelled as fractional polynomial (powers: 0, 0.5)*	OR=12.23*Lym$^{0.5}$	

Abbreviations: OR=odds ratio, RR=risk ratio.

[1]Multivariable effect estimate, adjusted for: haematocrit, mean corpuscular volume, neutrophil-lymphocyte ratio.

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A18. Lymphocyte proportion (Aim 1)

Table A16: Lymphocyte proportion for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:						
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No		
China	Everyone	Zhou 2017 (At diagnosis)	Mann-Whitney U	Yes, n=242 No, n=262	Median=23.95% Median=35.15%	<0.001
		Zhou 2017 (At diagnosis)	Mann-Whitney U	Yes, n=242 Polyp, n=248	Median=23.95% Median=31.50%	<0.001
		Zhou 2017 (At diagnosis)	Kruskal-Wallis	Yes, n=242 Polyp, n=248 No, n=262	Median=23.95% Median=31.50% Median=35.15%	<0.001

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A19. Monocyte count (Aim 1)

Table A17: Monocyte count for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
China	Everyone	Wu 2019 (At diagnosis)	T-test	Yes, n=186 No, n=108		Mean=0.53 10^9/L (SD=0.19) Mean=0.45 10^9/L (SD=0.15)	<0.05
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 Polyp, n=132		Mean=0.53 10^9/L (SD=0.19) Mean=0.50 10^9/L (SD=0.17)	≥0.05
		Wu 2019 (At diagnosis)	ANOVA	Yes=186 Polyp=132 Healthy=108		Mean=0.53 10^9/L Mean=0.50 10^9/L Mean=0.45 10^9/L	0.001
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			
	Males	Goshen 2017	T-test	Yes, n=936		Mean=0.68 10^9/L	<0.0001

		(1-6 months)		No, n=28491		Mean=0.61 10^9/L	
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=2.11 (95% CI=1.74, 2.8) Reference	
		Goshen 2017[1] (1-6 months)	Risk ratio	Yes, n=936 No, n=28491	Highest-risk quintile Lowest-risk quintile	RR=1.85 (95% CI=1.6, 2.12) Reference	
	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=0.56 10^9/L Mean=0.51 10^9/L	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=1.99 (95% CI=1.63, 2.65) Reference	
6 < outcome time window ≤ 12 months:							
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as continuous*	OR=1.08 (95% CI=1.03, 1.14)	

Abbreviations: OR=odds ratio, RR=risk ratio.

[1]Multivariable effect estimate, adjusted for: haemoglobin, mean corpuscular volume, platelets, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, iron, ferritin

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A20. Monocyte proportion (Aim 1)

Table A18: Monocyte proportion for bowel cancer

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No			

A21. Neutrophil count (Aim 1)

Table A19: Neutrophil count for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:							
China	Everyone	Wu 2019 (At diagnosis)	T-test	Yes, n=186 No, n=108		Mean=3.92 10^9/L (SD=1.26) Mean=3.40 10^9/L (SD=0.79)	<0.05
		Wu 2019 (At diagnosis)	T-test	Yes, n=186 Polyp, n=132		Mean=3.92 10^9/L (SD=1.26) Mean=3.57 10^9/L (SD=1.26)	<0.05
		Wu 2019 (At diagnosis)	ANOVA	Yes=186 Polyp=132 Healthy=108		Mean=3.92 10^9/L Mean=3.57 10^9/L Mean=3.40 10^9/L	<0.001
		Yang 2018 (At admission)	Mann-Whitney U	Yes, n=85 Polyp, n=54		Median=3.6 10^9/L Median=3.2 10^9/L	0.136
Israel	Everyone	Kinar 2016		Yes, n=2437			

		(3-6 months)	Machine-learning	No			
	Males	Goshen 2017 (1-6 months)	T-test	Yes, n=936 No, n=28491		Mean=4.69 10^9/L Mean=4.13 10^9/L	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=2.29 (95% CI=1.73, 2.96) Reference	
		Goshen 2017[1] (1-6 months)	Risk ratio	Yes, n=936 No, n=28491	Highest-risk quintile Lowest-risk quintile	RR=2.03 (95% CI=1.82, 2.35) Reference	<0.0001
	Females	Goshen 2017 (1-6 months)	T-test	Yes, n=819 No, n=26239		Mean=4.33 10^9/L Mean=3.7 10^9/L	<0.0001
		Goshen 2017 (1-6 months)	Risk ratio	Yes No	Highest-risk quintile Lowest-risk quintile	RR=1.99 (95% CI=1.63, 2.65) Reference	
6 < outcome time window ≤ 12 months:							
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes, n=4929 No, n=11311	*Modelled as continuous*	OR=1.24 (95% CI=1.21, 1.27)	

Abbreviations: OR=odds ratio, RR=risk ratio.

[1]Multivariable effect estimate, adjusted for: haemoglobin, mean corpuscular volume, platelets, red blood cell distribution width, alanine aminotransferase, protein, iron, ferritin

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A22. Neutrophil proportion (Aim 1)

Table A20: Neutrophil proportion for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:						
Israel	Everyone	Kinar 2016 (3-6 months)	Machine-learning	Yes, n=2437 No		
China	Everyone	Zhou 2017 (At diagnosis)	Mann-Whitney U	Yes, n=242 No, n=262	Median=66.50% Median=56.75%	<0.001
		Zhou 2017 (At diagnosis)	Mann-Whitney U	Yes, n=242 Polyp, n=248	Median=66.50% Median=58.15%	<0.001
		Zhou 2017 (At diagnosis)	Kruskal-Wallis	Yes, n=242 Polyp, n=248 No, n=262	Median=66.50% Median=58.15% Median=56.75%	<0.001

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A23. Combination of parameters (Aim 1)

Table A21: Parameters combined for bowel cancer (analyses sorted by outcome time window, country, and then strata)*

Country	Strata	Article (Study outcome window)	Analysis	Outcome groups, no. per group	Parameter combination	Blood level categories and no. patients and events per group	Analysis estimates	P-value
0 < outcome time window ≤ 6 months:								
China	Everyone	Huang 2019 (At admission)	T-test	Yes No	Red blood cell distribution width-lymphocyte ratio		Mean=8.21 Mean=7.20	<0.05
		Huang 2019 (At admission)	T-test	Yes Polyp	Red blood cell distribution width-lymphocyte ratio		Mean=8.21 Mean=7.59	≥0.05
		Huang 2019 (At admission)	ROC	Yes No	Red blood cell distribution width-lymphocyte ratio	8.91 cut-off	AUC=0.57 (95% CI=0.73, 0.83) Sensitivity=41% Specificity=72%	
		Wu 2019 (At diagnosis)	T-test	Yes No	Mean platelet volume-platelet ratio		Mean=0.0330 Mean=0.0447	<0.05
		Wu 2019	T-test	Yes			Mean=0.0330	<0.05

		(At diagnosis)		Polyp	Mean platelet volume-platelet ratio		Mean=0.0411	
		Wu 2019 (At diagnosis)	ANOVA	Yes Polyp Healthy	Mean platelet volume-platelet ratio		Mean=0.0330 Mean=0.0411 Mean=0.0447	<0.001
		Wu 2019 (At diagnosis)	ROC	Yes No	Mean platelet volume-platelet ratio		AUC=0.81 (95% CI=0.76, 0.86)	
		Wu 2019 (At diagnosis)	T-test	Yes No	Neutrophil-lymphocyte ratio		Mean=1.98 Mean=1.57	<0.05
		Wu 2019 (At diagnosis)	T-test	Yes Polyp	Neutrophil-lymphocyte ratio		Mean=1.98 Mean=1.67	<0.05
		Wu 2019 (At diagnosis)	ANOVA	Yes Polyp Healthy	Neutrophil-lymphocyte ratio		Mean=1.98 Mean=1.67 Mean=1.57	<0.001
		Wu 2019 (At diagnosis)	ROC	Yes No	Neutrophil-lymphocyte ratio		AUC=0.67 (95% CI=0.62, 0.73)	
		Wu 2019	T-test	Yes	Platelet-lymphocyte ratio		Mean=140.26	<0.05

		(At diagnosis)		No			Mean=94.55	
		Wu 2019 (At diagnosis)	T-test	Yes Polyp	Platelet-lymphocyte ratio		Mean=140.26 Mean=113.03	<0.05
		Wu 2019 (At diagnosis)	ANOVA	Yes Polyp Healthy	Platelet-lymphocyte ratio		Mean=140.26 Mean=113.03 Mean=94.55	<0.001
		Wu 2019 (At diagnosis)	ROC	Yes No	Platelet-lymphocyte ratio		AUC=0.78 (95% CI=0.73, 0.82)	
		Yang 2018 (At admission)	Mann-Whitney U	Yes Polyp	Neutrophil-lymphocyte ratio		Median=2.08 Median=1.87	0.091
		Yang 2018 (At admission)	Mann-Whitney U	Yes Polyp	Platelet-lymphocyte ratio		Median=124.48 Median=113.19	0.059
		Zhou 2017 (At diagnosis)	Mann-Whitney U	Yes No	Neutrophil-lymphocyte ratio		Median=2.76 Median=1.60	<0.001
		Zhou 2017 (At diagnosis)	Mann-Whitney U	Yes Polyp	Neutrophil-lymphocyte ratio		Median=2.76 Median=1.875	<0.001
		Zhou 2017		Yes			Median=2.76	<0.001

		(At diagnosis)	Kruskal-Wallis	Polyp Healthy	Neutrophil-lymphocyte ratio		Median=1.875 Median=1.60	
		Zhou 2017 (At diagnosis)	ROC	Yes No	Neutrophil-lymphocyte ratio	2.33 cut-off	Sensitivity=66.9% Specificity=77.6%	
Turkey	Everyone	Cakmak 2017 (6 months)	T-test	Yes No	Neutrophil-lymphocyte ratio		Mean=2.9 Mean=2.0	<0.001
		Cakmak 2017 (6 months)	ROC	Yes No	Neutrophil-lymphocyte ratio	2.05 cut-off	AUC=0.740 Sensitivity=78% Specificity=66%	
		Cakmak 2017 (6 months)	T-test	Yes No	Platelet-lymphocyte ratio		Mean=163.6 Mean=118.5	<0.001
		Cakmak 2017 (6 months)	ROC	Yes No	Platelet-lymphocyte ratio	130 cut-off	AUC=0.702 Sensitivity=65% Specificity=72%	
		Kilincalp 2015 (At diagnosis)	T-test	Yes No	Neutrophil-lymphocyte ratio		Mean=6.1 Mean=1.5	<0.001
		Kilincalp 2015	ROC	Yes		2.02 cut-off	AUC=0.921	

		(At diagnosis)		No	Neutrophil-lymphocyte ratio		Sensitivity=86% Specificity=84%	
		Kilincalp 2015 (At diagnosis)	T-test	Yes No	Platelet-lymphocyte ratio		Mean=230.5 Mean=106.3	<0.001
		Kilincalp 2015 (At diagnosis)	ROC	Yes No	Platelet-lymphocyte ratio	135 cut-off	AUC=0.854 Sensitivity=70% Specificity=90%	
6 < outcome time window ≤ 12 months:								
UK	Everyone	Boursi 2016 (1 year)	Odds ratio	Yes No	Neutrophil-lymphocyte ratio	*Modelled as continuous*	OR=1.21 (95% CI=1.18, 1.24)	<0.001
		Boursi 2016[1] (1 year)	Odds ratio	Yes No	Neutrophil-lymphocyte ratio	*Modelled as fractional polynomials (powers: -0.5, -0.5)*	$OR=0.11*NLR^{-0.5}$ $OR=0.70*NLR^{-0.5} \times \ln(NLR)$	
		Boursi 2016[2] (1 year)	Odds ratio	Yes No	Neutrophil-lymphocyte ratio	*Modelled as fractional polynomials (powers: -0.5, -0.5)*	$OR=0.31*NLR^{-0.5}$ $OR=0.77*NLR^{-0.5} \times \ln(NLR)$	
	Centre A	Panagiotopoulou 2014		Yes		MCV <80 fL & anaemia		0.285

		(3 months)	Chi-squared	No	Mean platelet volume & haemoglobin	MCV ≥80 fL & no anaemia		
		Panagiotopoulou 2014 (3 months)	Chi-squared	Yes No	Mean platelet volume & haemoglobin	MCV <80 fL & anaemia MCV ≥80 fL & anaemia		0.781
		Panagiotopoulou 2014 (3 months)	Odds ratio	Yes No	Mean platelet volume & haemoglobin	MCV <80 fL & anaemia MCV ≥80 fL & anaemia	OR=1.3 (95% CI=0.5, 3.9)	
	Centre B	Panagiotopoulou 2014 (3 months)	Chi-squared	Yes No	Mean platelet volume & haemoglobin	MCV <80 fL & anaemia MCV ≥80 fL & no anaemia		0.285 0.019
		Panagiotopoulou 2014 (3 months)	Chi-squared	Yes No	Mean platelet volume & haemoglobin	MCV <80 fL & anaemia MCV ≥80 fL & anaemia		0.196
		Panagiotopoulou 2014 (3 months)	Odds ratio	Yes No	Mean platelet volume & haemoglobin	MCV <80 fL & anaemia MCV ≥80 fL & anaemia	OR=1.6 (95% CI=0.8, 3.3)	

* This table is taken and adapted from: Virdee *et al* (2020). The Full Blood Count Blood Test for Colorectal Cancer Detection: A Systematic Review, Meta-Analysis, and Critical Appraisal. *Cancers, 12, 2348: 1-37*. DOI: 10.3390/cancers12092348[51]. Published under a CC-BY licence – permissions not required.

A24. Development details of FBC-based prediction models (Aim 2)

Table A22: Development details of FBC-based prediction models

Article	Model name/description	Data source	Modelling type	Variable selection approach	No. FBC parameters in model	Correlation assessed and method
Boursi 2016	Laboratory model	Electronic records	Logistic regression	Significant in univariable & backward selection	4	Yes - unclear
Boursi 2016	Combined model	Electronic records	Logistic regression	Significant in univariable & backward selection	6	Yes - unclear
Cubiella 2016	COLONPREDICT	Centres	Logistic regression	Significant in univariable & backward selection	1	Yes - unclear
Firat 2016		Electronic records	Machine-learning	Significant in univariable	2	No
Goshen 2017	Model for males	Electronic records	Logistic regression	Significant in univariable	4	Yes - unclear
Goshen 2017	Model for females	Electronic records	Logistic regression	Significant in univariable	5	Yes - unclear
Hippisley-Cox 2012	QCancer Colorectal males	Electronic records	Cox regression	Backward selection	1	No
Hippisley-Cox 2012	QCancer Colorectal females	Electronic records	Cox regression	Backward selection	1	No
Hippisley-Cox 2013	QCancer females	Electronic records	Logistic regression	Backward selection	1	No
Hippisley-Cox 2013	QCancer males	Electronic records	Logistic regression	Backward selection	1	No
Kinar 2016	ColonFlag	Electronic records	Machine-learning	All candidate variables	20	No

Marshall 2011	Bristol-Birmingham	Electronic records	Logistic regression	Significant in univariable & backward selection	2	No
Thompson 2017		Unclear	Logistic regression	Forward & backward selection	1	No

Abbreviations: FBC-full blood count

Appendix B

Appendices for Chapter 3 (methodology to assess EHR data quality)

Most of the contents in this appendix have previously been published (in supplementary material)[128]:

Virdee P.S. *et al*. Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96* (2020). DOI: 10.1186/s40537-020-00375-w. Published under a CC-BY licence – permissions not required.

B1. Permission for re-use of work

Figure 3.3

Request:

From: Pradeep Virdee <pradeep.virdee@csm.ox.ac.uk>
Sent: 01 August 2021 20:36
To: Platt Andrew (RTH) OUH <Andrew.Platt@ouh.nhs.uk>
Subject: OUH Haematology Laboratory Handbook - request

Dear Andrew,

I'm a DPhil student at Oxford currently writing my thesis. You're department's OUH Haematology Laboratory Handbook has been quite useful for me to understand the FBC blood test. I was hoping to use the image of the haematology request form example in Appendix 3 of your handbook in my thesis.

Handbook link:
https://www.ouh.nhs.uk/services/departments/laboratory-medicine/haematology-laboratories/documents/haematology-handbook.pdf

I'm wondering whether there is an copyright for this document on re-use of images? Could I please request your permission for me to reuse this image by including it in my thesis? It will not be used for anything else.

Pradeep

Mr Pradeep Virdee
NIHR Doctoral Research Fellow; Medical Statistician
Centre for Statistics in Medicine (CSM),
Botnar Research Centre, NDORMS,
Nuffield Orthopaedic Centre,
University of Oxford,
Windmill Road,
Oxford,
OX3 7LD

Tel: +44 (0)1865 737950
Email: pradeep.virdee@csm.ox.ac.uk
Web: https://www.ndorms.ox.ac.uk/team/pradeep-virdee

Permission:



Sun 01/08/2021 22:53

Platt Andrew (RTH) OUH <Andrew.Platt@ouh.nhs.uk>

RE: OUH Haematology Laboratory Handbook - request

To Pradeep Virdee

You replied to this message on 02/08/2021 10:31.

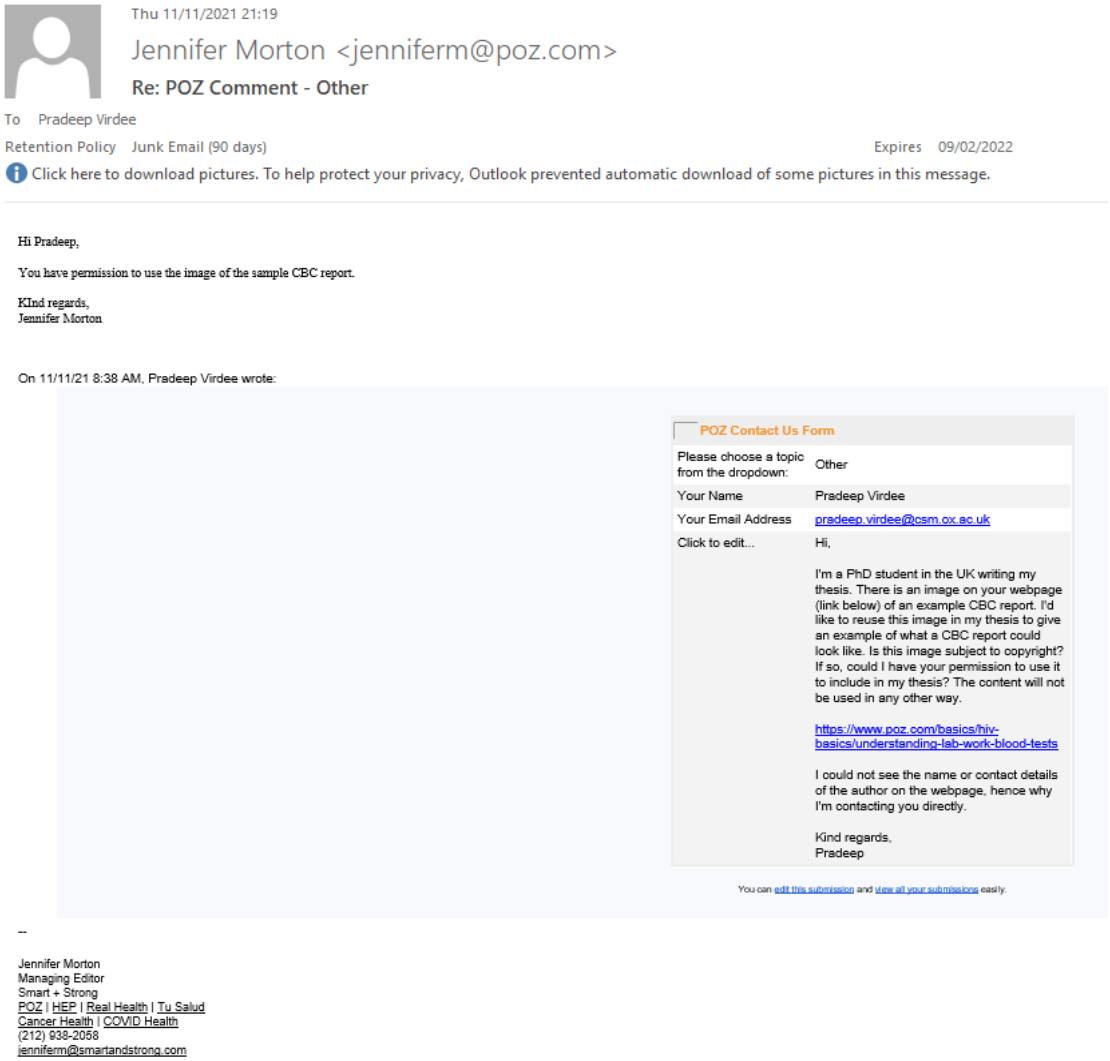
Dear Pradeep,
I would not consider that the image is copyright, so yes, feel free to use it. I would however ask you to remember that it is an example of a paper request form, this method of ordering tests is now not as common as it was, with electronic methods such as EPR becoming more frequently used.
Bw
Andrew

Andrew Platt,
Quality Manager, Training Lead
The Department of Laboratory Haematology & Clinical Biochemistry
Oxford University Hospitals NHS Foundation Trust

Tel = 01865 (8)57663
email = andrew.platt@ouh.nhs.uk <mailto:andrew.platt@ouh.nhs.uk>

For emails involving confidential information please use = andrew.platt@nhs.net

Figure 3.4



Thu 11/11/2021 21:19

Jennifer Morton <jenniferm@poz.com>

Re: POZ Comment - Other

To Pradeep Virdee

Retention Policy Junk Email (90 days) Expires 09/02/2022

Click here to download pictures. To help protect your privacy, Outlook prevented automatic download of some pictures in this message.

Hi Pradeep,

You have permission to use the image of the sample CBC report.

KInd regards,
Jennifer Morton

On 11/11/21 8:38 AM, Pradeep Virdee wrote:

POZ Contact Us Form

Please choose a topic from the dropdown:	Other
Your Name	Pradeep Virdee
Your Email Address	pradeep.virdee@csm.ox.ac.uk
Click to edit...	Hi, I'm a PhD student in the UK writing my thesis. There is an image on your webpage (link below) of an example CBC report. I'd like to reuse this image in my thesis to give an example of what a CBC report could look like. Is this image subject to copyright? If so, could I have your permission to use it to include in my thesis? The content will not be used in any other way. https://www.poz.com/basics/hiv-basics/understanding-lab-work-blood-tests I could not see the name or contact details of the author on the webpage, hence why I'm contacting you directly. Kind regards, Pradeep

You can edit this submission and view all your submissions easily.

--

Jennifer Morton
Managing Editor
Smart + Strong
POZ | HEP | Real Health | Tu Salud
Cancer Health | COVID Health
(212) 938-2058
jenniferm@smartandstrong.com

B2. Functional equations for FBC parameters

$$\boldsymbol{Mean\ corpuscular\ volume\ (fl)} = \frac{\boldsymbol{Haematocrit\ (L/L)}\times\mathbf{1000}}{\boldsymbol{Red\ blood\ cell\ count\ (10^{12}/L)}} \quad \text{Eq. B1}$$

$$Haematocrit\ (L/L) = \frac{Mean\ corpuscular\ volume\ (fl)\times Red\ blood\ cell\ count\ (10^{12}/L)}{1000} \quad \text{Eq. B2}$$

$$Haematocrit\ (L/L) = \frac{Haemoglobin\ (g/dL)}{Mean\ corpuscular\ haemoglobin\ concentration\ (g/dL)} \quad \text{Eq. B3}$$

$$Red\ blood\ cell\ count\ (10^{12}/L) = \frac{Haematocrit\ (L/L)\times 1000}{Mean\ corpuscular\ volume\ (fl)} \quad \text{Eq. B4}$$

$$Red\ blood\ cell\ count\ (10^{12}/L) = \frac{Haemoglobin\ (g/dL)\times 10}{Mean\ corpuscular\ haemoglobin\ (pg)} \quad \text{Eq. B5}$$

$$Haemoglobin\ (g/dL) = \frac{Red\ blood\ cell\ count\ (10^{12}/L)\times Mean\ corpuscular\ haemoglobin\ (pg)}{10} \quad \text{Eq. B6}$$

$$Haemoglobin\ (g/dL) = Haematocrit\ (L/L) \times$$
$$Mean\ corpuscular\ haemoglobin\ concentration\ (g/dL) \quad \text{Eq. B7}$$

$$Mean\ corpuscular\ haemoglobin\ (pg) = \frac{Haemoglobin\ (g/dL)\times 10}{Red\ blood\ cell\ count\ (10^{12}/L)} \quad \text{Eq. B8}$$

$$Mean\ corpuscular\ haemoglobin\ concentration\ (g/dL) = \frac{Haemoglobin\ (g/dL)}{Haematocrit\ (L/L)} \quad \text{Eq. B9}$$

$$WBC\ (10^9/L) = Basophils\ (10^9/L) + Eosinophils\ (10^9/L) + Lymphocytes\ (10^9/L) +$$
$$Monocytes\ (10^9/L) + Neutrophils\ (10^9/L) \quad \text{Eq. B10}$$

$$Basophils\ (10^9/L)\ = WBC\ (10^9/L) -\ Eosinophils\ (10^9/L)\ -\ Lymphocytes\ (10^9/L)\ -\ Monocytes\ (10^9/L)\ -\ Neutrophils\ (10^9/L)$$ Eq. B11

$$Basophils\ (10^9/L) = \frac{Basophil\ proportion\ (\%)}{100} \times WBC\ (10^9/L)$$ Eq. B12

$$Basophil\ proportion\ (\%) = \frac{Basophils\ (10^9/L)}{WBC\ (10^9/L)} \times 100$$ Eq. B13

$$Eosinophils\ (10^9/L)\ = WBC\ (10^9/L) -\ Basophils\ (10^9/L)\ -\ Lymphocytes\ (10^9/L)\ -\ Monocytes\ (10^9/L)\ -\ Neutrophils\ (10^9/L)$$ Eq. B14

$$Eosinophils\ (10^9/L) = \frac{Eosinophils\ proportion\ (\%)}{100} \times WBC\ (10^9/L)$$ Eq. B15

$$Eosinophils\ proportion\ (\%) = \frac{Eosinophils\ (10^9/L)}{WBC\ (10^9/L)} \times 100$$ Eq. B16

$$Lymphocytes\ (10^9/L) = WBC\ (10^9/L) -\ Basophils\ (10^9/L)\ -\ Eosinophils\ (10^9/L)\ -\ Monocytes\ (10^9/L)\ -\ Neutrophils\ (10^9/L)$$ Eq. B17

$$Lymphocytes\ (10^9/L) = \frac{Lymphocytes\ proportion\ (\%)}{100} \times WBC\ (10^9/L)$$ Eq. B18

$$Lymphocytes\ proportion\ (\%) = \frac{Lymphocytes\ (10^9/L)}{WBC\ (10^9/L)} \times 100$$ Eq. B19

$$Monocytes\ (10^9/L) = WBC\ (10^9/L) -\ Basophils\ (10^9/L)\ -\ Eosinophils\ (10^9/L)\ -\ Lymphocytes\ (10^9/L)\ -\ Neutrophils\ (10^9/L)$$ Eq. B20

$$Monocytes\ (10^9/L) = \frac{Monocytes\ proportion\ (\%)}{100} \times WBC\ (10^9/L)$$ Eq. B21

$$Monocytes\ proportion\ (\%) = \frac{Monocytes\ (10^9/L)}{WBC\ (10^9/L)} \times 100$$ Eq. B22

$$Neutrophils\ (10^9/L) = WBC\ (10^9/L) - Basophils\ (10^9/L) - Eosinophils\ (10^9/L) - Lymphocytes\ (10^9/L) - Monocytes\ (10^9/L)$$ Eq. B23

$$Neutrophils\ (10^9/L) = \frac{Neutrophils\ proportion\ (\%)}{100} \times WBC\ (10^9/L)$$ Eq. B24

$$Neutrophils\ proportion\ (\%) = \frac{Neutrophils\ (10^9/L)}{WBC\ (10^9/L)} \times 100$$ Eq. B25

B3. FBC entity codes with corresponding medical code =0

Table B1: Number of FBC parameters assigned medical code=0 (i.e. missing)

FBC parameter	Entity code	No. of records
RBC	194	212
WBC	207	220
Haemoglobin	173	22
Haematocrit	312	9
MCV	182	9
MCH	180	9
MCHC	181	38
RDW	361	2
Platelets	189	9
Basophil count	313	12
Eosinophil count	168	19
Lymphocyte count	208	16
Monocyte count	183	16
Neutrophil count	184	16
Total		**609**

Abbreviations: FBC=full blood count, RBC=red blood cell count, WBC=white blood cell count, MCV=mean corpuscular volume, MCH=mean corpuscular haemoglobin, MCHC=mean corpuscular haemoglobin concentration, RDW=red blood cell distribution width

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

B4. Details of rectified mismatches

Table B2: Number of rectified and non-rectified mismatches in strata 1[1]

Mismatch classified as	No. mismatches
MCH	48
MCV	7
Can't tell	34,059
No values	10,235
Total	**44,349**

Abbreviations: MCH=Mean corpuscular haemoglobin, MCV=Mean corpuscular volume
[1]Where existing medical or entity codes for both parameters could have been assigned
* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B3: Number of rectified and non-rectified mismatches in strata 2[1]

Mismatch classified as	No. mismatches
Basophil count	79,517
Basophil proportion	1,477,870
Eosinophil count	10,379
Eosinophil proportion	1,485,608
Lymphocyte count	20,389
Lymphocyte proportion	1,486,146
MPV	2,041,309
Monocyte count	24,500
Monocyte proportion	1,487,607
Neutrophil count	10,375
Neutrophil proportion	1,493,291

Platelets	8,430
Can't tell	602,633
No values	43,953
Total	**10,272,007**

Abbreviations: MPV=Mean platelet volume
[1]Where the medical code suggested one of the six FBC parameters without an existing entity code
* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B4: Number of rectified and non-rectified mismatches in strata 3[1]

Mismatch classified as	**No. mismatches**
Basophil count	997
Basophil proportion	20,820
Eosinophil count	989
Eosinophil proportion	19,988
Haemoglobin	7,354
Haematocrit	8
Lymphocyte count	2,621
Lymphocyte proportion	22,226
MCH	23
MCHC	13
MCV	17
MPV	51
Monocyte count	3,174
Monocyte proportion	22,635
Neutrophil count	8,211

Neutrophil proportion	28,251
Platelets	229
Red blood cell count	35
RDW	48
White blood cell count	300
Can't tell	344,438
No values	156,222
Total	**638,650**

Abbreviations: MCH=Mean corpuscular haemoglobin, MCHC=Mean corpuscular haemoglobin concentration, MCV=Mean corpuscular volume, MPV=Mean platelet volume, RDW=Red blood cell distribution width

[1]Where one code suggested a FBC parameter but the other suggested the measurement was not the FBC-related

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

B5. Number of FBC parameters for each unit of measure

Table B5: Number of red blood cell count records per unit of measure

Unit code-description	Frequency
0-No Data Entered	894,896
1-%	62,086
104-ng	29
113-Pa	2
114-pg	18
127-U/L	22
14-/(hpf)	3
148-year	1
149-MicroU/L	1
150-cms	14
153-10*9	139,253
154-L	15,928
155-10*6	4
157-10*3	30
158-mPas	1
159-cP	4
160-mU/L	1
161-1/1	9
162-/100(WBC)	72
164-/uL	33,090
17-/L	69,382
18-/m3	1
183-10*-2	3,502
184-10*-3	3
185-L/L	7
187-mmol/mmol	2
191-nmol/mmol	3
195-titre	1
198-/cu.mm	10
209-#/[tot]	11
21-/mL	7,570
210-#/HPF	14
219-cell/L	2
nmol/10*9(platelets)	11
26-1	5,739
268-Tcell/L	1
27-1(tot)	2
277-uL	2
29-10*-2(%)	1
31-10*12/L	17,842
32-10*3(rbc)	1
33-10*3/L	23,455
34-10*3/mL	1,919

35-10*6/L	33,559
36-10*6/mL	52,504
37-10*9/L	16,098,918
38-10*9/mL	22,734
4-(cfU)	1
40-cm(H2O)	4
41-d	1
46-fg	1
47-fL	464
56-g/dL	25
57-g/L	8,199
58-h	15
59-IU	321
6-(mclgu)	22
61-IU/L	4
62-IU/mL	1
7-OD	50
70-kU/L	9
73-lm/m2	34
74-m	2
84-mg/m3	2
89-mL	1
9-pH	14,462
92-mm	1,878
96-mmol/L	2,247
97-mmol/mol	7
Total	**17,510,410**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B6: Number of white blood cell count records per unit of measure entered

Unit code-description	**Frequency**
0-No Data Entered	894,896
1-%	62,086
104-ng	29
113-Pa	2
114-pg	18
127-U/L	22
14-/(hpf)	3
148-year	1
149-MicroU/L	1
150-cms	14
153-10*9	139,253
154-L	15,928
155-10*6	4

157-10*3	30
158-mPas	1
159-cP	4
160-mU/L	1
161-1/1	9
162-/100(WBC)	72
164-/uL	33,090
17-/L	69,382
18-/m3	1
183-10*-2	3,502
184-10*-3	3
185-L/L	7
187-mmol/mmol	2
191-nmol/mmol	3
195-titre	1
198-/cu.mm	10
209-#/[tot]	11
21-/mL	7,570
210-#/HPF	14
219-cell/L	2
258-nmol/10*9(platelets)	11
26-1	5,739
268-Tcell/L	1
27-1(tot)	2
277-uL	2
29-10*-2(%)	1
31-10*12/L	17,842
32-10*3(rbc)	1
33-10*3/L	23,455
34-10*3/mL	1,919
35-10*6/L	33,559
36-10*6/mL	52,504
37-10*9/L	16,098,921
38-10*9/mL	22,734
4-(cfU)	1
40-cm(H2O)	4
41-d	1
46-fg	1
47-fL	464
56-g/dL	25
57-g/L	8,199
58-h	15
59-IU	321
6-(mclg u)	22
61-IU/L	4
62-IU/mL	1
7-OD	50
70-kU/L	9
73-lm/m2	34

74-m	2
84-mg/m3	2
89-mL	1
9-pH	14,462
92-mm	1,878
96-mmol/L	2,247
97-mmol/mol	7
Total	**17,510,413**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B7: Number of haemoglobin records per unit of measure entered

Unit code-description	Frequency
0-No Data Entered	921,005
1-%	912
104-ng	19
105-ng/L	3
106-ng/mL	6
11-(ppm)	1
114-pg	9
122-rad	2
126-u	2
127-U/L	6
131-ug/dL	1
133-ug/L	1
14-/(hpf)	3
141-umol/h/g(Hb)	1
142-umol/L	5
145-umol/mmol	1
148-year	1
149-MicroU/L	2
150-cms	1
151-ratio	6
153-10*9	3
154-L	88
156-mmol	5
159-cP	4
161-1/1	1
164-/uL	1
17-/L	1
185-L/L	10
187-mmol/mmol	2
194-g/mol	144
199-GPL/ml	1
203-GPL U/ml	2

205-mmol/mol Hb	1
210-#/HPF	1
211-#G ( )/L	3
219-cell/L	3
220-cm	1
224-g/h	31
225-g/kg	9
226-Gcell/L	1
227-Gcopy/L	18
26-1	96
27-1(tot)	23
28-1/mL	1
29-10*-2(%)	1
3-(bsa)	2
31-10*12/L	3
37-10*9/L	180
38-10*9/mL	1
40-cm(H2O)	2
41-d	9
42-dB	2
45-ft	7
47-fL	6
49-g	133
51-g(hgb)	4
52-g(tot_nit)	3
53-g(tot_prot)	12
54-g(wet_tis)	5
55-g/d	67
56-g/dL	13,241,660
57-g/L	3,970,253
58-h	16
59-IU	3
6-(mclg u)	3
60-IU/d	1
61-IU/L	1
62-IU/mL	1
74-m	1
80-mg	3
81-mg/d	1
82-mg/dL	9
83-mg/L	8
86-mg/mmol	4
93-mm(Hg)	1
94-mm/h	1
96-mmol/L	49
97-mmol/mol	1,689
Total	**13,241,660**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B8: Number of haematocrit records per unit of measure entered

Unit code-description	Frequency
0-No Data Entered	7,547,339
1-%	1,364,620
102-mosmol/kg	1
104-ng	18
110-nmol/L	1
114-pg	7
120-pmol/L	1
122-rad	3
123-rem	1
127-U/L	27,091
133-ug/L	69
14-/(hpf)	1
142-umol/L	6
148-year	1
151-ratio	1,693,195
153-10*9	2
154-L	38,515
155-10*6	1
156-mmol	2
159-cP	2
161-1/1	1,409,173
17-/L	145,971
172-mPa.s	1
175-U/dL	1
184-10*-3	22
185-L/L	3,901,766
210-#/HPF	3
211-#G ( )/L	83
234-L/h	1
24-/week	2
25-/year	3
26-1	5,688
27-1(tot)	7,879
28-1/mL	24,542
3-(bsa)	1
31-10*12/L	146
32-10*3(rbc)	71
34-10*3/mL	32
36-10*6/mL	159
37-10*9/L	2,279
38-10*9/mL	74

4-(cfU)	3
40-cm(H2O)	30
41-d	1
42-dB	4
47-fL	2,306
49-g	2
5-(knk u)	1
56-g/dL	20
57-g/L	33,306
58-h	4
59-IU	1
61-IU/L	24,382
63-kat	1
67-/kg(body wt)	2
71-L/min	648
72-lm	1
73-lm/m2	1
74-m	2
83-mg/L	4
84-mg/m3	1
9-pH	1
96-mmol/L	8
Total	**16,229,502**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B9: Number of mean corpuscular volume records per unit of measure entered

Unit code-description	**Frequency**
0-No Data Entered	755,880
1-%	3,306
104-ng	22
106-ng/mL	1
113-Pa	1
114-pg	12,224
127-U/L	1
14-/(hpf)	3
142-umol/L	1
148-year	1
149-MicroU/L	3
150-cms	3
151-ratio	1
153-10*9	1
154-L	12
157-10*3	1
161-1/1	2

178-U/mL	5
185-L/L	6
194-g/mol	1
219-cell/L	1
26-1	51
27-1(tot)	6
31-10*12/L	3
37-10*9/L	1,212
39-degC	1
4-(cfU)	1
40-cm(H2O)	1
41-d	15
44-deg	1
46-fg	8
47-fL	16,190,073
48-fmol	153
49-g	3
50-g(creat)	3
56-g/dL	9,584
57-g/L	6
58-h	17
62-IU/mL	34
71-L/min	1
74-m	8
83-mg/L	1
86-mg/mmol	2
88-mU/mL	1
Total	**16,972,661**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B10: Number of mean corpuscular haemoglobin records per unit of measure entered

Unit code-description	**Frequency**
0-No Data Entered	609,337
1-%	56
104-ng	11
112-ns	1
113-Pa	4
114-pg	14,776,728
115-pg/L	6
116-pg/mL	268,581
117-pkat	2
118-pm	1
149-MicroU/L	1
151-ratio	2

154-L	11
161-1/1	1
179-ug/mL	2
185-L/L	3
223-g/24h	1
224-g/h	1
26-1	16
31-10*12/L	1
32-10*3(rbc)	6
34-10*3/mL	1
37-10*9/L	5
47-fL	68
48-fmol	5
49-g	4
5-(knk u)	1
51-g(hgb)	4
55-g/d	2
56-g/dL	169,003
57-g/L	4,590
58-h	21
7-OD	1
8-(pa)	1
80-mg	2
88-mU/mL	14
92-mm	4
96-mmol/L	112
Total	**15,828,610**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B11: Number of mean corpuscular haemoglobin concentration records per unit of measure entered

Unit code-description	Frequency
0-No Data Entered	367,274
1-%	453,269
100-mol/s	1
104-ng	13
114-pg	29,499
115-pg/L	8,414
131-ug/dL	4
139-umol/d	1
14-/(hpf)	1
151-ratio	3,800
154-L	9

161-1/1	44,855
17-/L	1,896
171-mmol/mmol(creat)	1
175-U/dL	1
179-ug/mL	4
185-L/L	46,388
253-mmol/g(dry wt)	1
26-1	19
27-1(tot)	3
37-10*9/L	10
40-cm(H2O)	1
47-fL	17
49-g	16
50-g(creat)	1
51-g(hgb)	1
54-g(wet_tis)	2
55-g/d	18
56-g/dL	7,872,690
57-g/L	2,569,397
58-h	11
82-mg/dL	1
83-mg/L	1
96-mmol/L	202,810
97-mmol/mol	2
98-mol/kg	1
99-mol/L	2
Total	**11,600,434**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B12: Number of red blood cell distribution width records per unit of measure entered

Unit code-description	**Frequency**
0-No Data Entered	6,865,155
1-%	319,775
115-pg/L	7
126-u	1
154-L	1
161-1/1	41
164-/uL	1
214-#T ( )/L	1
215-%Hb	1
38-10*9/mL	4
47-fL	2,781
90-mL/min	1

Total	**7,187,769**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B13: Number of platelet records per unit of measure entered

Unit code-description	Frequency
0-No Data Entered	770,031
1-%	73,633
104-ng	29
114-pg	8
12-(pu)	4
124-s	2
126-u	1
127-U/L	4
150-cms	6
151-ratio	15
153-10*9	76,541
154-L	15,827
157-10*3	31
159-cP	4
161-1/1	3,800
162-/100(WBC)	72
164-/uL	12,727
17-/L	92,836
183-10*-2	3,481
184-10*-3	4
185-L/L	3
187-mmol/mmol	2
191-nmol/mmol	6
195-titre	1
198-/cu.mm	1
209-#/[tot]	11
21-/mL	7,378
210-#/HPF	11
219-cell/L	1
258-nmol/10*9(platelets)	11
26-1	8,398
3-(bsa)	1
31-10*12/L	25,436
33-10*3/L	23,528
34-10*3/mL	1,130
35-10*6/L	4
36-10*6/mL	52,377
37-10*9/L	16,110,377
38-10*9/mL	15,195

4-(cfU)	1
40-cm(H2O)	4
41-d	5
46-fg	1
47-fL	3,993
56-g/dL	15
57-g/L	8,234
58-h	11
59-IU	320
61-IU/L	3
7-OD	50
70-kU/L	7
73-lm/m2	34
9-pH	13,611
94-mm/h	1
96-mmol/L	2,246
97-mmol/mol	7
Total	**17,321,470**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B14: Number of mean platelet volume records per unit of measure entered

Unit code-description	Frequency
37-10*9/L	8,635
38-10*9/mL	1
47-fL	2,041,360
Total	**2,049,996**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B15: Number of basophil count records per unit of measure entered

Unit code-description	Frequency
0-No Data Entered	603,616
1-%	239,970
104-ng	3
114-pg	1
126-u	13
127-U/L	1
133-ug/L	5
147-week	1
151-ratio	2

153-10*9	74,152
154-L	27,432
155-10*6	1
157-10*3	17
158-mPas	1
159-cP	3
161-1/1	8
162-/100(WBC)	52
164-/uL	1
17-/L	107,127
183-10*-2	3,463
184-10*-3	1
185-L/L	5
187-mmol/mmol	2
191-nmol/mmol	2
2-(bdsk u)	1
209-#/[tot]	11
21-/mL	314
210-#/HPF	11
214-#T ( )/L	1
26-1	5,745
27-1(tot)	2
30-10*-3(%)	1
31-10*12/L	17,197
33-10*3/L	8,771
34-10*3/mL	354
35-10*6/L	5
36-10*6/mL	16
37-10*9/L	13,939,287
38-10*9/mL	14,127
40-cm(H2O)	1
47-fL	500
56-g/dL	5
57-g/L	5,748
59-IU	319
61-IU/L	3
7-OD	42
70-kU/L	2
73-lm/m2	30
9-pH	13,567
96-mmol/L	2,173
97-mmol/mol	7
Total	**15,064,119**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B16: Number of basophil proportion records per unit of measure entered

Unit code-description	Frequency
1-%	1,498,690
153-10*9	1
17-/L	1,361
37-10*9/L	79,115
Total	**1,579,167**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B17: Number of eosinophil count records per unit of measure entered

Unit code-description	Frequency
0-No Data Entered	612,199
1-%	296,924
104-ng	4
122-rad	1
127-U/L	1
133-ug/L	1
151-ratio	1
153-10*9	75,539
154-L	26,763
157-10*3	14
158-mPas	1
161-1/1	10
162-/100(WBC)	52
17-/L	117,487
183-10*-2	3,471
184-10*-3	4
187-mmol/mmol	2
191-nmol/mmol	4
195-titre	1
209-#/[tot]	11
21-/mL	313
210-#/HPF	10
26-1	5,674
268-Tcell/L	1
27-1(tot)	1
3-(bsa)	2
31-10*12/L	17,222
33-10*3/L	8,800
34-10*3/mL	374
35-10*6/L	10
36-10*6/mL	148

37-10*9/L	14,761,804
38-10*9/mL	14,733
4-(cfU)	3
40-cm(H2O)	4
47-fL	462
56-g/dL	4
57-g/L	5,772
58-h	2
59-IU	319
61-IU/L	4
7-OD	50
70-kU/L	2
73-lm/m2	33
83-mg/L	2
84-mg/m3	1
89-mL	107
9-pH	13,398
96-mmol/L	2,183
97-mmol/mol	7
Total	**15,963,935**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B18: Number of eosinophil proportion records per unit of measure entered

Unit code-description	Frequency
1-%	1,505,596
153-10*9	15
17-/L	1,346
191-nmol/mmol	1
37-10*9/L	9,988
Total	**1,516,946**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B19: Number of lymphocyte count records per unit of measure entered

Unit code-description	Frequency
0-No Data Entered	667,073
1-%	247,851
104-ng	2
122-rad	1

124-s	1
127-U/L	1
142-umol/L	1
151-ratio	1
153-10*9	78,772
154-L	26,919
157-10*3	15
158-mPas	1
159-cP	1
161-1/1	7
162-/100(WBC)	52
164-/uL	7
17-/L	119,164
183-10*-2	3,477
184-10*-3	2
187-mmol/mmol	2
191-nmol/mmol	3
195-titre	1
209-#/[tot]	11
21-/mL	313
210-#/HPF	10
23-/second	1
26-1	5,722
31-10*12/L	17,231
33-10*3/L	8,786
34-10*3/mL	380
35-10*6/L	7
36-10*6/mL	148
37-10*9/L	14,950,124
38-10*9/mL	14,606
46-fg	1
47-fL	467
56-g/dL	1,710
57-g/L	5,894
58-h	1
59-IU	319
61-IU/L	3
7-OD	50
70-kU/L	4
73-lm/m2	30
84-mg/m3	1
9-pH	13,297
96-mmol/L	2,285
97-mmol/mol	7
Total	**16,164,762**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B20: Number of lymphocyte proportion records per unit of measure entered

Unit code-description	Frequency
1-%	1,508,370
153-10*9	4
17-/L	1,364
37-10*9/L	21,607
Total	**1,531,345**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B21: Number of monocyte count records per unit of measure entered

Unit code-description	Frequency
0-No Data Entered	645,114
1-%	307,782
104-ng	4
114-pg	1
124-s	1
127-U/L	1
14-/(hpf)	1
142-umol/L	1
151-ratio	4
153-10*9	187,247
154-L	26,891
157-10*3	18
161-1/1	13
162-/100(WBC)	72
17-/L	118,827
183-10*-2	3,477
185-L/L	165
187-mmol/mmol	2
191-nmol/mmol	3
2-(bdsk u)	2
209-#/[tot]	11
21-/mL	313
210-#/HPF	11
26-1	5,727
27-1(tot)	2
30-10*-3(%)	1
31-10*12/L	17,224
33-10*3/L	8,801

34-10*3/mL	366
35-10*6/L	8
36-10*6/mL	150
37-10*9/L	14,687,897
38-10*9/mL	15,258
4-(cfU)	1
41-d	1
47-fL	462
56-g/dL	5
57-g/L	5,814
58-h	4
59-IU	319
61-IU/L	3
69-kpal	43
7-OD	50
70-kU/L	4
73-lm/m2	32
84-mg/m3	1
9-pH	13,634
96-mmol/L	2,183
97-mmol/mol	7
Total	**16,047,958**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B22: Number of monocyte proportion records per unit of measure entered

Unit code-description	Frequency
1-%	1,510,233
153-10*9	11
154-L	46
17-/L	1,366
37-10*9/L	26,217
61-IU/L	1
Total	**1,537,874**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B23: Number of neutrophil count records per unit of measure entered

Unit code-description	Frequency

0-No Data Entered	645,923
1-%	239,351
104-ng	10
114-pg	1
122-rad	1
124-s	2
127-U/L	1
139-umol/d	1
150-cms	3
151-ratio	1
153-10*9	74,176
154-L	26,919
155-10*6	1
157-10*3	12
159-cP	1
161-1/1	7
162-/100(WBC)	52
17-/L	114,573
178-U/mL	1
18-/m3	1
183-10*-2	3,477
184-10*-3	5
185-L/L	6
187-mmol/mmol	2
191-nmol/mmol	5
2-(bdsk u)	1
209-#/[tot]	11
21-/mL	317
210-#/HPF	13
26-1	5,723
30-10*-3(%)	1
31-10*12/L	17,231
33-10*3/L	8,790
34-10*3/mL	437
35-10*6/L	4
36-10*6/mL	240
37-10*9/L	15,061,611
38-10*9/mL	14,599
40-cm(H2O)	1
46-fg	1
47-fL	463
56-g/dL	7
57-g/L	5,915
58-h	11
59-IU	319
61-IU/L	3
7-OD	50
70-kU/L	4
73-lm/m2	30

74-m	1
9-pH	13,353
96-mmol/L	2,185
97-mmol/mol	7
Total	**16,235,860**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B24: Number of neutrophil proportion records per unit of measure entered

Unit code-description	Frequency
1-%	1,521,541
153-10*9	4
154-L	1
157-10*3	1
17-/L	2,307
37-10*9/L	16,233
Total	**1,540,087**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

B6. Summary of quantiles to identify extent of extreme values per FBC parameter

Table B25: Summary statistics of red blood cell count (10 percentiles) entered as 10^{12}/L

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	1,478,446	3.4	-2.6	3.3	3.5	3.7	3.8
2	1,437,093	3.9	3.8	3.9	3.9	4.0	4.0
3	1,489,618	4.1	4.0	4.1	4.1	4.2	4.2
4	1,491,707	4.3	4.2	4.2	4.3	4.3	4.3
5	1,390,775	4.4	4.3	4.4	4.4	4.4	4.5
6	1,402,740	4.5	4.5	4.5	4.5	4.6	4.6
7	1,538,875	4.7	4.6	4.6	4.7	4.7	4.7
8	1,363,102	4.8	4.7	4.8	4.8	4.8	4.9
9	1,447,950	5.0	4.9	4.9	5.0	5.0	5.1
10	1,436,144	5.6	5.1	5.2	5.3	5.4	8,577.0
No values	7,506						
Total	**14,483,956**	**4.5**	**-2.6**	**4.1**	**4.5**	**4.8**	**8,577.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B26: Summary statistics of white blood cell count (10 percentiles) entered as 10^{9}/L

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	1,717,659	4	-3.9	3.8	4.2	4.5	4.7
2	1,501,619	5.1	4.7	4.9	5.1	5.2	5.4
3	1,830,881	5.7	5.4	5.5	5.7	5.8	5.9
4	1,668,254	6.2	5.9	6.1	6.2	6.3	6.4

5	1,326,666	6.6	6.4	6.5	6.6	6.7	6.9
6	1,768,862	7.1	6.9	7	7.1	7.3	7.4
7	1,496,511	7.7	7.4	7.6	7.7	7.9	8.0
8	1,689,474	8.5	8	8.2	8.4	8.7	8.9
9	1,537,474	9.6	8.9	9.2	9.5	9.9	10.3
10	1,548,636	15.1	10.3	10.9	11.8	13.4	123,000.0
No values	12,882						
Total	**16,098,918**	**7.5**	**-3.9**	**5.6**	**6.9**	**8.4**	**123,000.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B27: Summary statistics of haemoglobin (10 percentiles) entered as g/dL

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	1,394,818	10.1	-10.0	9.6	10.4	10.9	11.2
2	1,418,585	11.7	11.2	11.5	11.8	12.0	12.1
3	1,212,471	12.4	12.1	12.3	12.4	12.5	12.6
4	1,525,900	12.9	12.6	12.8	12.9	13.0	13.1
5	1,331,924	13.4	13.1	13.3	13.4	13.4	13.5
6	1,326,151	13.7	13.5	13.6	13.7	13.8	13.9
7	1,194,615	14.1	13.9	14.0	14.1	14.2	14.3
8	1,258,449	14.6	14.3	14.5	14.6	14.7	14.8
9	1,255,958	15.2	14.8	15.0	15.2	15.3	15.5
10	1,282,578	33.9	15.5	15.8	16.2	17.0	4,095.0
No values	40,211						
Total	**13,241,660**	15.1	-10.0	12.4	13.5	14.5	4,095.0

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B28: Summary statistics of haematocrit (10 percentiles) entered as L/L

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	397,591	0.3	0.0	0.3	0.3	0.3	0.3
2	435,587	0.4	0.3	0.4	0.4	0.4	0.4
3	346,104	0.4	0.4	0.4	0.4	0.4	0.4
4	397,716	0.4	0.4	0.4	0.4	0.4	0.4
5	381,178	0.4	0.4	0.4	0.4	0.4	0.4
6	500,785	0.4	0.4	0.4	0.4	0.4	0.4
7	324,622	0.4	0.4	0.4	0.4	0.4	0.4
8	355,883	0.4	0.4	0.4	0.4	0.4	0.4
9	374,659	0.5	0.4	0.4	0.5	0.5	0.5
10	387,557	7.2	0.5	0.5	0.5	0.5	416.0
No values	87						
Total	**3,901,766**	**1.1**	**0.0**	**0.4**	**0.4**	**0.4**	**416.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B29: Summary statistics of mean corpuscular volume (10 percentiles) entered as fL

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	1,734,089	78.9	-74.0	77.9	81.2	83.0	84.0
2	1,565,428	85.6	84.0	85.0	85.7	86.1	86.7
3	1,571,833	87.5	86.8	87.0	87.6	88.0	88.2
4	1,617,004	89.0	88.3	88.7	89.0	89.3	89.7
5	1,895,207	90.4	89.7	90.0	90.4	90.9	91.0
6	1,362,215	91.7	91.0	91.4	91.8	92.0	92.1
7	1,591,060	92.9	92.2	92.6	93.0	93.2	93.6

8	1,669,684	94.5	93.6	94.0	94.4	95.0	95.3
9	1,654,286	96.6	95.4	96.0	96.6	97.2	98.0
10	1,521,461	102.2	98.0	99.1	100.7	103.0	10,730.0
No values	7,806						
Total	**16,190,073**	**90.8**	**-74.0**	**87.5**	**91.0**	**94.4**	**10,730.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B30: Summary statistics of mean corpuscular haemoglobin (10 percentiles) entered as fL

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	1,482,873	25.3	-28.9	24.6	26.1	26.9	27.4
2	1,502,316	28.1	27.4	27.9	28.2	28.4	28.6
3	1,547,209	29.0	28.7	28.9	29.0	29.2	29.3
4	1,671,415	29.7	29.3	29.5	29.7	29.8	29.9
5	1,301,121	30.1	30.0	30.0	30.1	30.2	30.3
6	1,547,553	30.6	30.3	30.5	30.6	30.7	30.8
7	1,473,574	31.1	30.9	31.0	31.1	31.2	31.3
8	1,396,037	31.6	31.4	31.5	31.6	31.8	31.9
9	1,432,598	32.3	32.0	32.1	32.3	32.6	32.8
10	1,416,694	34.3	32.9	33.2	33.7	34.5	3,508.0
No values	5,338						
Total	**14,776,728**	**30.2**	**-28.9**	**29.0**	**30.3**	**31.6**	**3,508.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B31: Summary statistics of mean corpuscular haemoglobin concentration (10 percentiles) entered as g/dL

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	871,682	30.8	0.0	30.5	31.0	31.4	31.6
2	803,578	32.0	31.7	31.8	32.0	32.1	32.2
3	752,678	32.5	32.3	32.4	32.5	32.6	32.6
4	981,497	32.9	32.7	32.8	32.9	33.0	33.0
5	815,197	33.2	33.1	33.1	33.2	33.3	33.3
6	503,056	33.5	33.4	33.4	33.5	33.5	33.5
7	794,227	33.7	33.5	33.6	33.7	33.8	33.8
8	925,860	34.0	33.8	33.9	34.0	34.1	34.2
9	654,625	34.4	34.3	34.3	34.4	34.5	34.6
10	769,626	35.4	34.7	34.8	35.0	35.4	479.3
No values	664						
Total	**7,872,690**	**33.2**	**0.0**	**32.4**	**33.3**	**34.0**	**479.3**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B32: Summary statistics of red blood cell distribution width (10 percentiles) entered as proportions

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	36,809	12.1	0.0	11.9	12.2	12.4	12.5
2	34,913	12.8	12.6	12.7	12.8	12.9	12.9
3	24,539	13.0	13.0	13.0	13.0	13.1	13.1
4	38,604	13.3	13.2	13.2	13.3	13.4	13.4

5	33,680	13.6	13.5	13.5	13.6	13.7	13.7
6	30,632	13.9	13.8	13.8	13.9	14.0	14.0
7	31,034	14.2	14.1	14.1	14.2	14.3	14.4
8	27,847	14.7	14.5	14.6	14.7	14.8	14.9
9	29,791	15.4	15.0	15.1	15.3	15.6	15.9
10	31,924	17.9	16.0	16.4	17.2	18.6	43.8
No values	2						
Total	**319,775**	**14.0**	**0.0**	**13.0**	**13.7**	**14.6**	**43.8**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B33: Summary statistics of platelet (10 percentiles) entered as 10^9/L

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	1,628,527	143.4	109.0	131	154	167	176
2	1,595,159	190.5	177.0	184.0	191.0	197.0	202.0
3	1,671,535	212.8	203.0	208.0	213.0	218.0	222.0
4	1,594,669	231.1	223.0	227.0	231.0	235.0	239.0
5	1,622,895	248.0	240.0	244.0	248.0	252.0	256.0
6	1,613,494	265.3	257.0	261.0	265.0	270.0	274.0
7	1,553,958	284.2	275.0	279.0	284.0	289.0	294.0
8	1,618,971	307.2	295.0	300.0	307.0	314.0	321.0
9	1,597,767	341.0	322.0	330.0	340.0	351.0	365.0
10	1,599,479	442.0	366.0	384.0	411.0	462.0	319,000.0
No values	13,932						
Total	**16,110,377**	**266.1**	**109.0**	**213.0**	**256.0**	**307.0**	**319,000.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B34: Summary statistics of mean platelet volume (10 percentiles) entered as fL

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	215,699	7.2	0.0	7	7.3	7.6	7.7
2	235,143	8.0	7.8	7.9	8.0	8.1	8.2
3	189,376	8.5	8.3	8.3	8.4	8.6	8.6
4	216,267	8.9	8.7	8.8	8.9	9.0	9.0
5	196,009	9.2	9.1	9.1	9.2	9.3	9.4
6	187,726	9.6	9.5	9.5	9.6	9.7	9.8
7	195,817	10.0	9.9	9.9	10.0	10.1	10.2
8	204,374	10.5	10.3	10.4	10.5	10.6	10.7
9	214,249	11.1	10.8	10.9	11.1	11.2	11.4
10	186,700	14.0	11.5	11.7	12.0	12.4	1103.0
No values	0						
Total	**2,041,360**	**9.6**	**0.0**	**8.4**	**9.4**	**10.5**	**1,103.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B35: Summary statistics of basophil count (4 percentiles) entered as 10^9/L

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	5,503,616	0.0	0.0	0.0	0.0	0.0	0.0
2	1,529,203	0.0	0.0	0.0	0.0	0.0	0.0
3	3,292,989	0.0	0.0	0.0	0.0	0.1	0.1
4	3,343,866	0.1	0.1	0.1	0.1	0.1	1,218.0
No values	269,613						
Total	**13,939,287**	**0.0**	**0.0**	**0.0**	**0.0**	**0.1**	**1,218.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B36: Summary statistics of basophil proportion (4 percentiles) entered as proportions

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	398,537	0.1	0.0	0.0	0.0	0.2	0.3
2	383,789	0.6	0.4	0.4	0.5	0.7	0.8
3	577,797	1.0	0.9	1.0	1.0	1.0	1.0
4	138,567	1.8	1.1	1.3	2.0	2.0	183.0
No values	0						
Total	**1,498,690**	**0.7**	**0.0**	**0.3**	**0.8**	**1.0**	**183.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B37: Summary statistics of eosinophil count (4 percentiles) entered as 10^9/L

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	5,412,643	0.1	0.0	0.1	0.1	0.1	0.1
2	4,781,444	0.2	0.1	0.2	0.2	0.2	0.2
3	868,543	0.2	0.2	0.2	0.2	0.3	0.3
4	3,661,853	0.4	0.3	0.3	0.3	0.4	1,218.0
No values	37,321						
Total	**14,761,804**	**0.2**	**0.0**	**0.1**	**0.2**	**0.3**	**1,218.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B38: Summary statistics of eosinophil proportion (4 percentiles) entered as proportions

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	377,271	1.0	0.0	1.0	1.0	1.3	1.7
2	383,208	2.1	1.8	2.0	2.0	2.0	2.5
3	465,679	3.3	2.6	3.0	3.0	3.9	4.0
4	279,438	6.3	4.1	5.0	5.6	7.0	100.0
No values	0						
Total	**1,505,596**	**3.0**	**0.0**	**1.7**	**2.5**	**4.0**	**100.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B39: Summary statistics of lymphocyte count (4 percentiles) entered as 10^9/L

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	3,745,534	1.1	-35.0	1.0	1.2	1.3	1.5
2	4,336,541	1.7	1.5	1.6	1.7	1.8	1.9
3	3,129,118	2.1	1.9	2.0	2.1	2.2	2.4
4	3,733,174	3.4	2.4	2.5	2.8	3.2	275,000.0
No values	5,757						
Total	**14,950,124**	**2.1**	**-35.0**	**1.5**	**1.9**	**2.4**	**275,000.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B40: Summary statistics of lymphocyte proportion (4 percentiles) entered as proportions

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	382,015	17.0	0.0	14.8	18.0	20.2	22.0
2	376,908	25.4	22.1	24.0	25.4	27.0	28.0
3	374,921	31.2	28.1	29.9	31.0	32.8	34.0
4	374,526	40.3	34.1	36.0	38.8	42.4	285.1
No values	0						
Total	**1,508,370**	**28.4**	**0.0**	**22.0**	**28.0**	**34.0**	**285.1**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B41: Summary statistics of monocyte count (4 percentiles) entered as 10^9/L

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	4,820,118	0.3	0.0	0.3	0.4	0.4	0.4
2	3,135,589	0.5	0.4	0.5	0.5	0.5	0.5
3	3,088,156	0.6	0.5	0.6	0.6	0.6	0.7
4	3,639,532	0.9	0.7	0.7	0.8	0.9	1,218.0
No values	4,502						
Total	**14,687,897**	**0.6**	**0.0**	**0.4**	**0.5**	**0.7**	**1,218.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B42: Summary statistics of monocyte proportion (4 percentiles) entered as proportions

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max

1	480,522	5.0	0.0	4.4	5.0	6.0	6.0
2	275,444	6.8	6.1	6.7	7.0	7.0	7.1
3	426,274	8.2	7.2	8.0	8.0	8.9	9.0
4	327,993	11.3	9.1	10.0	10.7	12.0	100.0
No values	0						
Total	**1,510,233**	**7.6**	**0.0**	**6.0**	**7.1**	**9.0**	**100.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B43: Summary statistics of neutrophil count (4 percentiles) entered as 10^9/L

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	3,782,760	2.5	-69.0	2.2	2.6	2.9	3.1
2	3,751,045	3.6	3.1	3.4	3.6	3.8	4.1
3	3,879,582	4.6	4.1	4.3	4.6	4.9	5.3
4	3,640,029	7.4	5.3	5.8	6.5	7.8	12,200.0
No values	8,195						
Total	**15,061,611**	**4.5**	**-69.0**	**3.1**	**4.1**	**5.3**	**12,200.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Table B44: Summary statistics of neutrophil proportion (4 percentiles) entered as proportions

Percentile	Frequency	Mean	Min	Lower quartile	Median	Upper quartile	Max
1	381,823	47.2	0.0	45.0	49.0	51.8	53.9
2	382,254	57.1	54.0	55.6	57.0	59.0	60.0
3	397,590	63.6	60.1	62.0	63.6	65.1	67.0

4	359,874	73.4	67.1	69.3	72.0	76.0	100.0
No values	0						
Total	**1,521,541**	**60.2**	**0.0**	**53.9**	**60.0**	**67.0**	**100.0**

* This table is taken and adapted from: Virdee *et al* (2020). Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data 7, 96*. DOI: 10.1186/s40537-020-00375-w[128]. Published under a CC-BY licence – permissions not required.

Appendix C

Appendices for Chapter 4 (trajectories in the FBC)

C1. Biologically plausible FBC ranges

Table C1: Biologically plausible ranges for each FBC parameter

FBC parameter	Biologically plausible range
Red blood cell count (10^{12}/L)	2.5 – 7.5
Haemoglobin (g/dL)	0.3 – 21.0
Haematocrit/packed cell volume (L/L)	0.1 – 0.6
Mean corpuscular volume (fL)	53 – 125
Mean corpuscular haemoglobin (pg)	22 – 36
Mean corpuscular haemoglobin concentration (g/dL)	27 – 37
Red cell distribution width (%)	9 – 15
Platelet count (10^9/L)	0 – 1500
Mean platelet volume (fL)	5 – 14
White blood cell count (10^9/L)	0 – 850
Basophil count (10^9/L)	0 – 1.5
Eosinophil count (10^9/L)	0 – 8
Lymphocyte count (10^9/L)	0 – 850
Monocyte count (10^9/L)	0 – 40
Neutrophil count (10^9/L)	0 – 850

Abbreviations: FBC=full blood count.

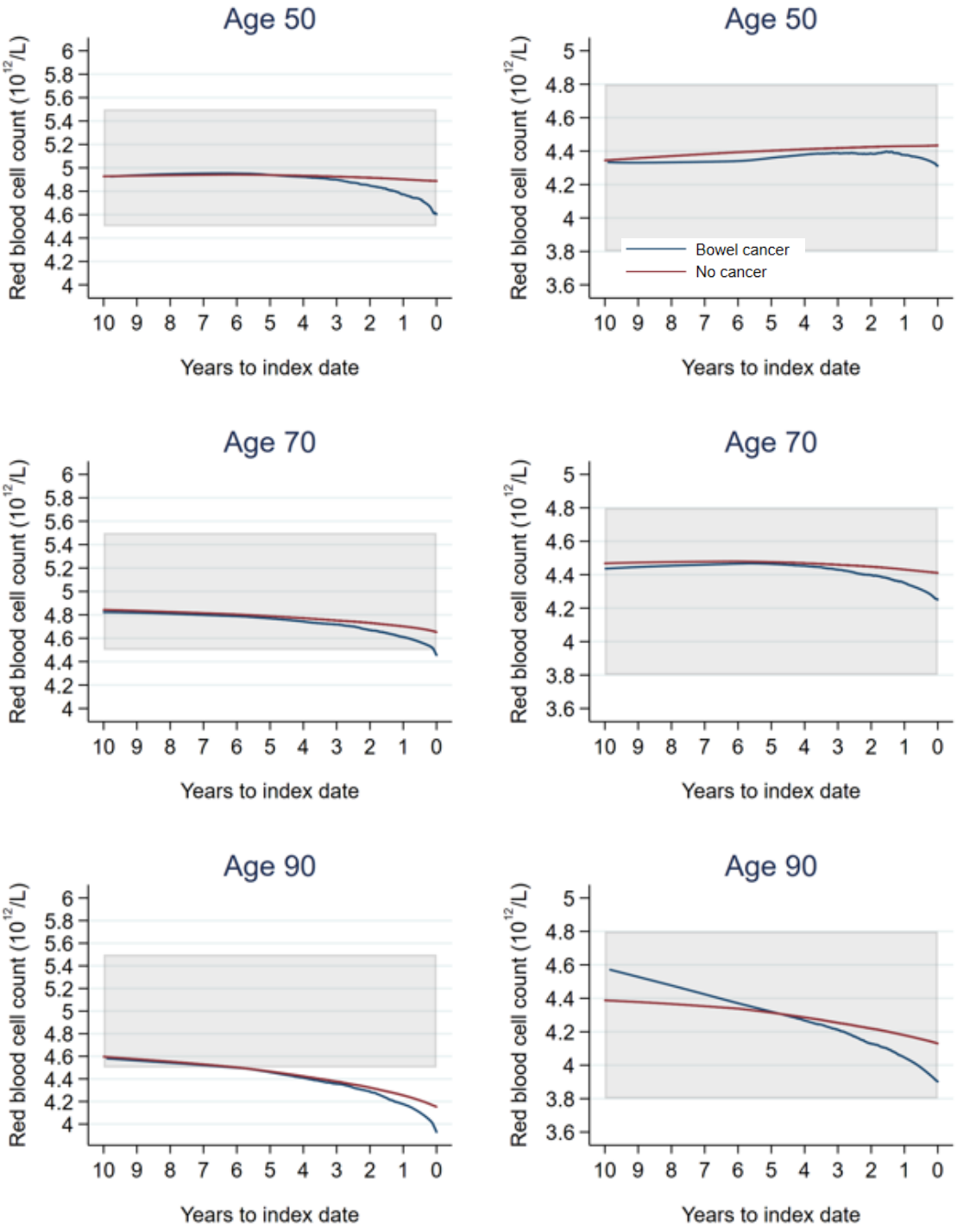
C2. Trajectories & FBC reference ranges

Table C2: Reference range for patients aged 13+ years for each FBC parameter[1]

FBC parameter	Males normal range	Females normal range
Red blood cell count (10^9/L)	4.5-5.5	3.8-4.8
Haemoglobin (g/dL)	13-17	12-15
Haematocrit (L/L)	0.40-0.50	0.36-0.46
Mean corpuscular volume (fL)	83-101	83-101
Mean corpuscular haemoglobin (pg)	27-32	27-32
Mean corpuscular haemoglobin concentration (g/dL)	31.5-34.5	31.5-34.5
Platelets (10^{12}/L)	150-400	150-400
Mean platelet volume (fL)	9.0-12.1	9.0-12.1
White blood cell count (10^{12}/L)	4-11	4-11
Basophil count (10^{12}/L)	0.02-0.10	0.02-0.10
Eosinophil count (10^{12}/L)	0.02-0.50	0.02-0.50
Lymphocyte count (10^{12}/L)	1-4	1-4
Monocyte count (10^{12}/L)	0.2-1.0	0.2-1.0
Neutrophil count (10^{12}/L)	2-7	2-7

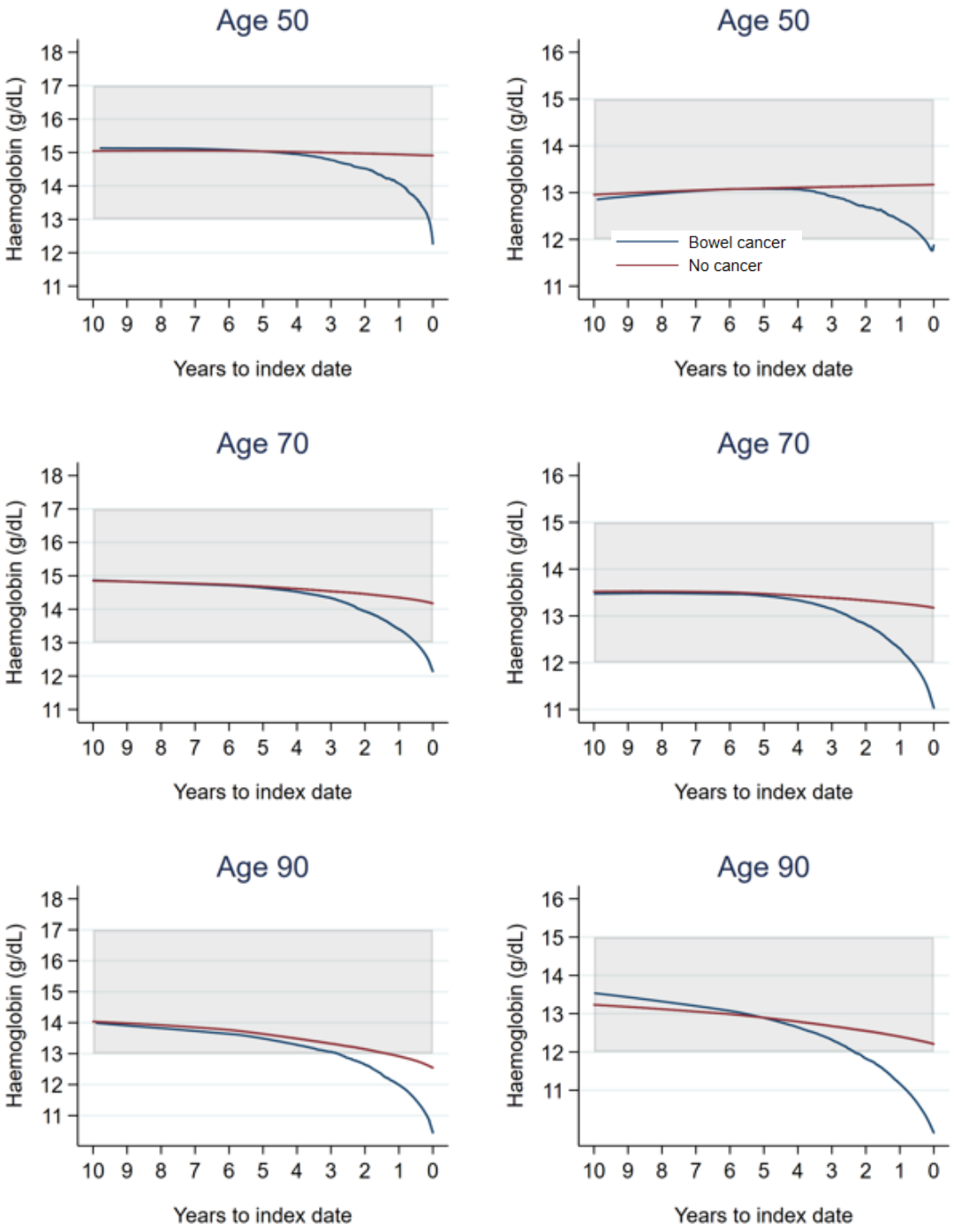
[1]References ranges are taken from: Platt, A. The Department of Laboratory Haematology – OUH Haematology Laboratory Handbook. Available at <https://www.ouh.nhs.uk/services/departments/laboratory-medicine/haematology-laboratories/documents/haematology-handbook.pdf>. Last accessed 20 October 2021 (2020)

Figure C1: Trajectories (LOWESS) in red blood cell count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



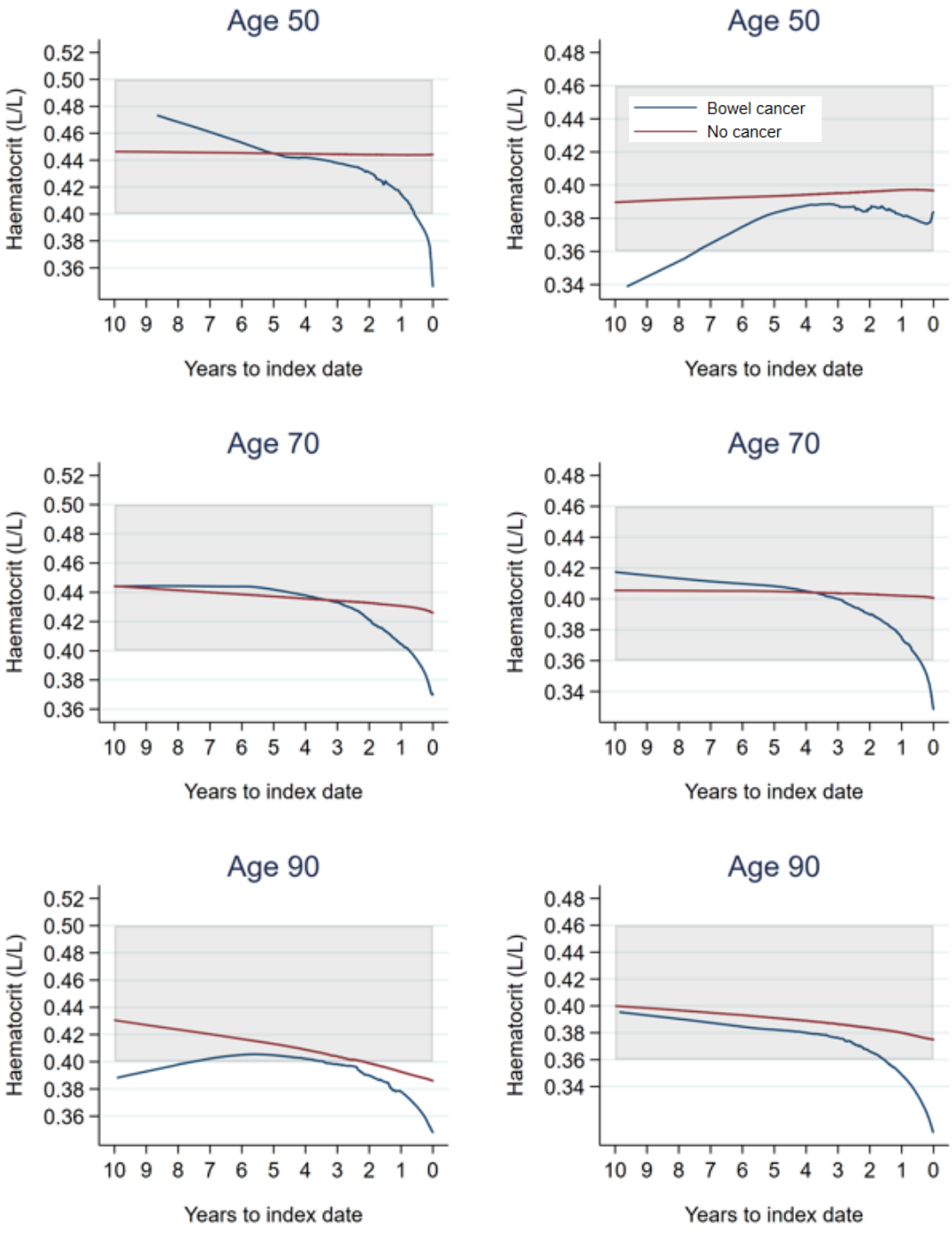
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C2: Trajectories (LOWESS) in haemoglobin for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



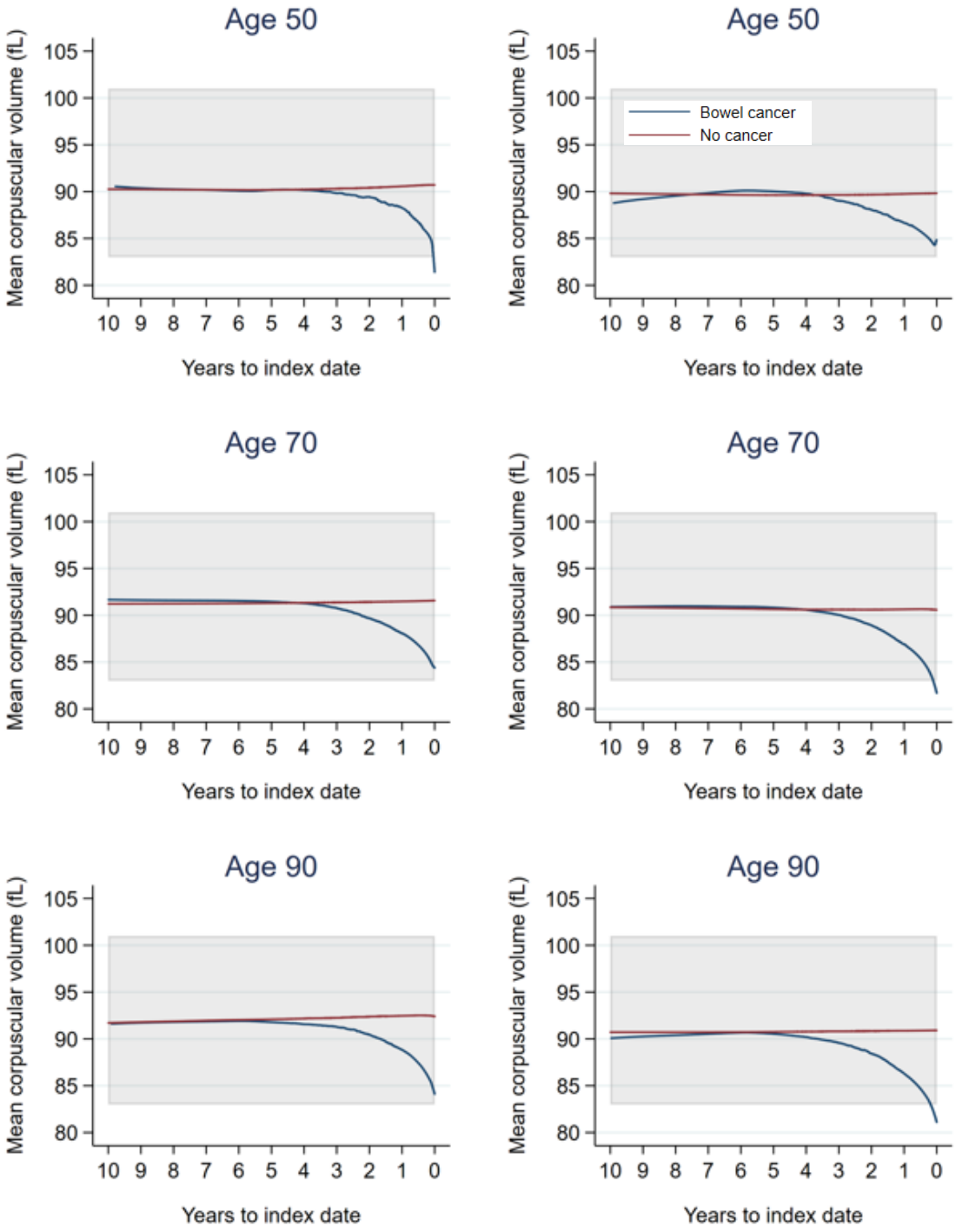
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C3: Trajectories (LOWESS) in haematocrit for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2,3]



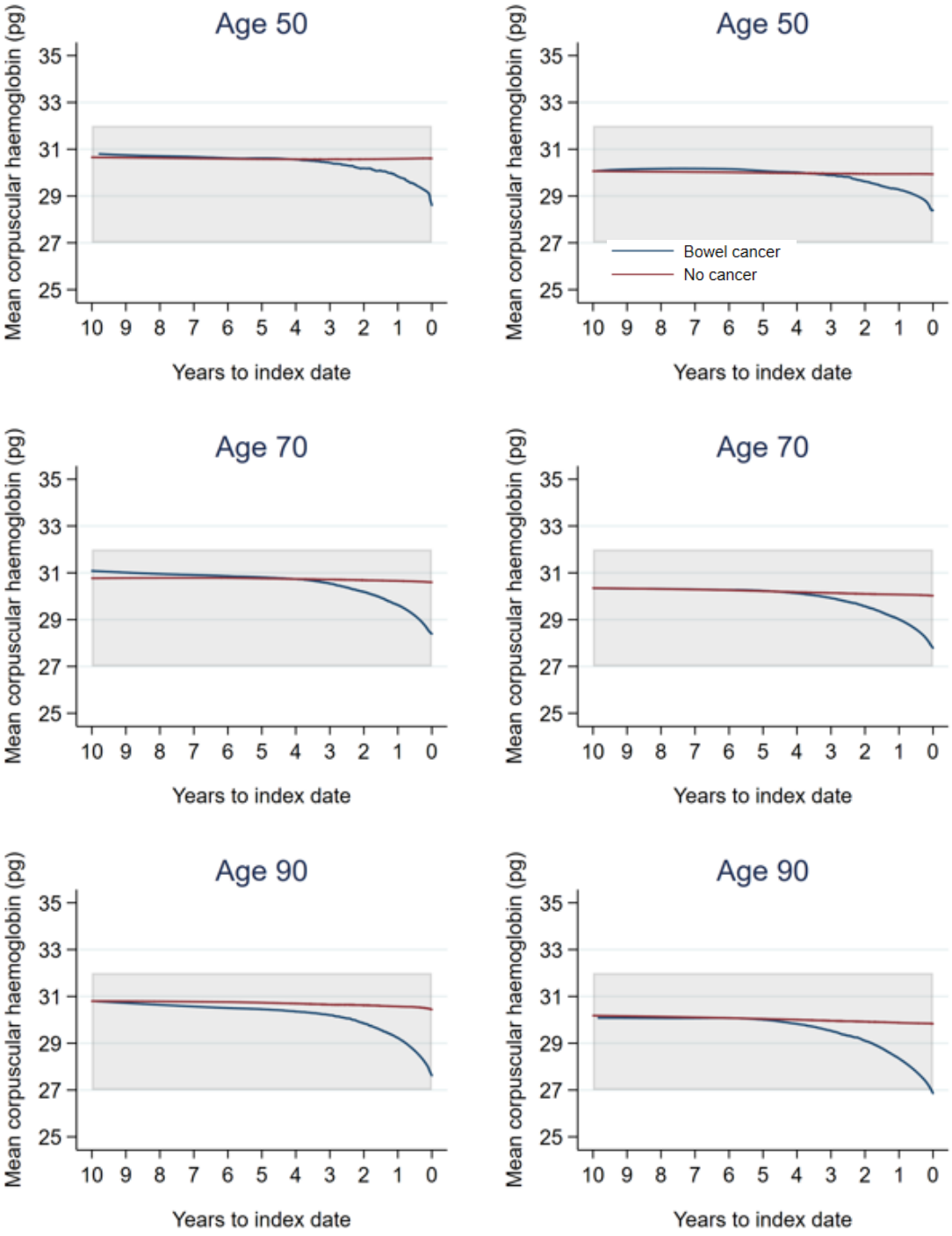
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).
[3]Age 50 – there is a small sample of patients with data available around the 10-year mark, effecting the shape of trajectories.

Figure C4: Trajectories (LOWESS) in mean corpuscular volume for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



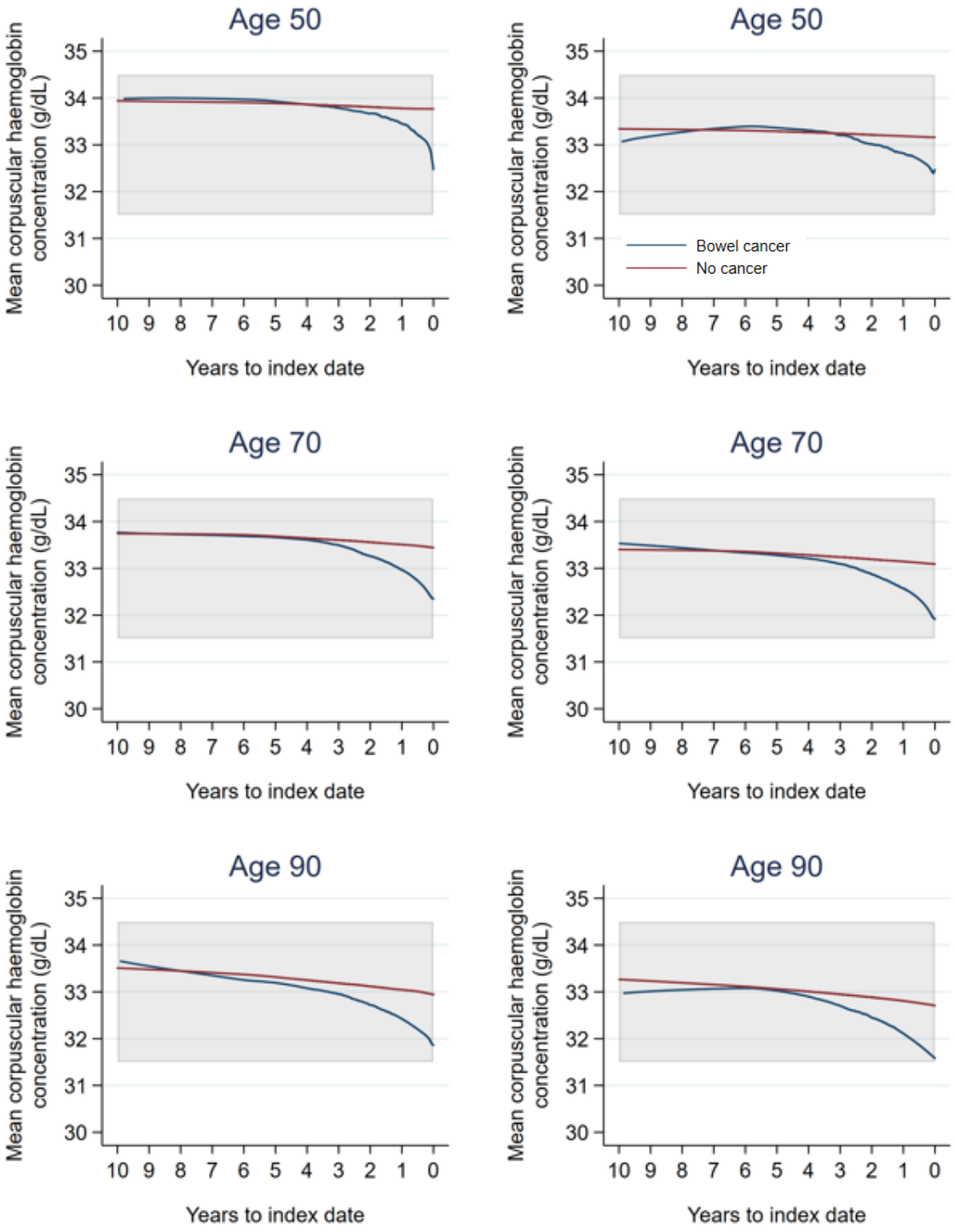
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C5: Trajectories (LOWESS) in mean corpuscular haemoglobin for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



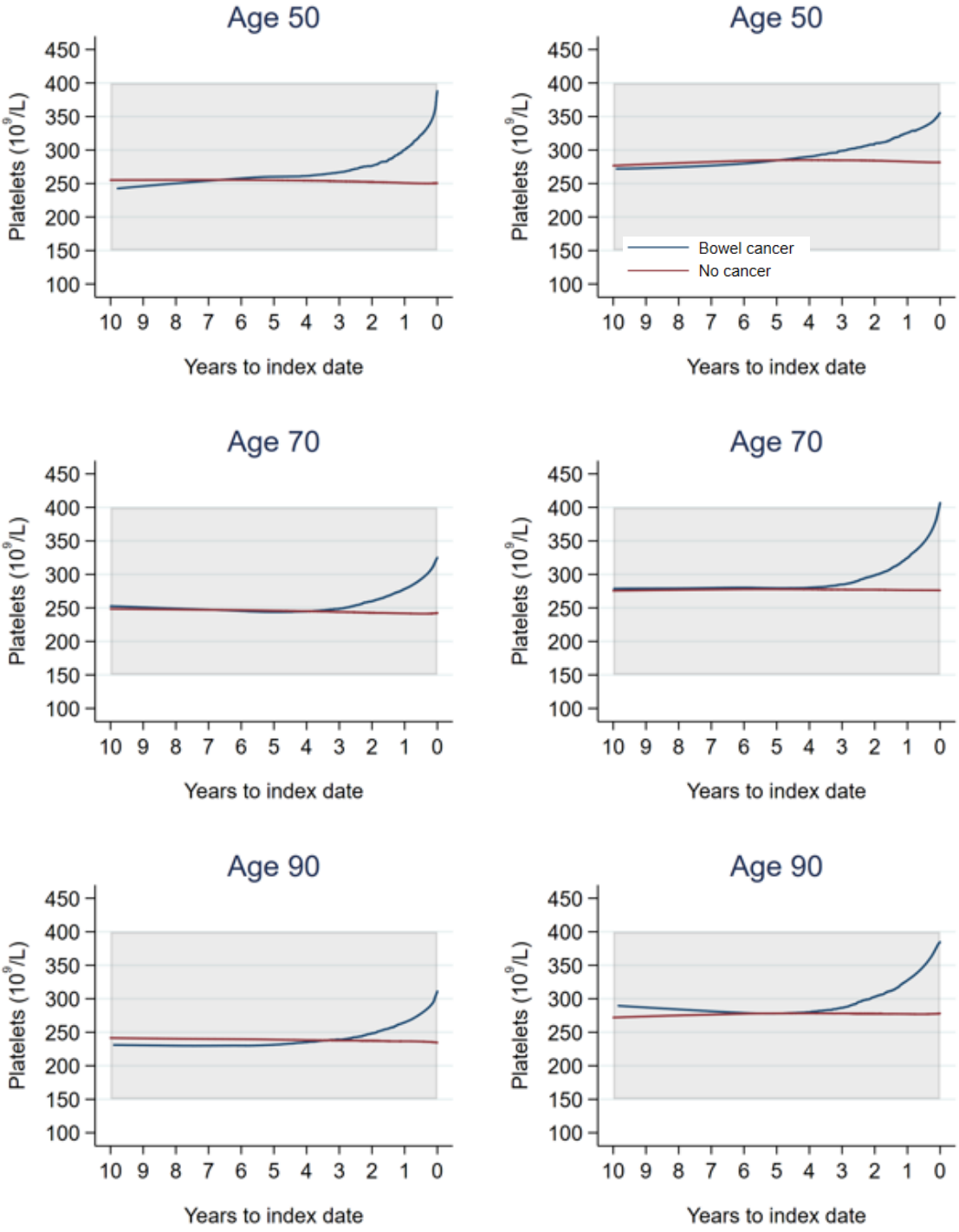
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C6: Trajectories (LOWESS) in mean corpuscular haemoglobin concentration for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



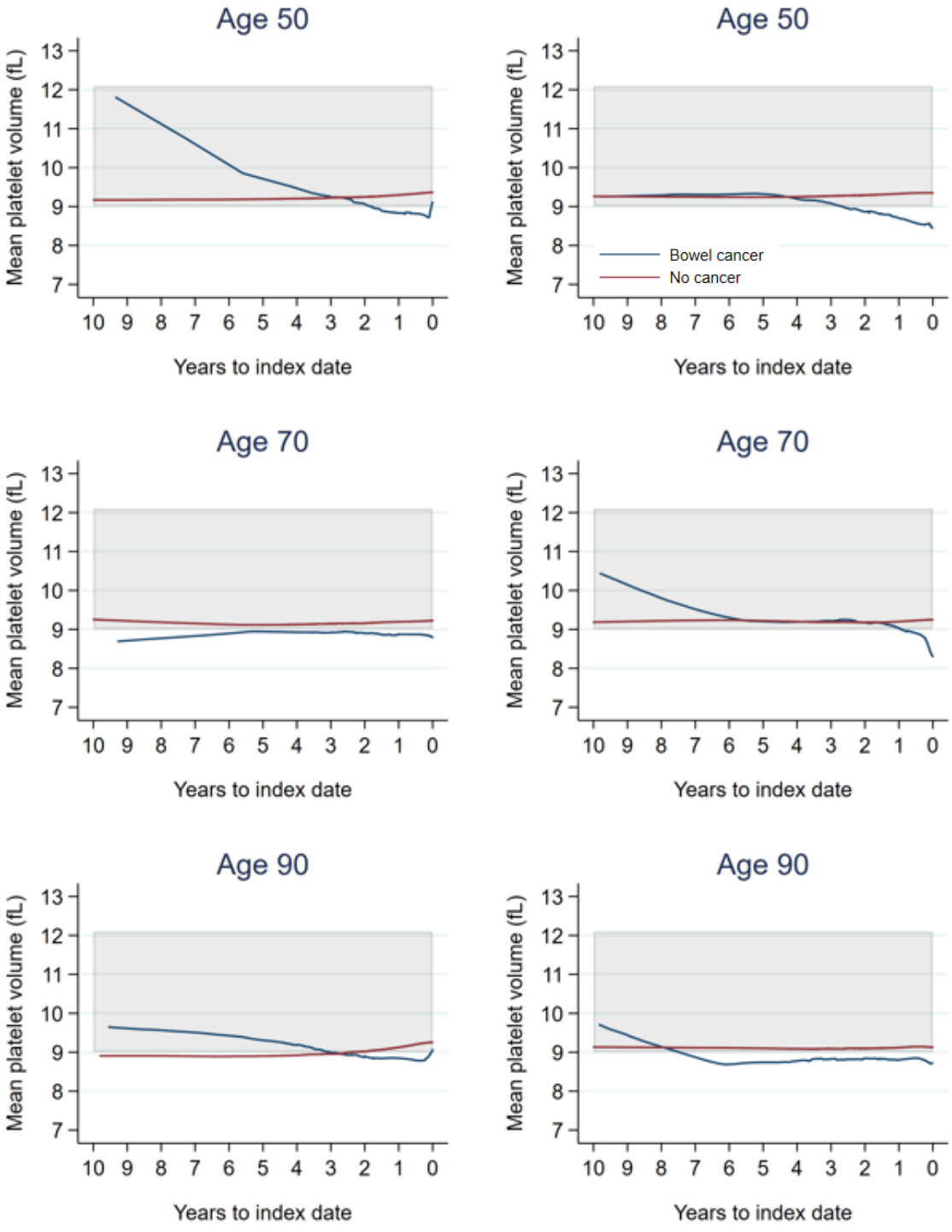
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C7: Trajectories (LOWESS) in platelets for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



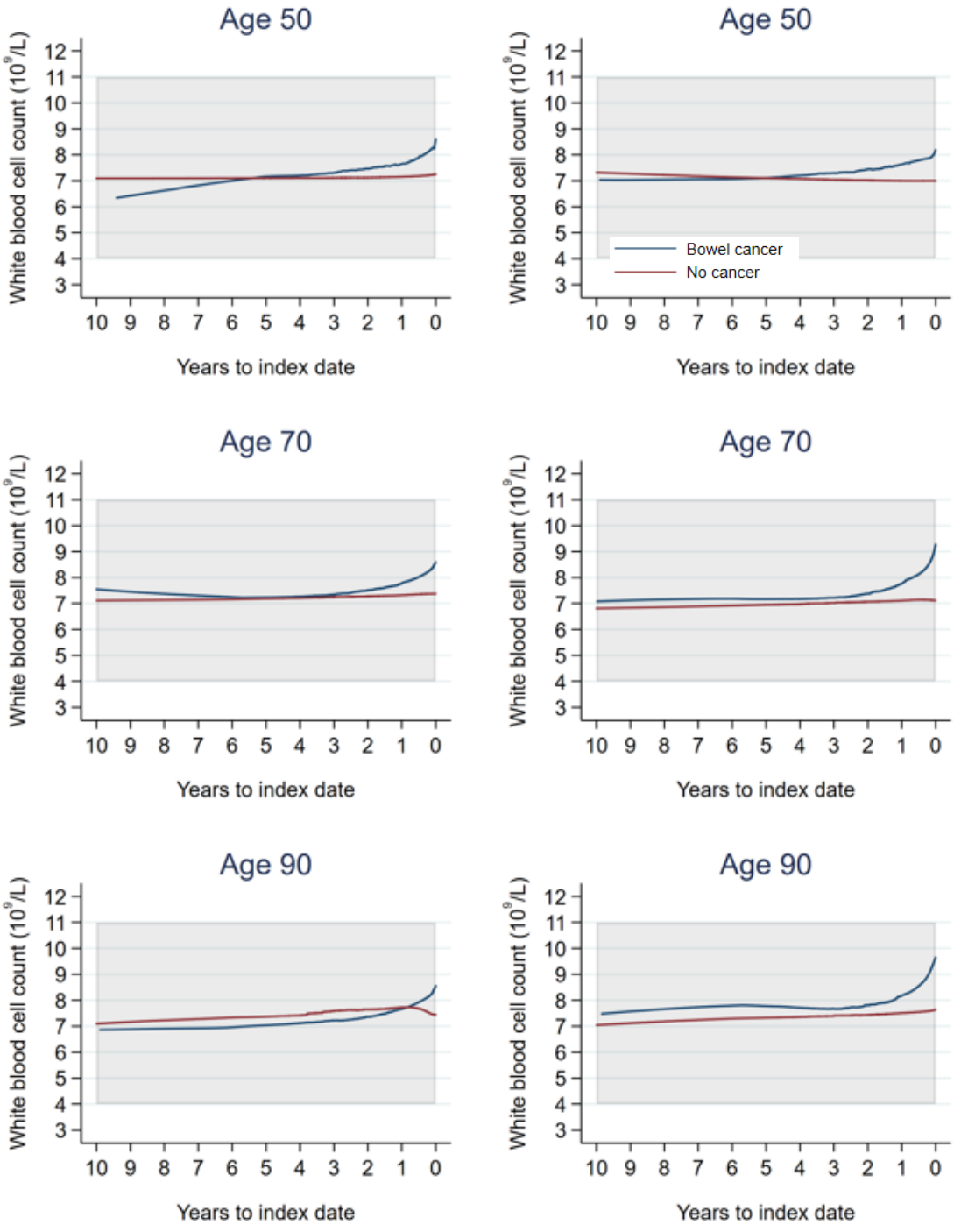
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C8: Trajectories (LOWESS) in mean platelet volume for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2, 3]



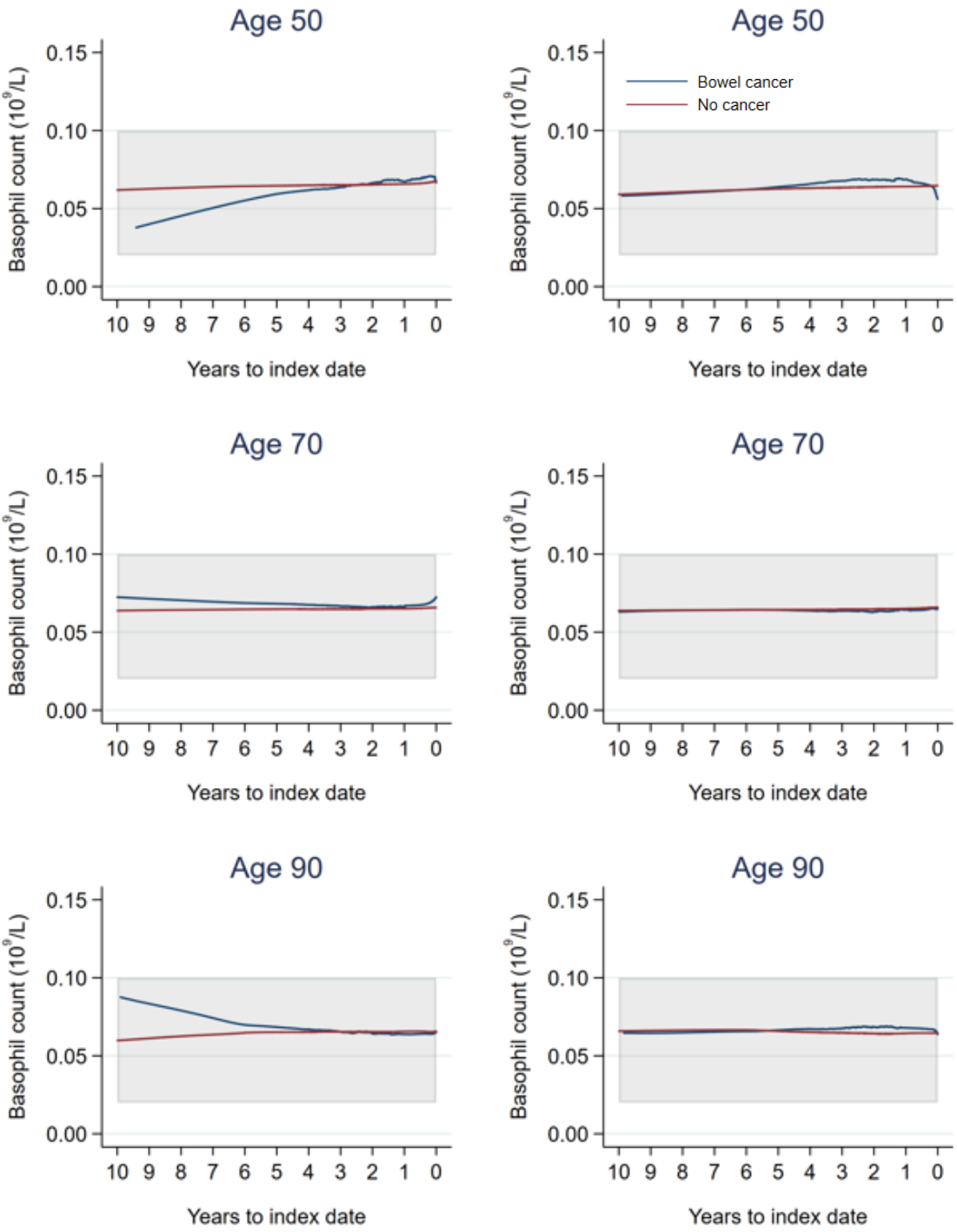
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).
[3]Age 50 – there is a small sample of male patients with data available around the 10-year mark, effecting the shape of trajectories.

Figure C9: Trajectories (LOWESS) in white blood cell count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



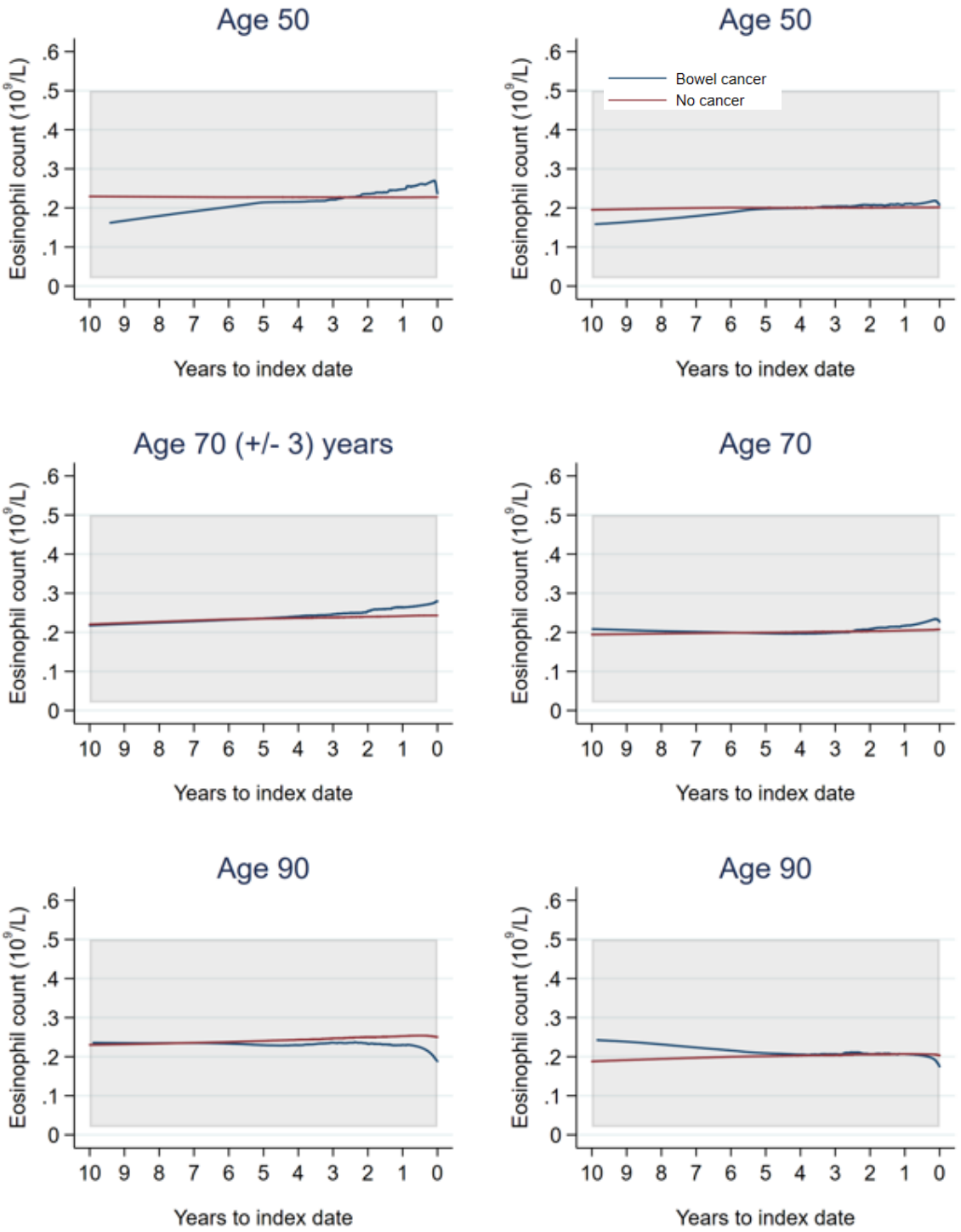
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C10: Trajectories (LOWESS) in basophil count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2, 3]



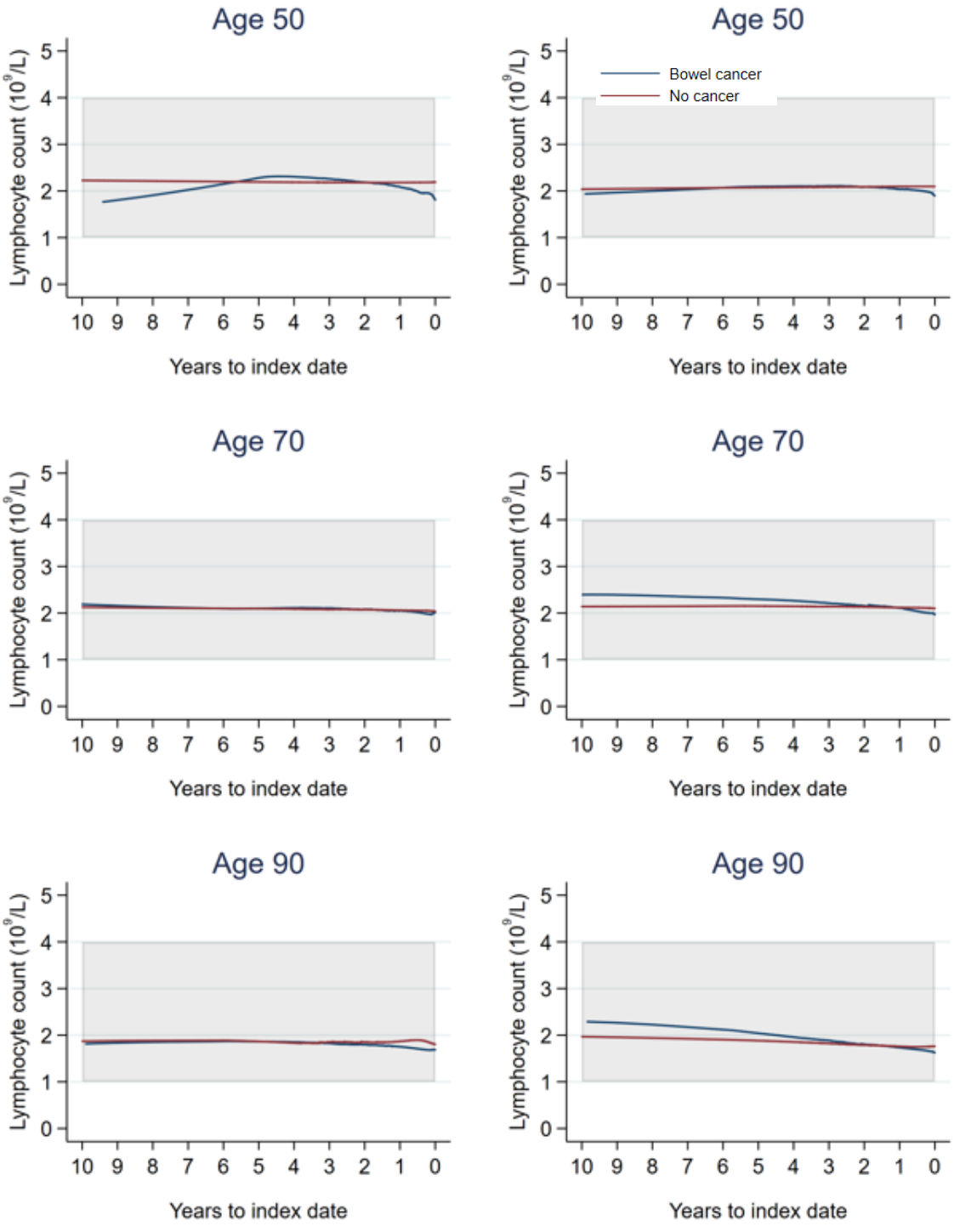
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).
[3]Age 50 and 90 – there is a small sample of male patients with data available around the 10-year mark, effecting the shape of trajectories.

Figure C11: Trajectories (LOWESS) in eosinophil count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



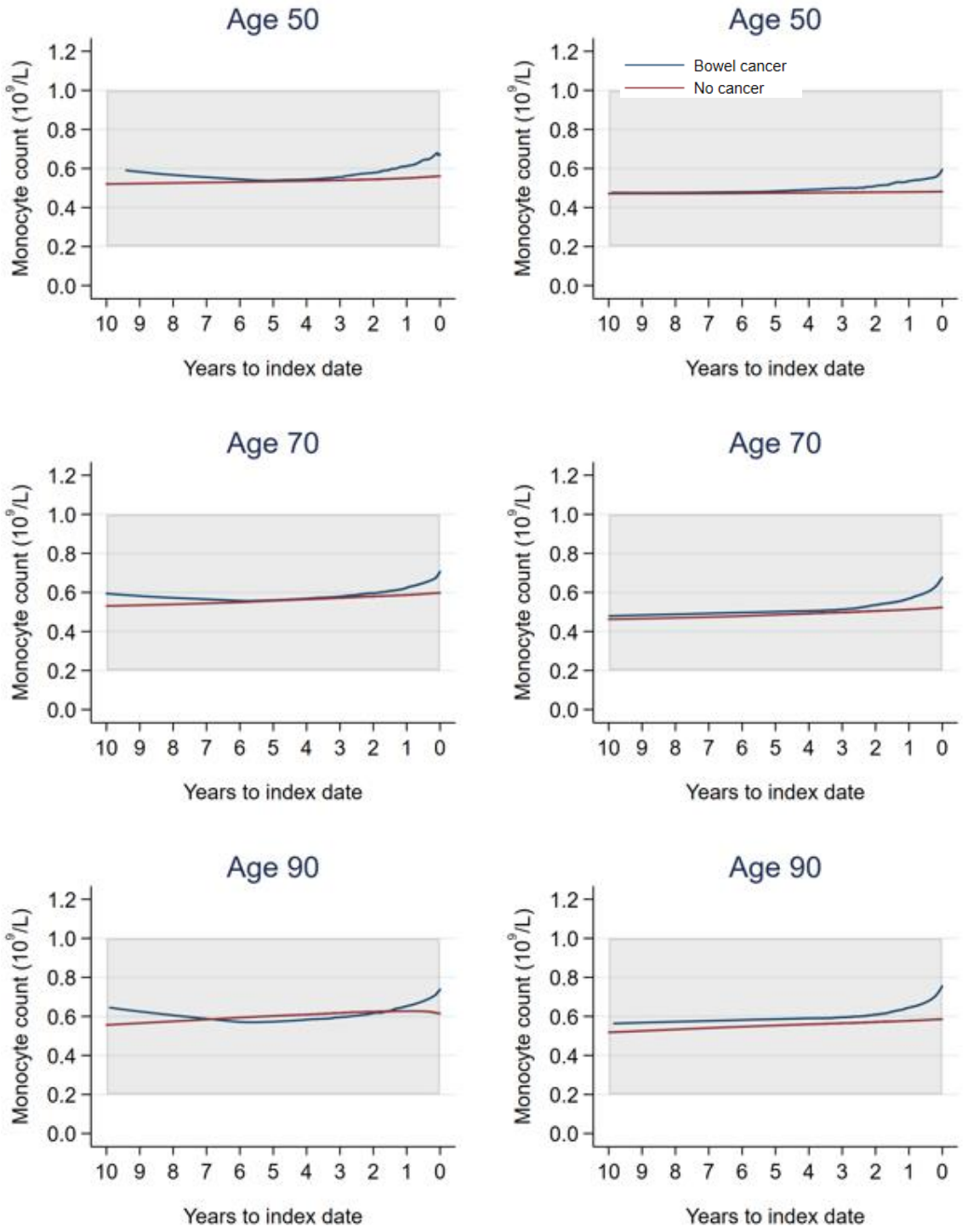
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C12: Trajectories (LOWESS) in lymphocyte count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



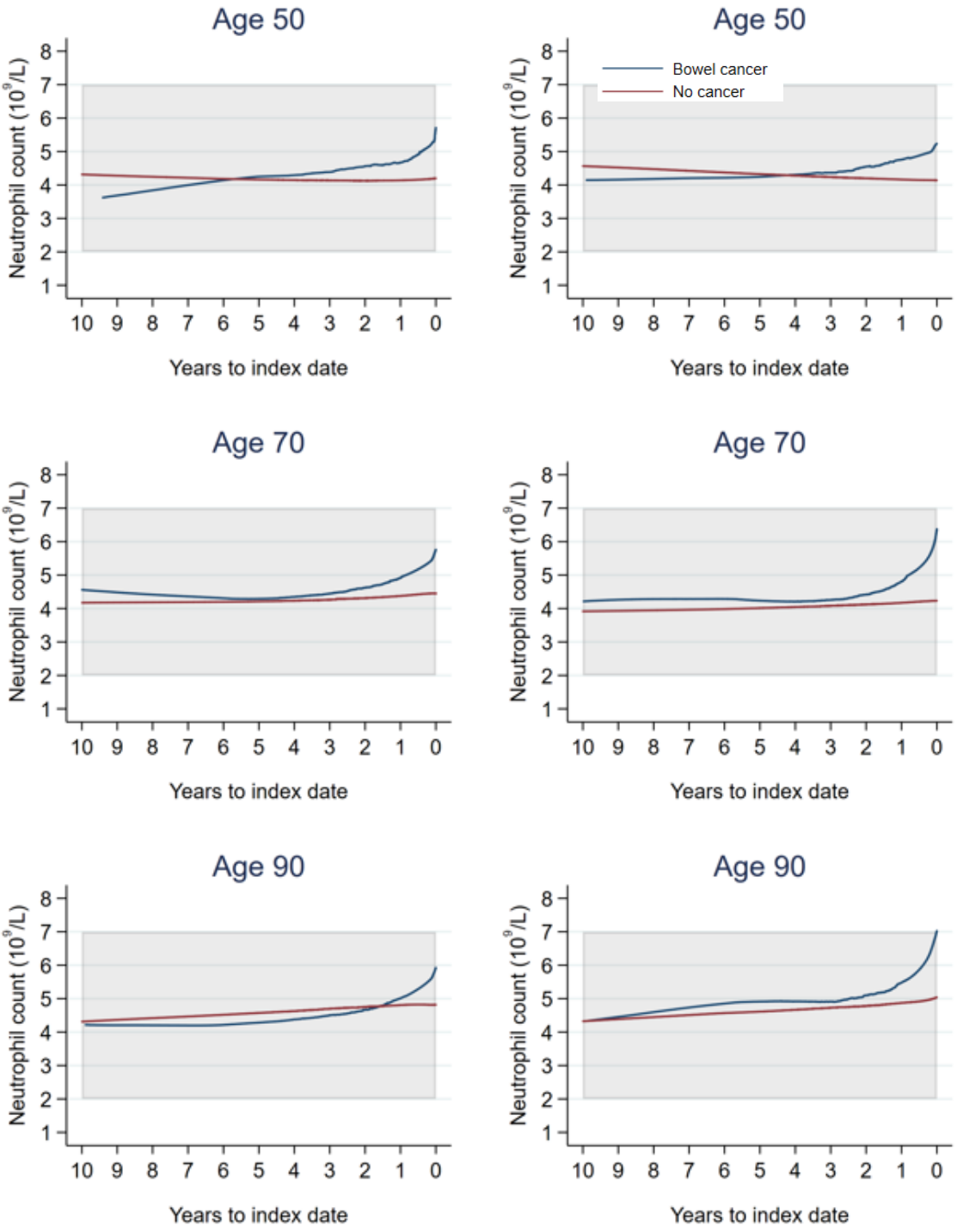
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C13: Trajectories (LOWESS) in monocyte count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

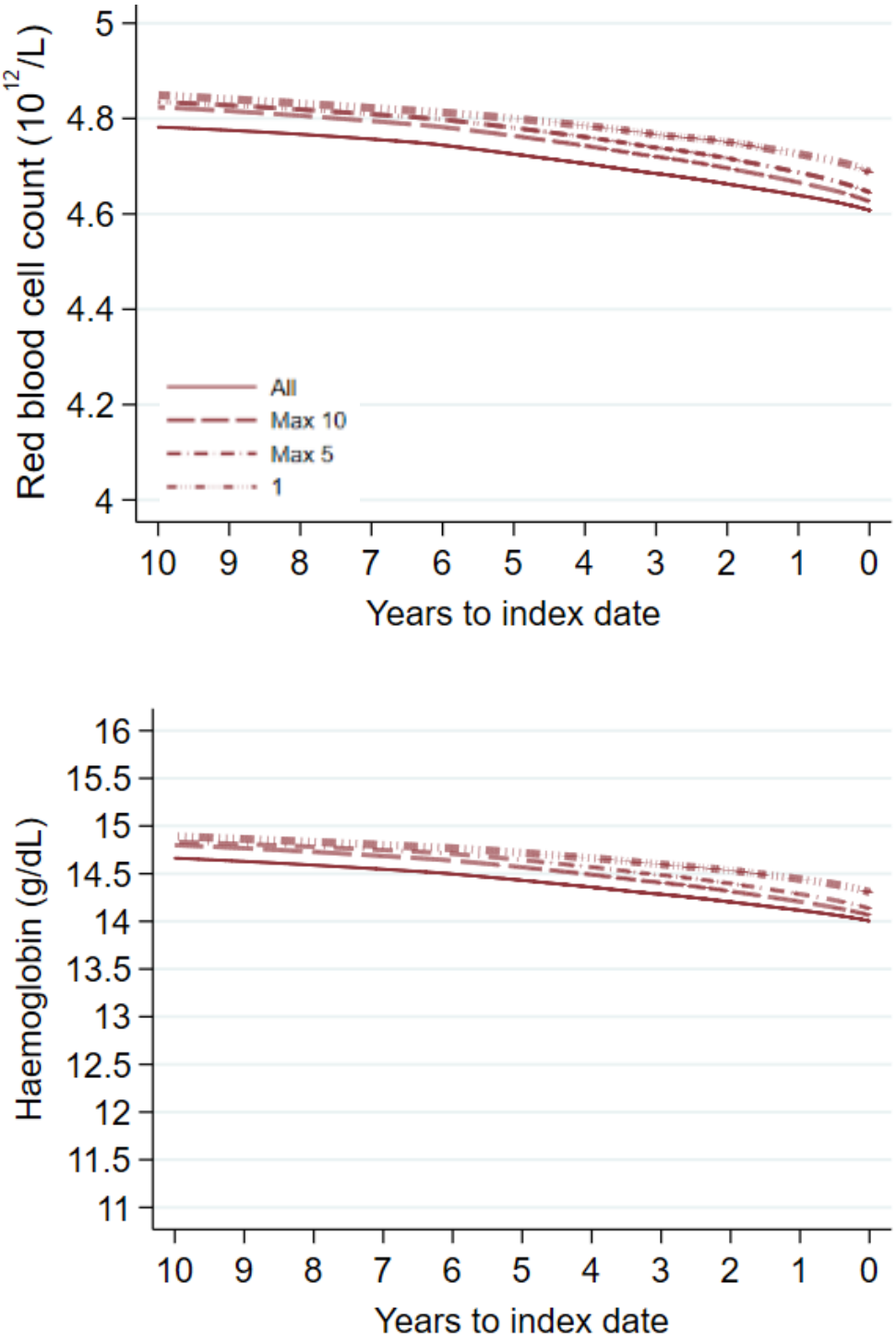
Figure C14: Trajectories (LOWESS) in neutrophil count for males (left) and females (right) with shaded reference range, by bowel cancer status and age at index date[1, 2]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

C3. Influence of number of FBCs on LOWESS trends

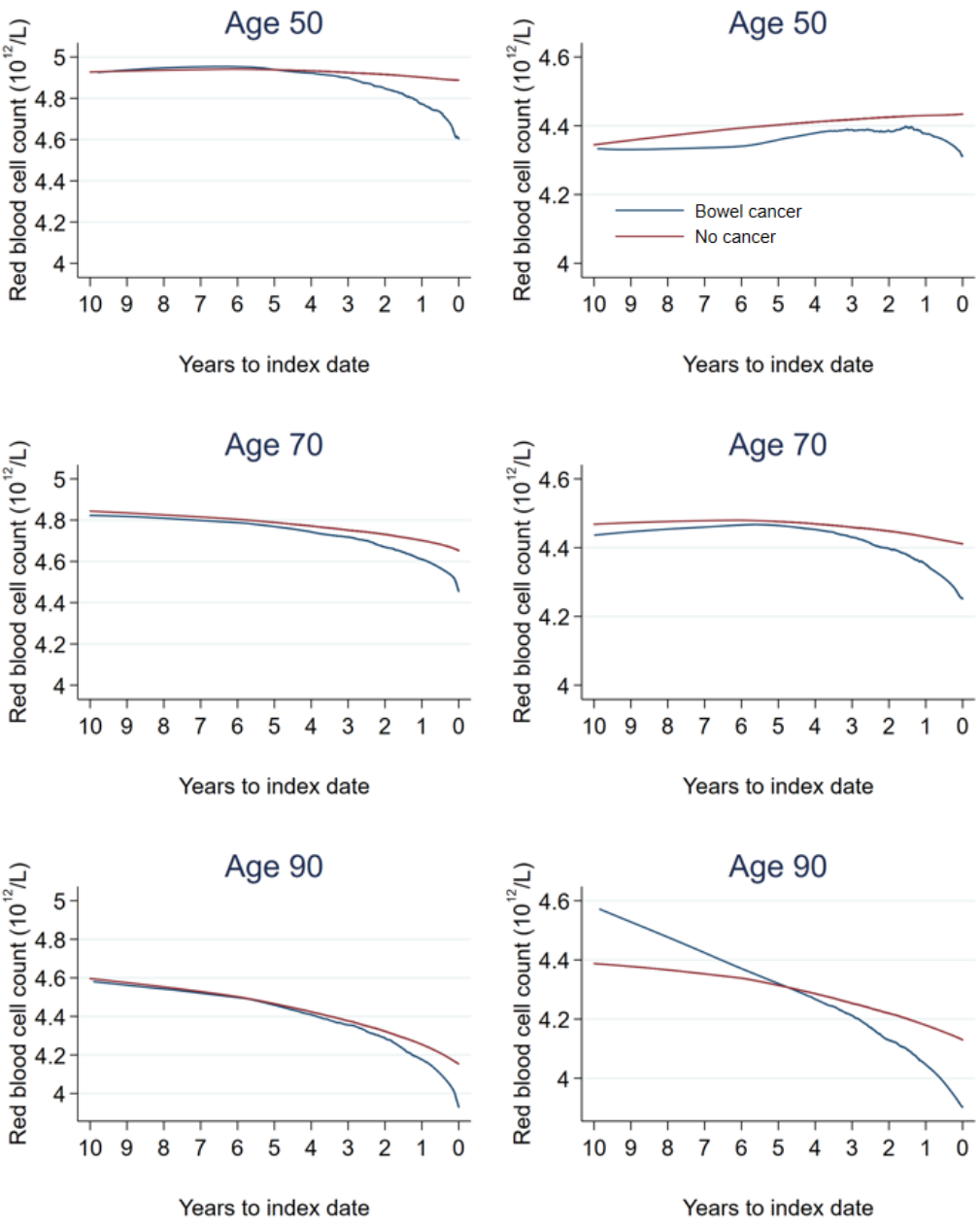
Figure C15: Trajectories (LOWESS) in red blood cell count (top) and haemoglobin (bottom) for male controls aged 70 years at index date, influenced by the number of FBCs[1, 2]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

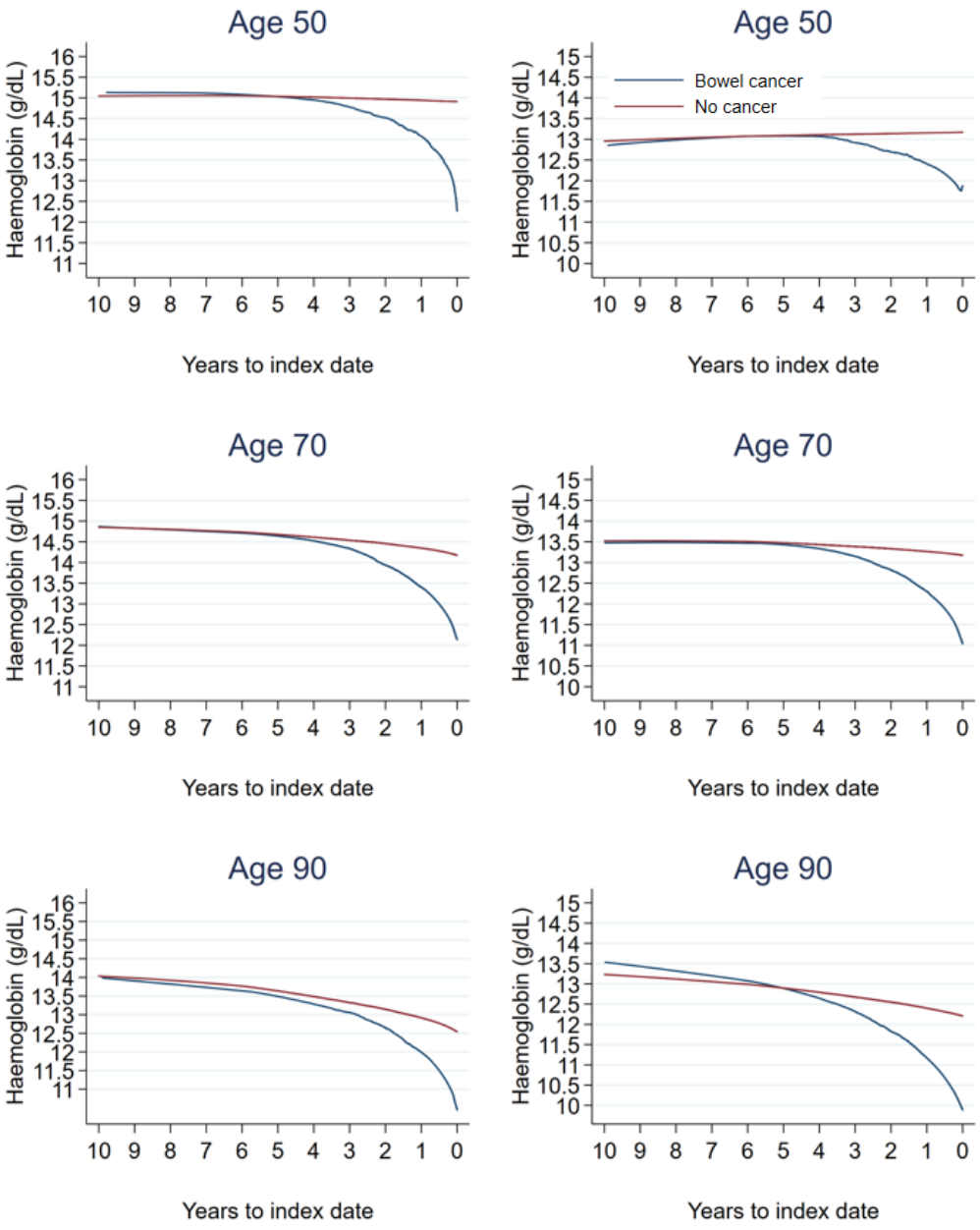
C4. Assessing non-linearity of time to index (years)

Figure C16: Trajectories (LOWESS) in red blood cell count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



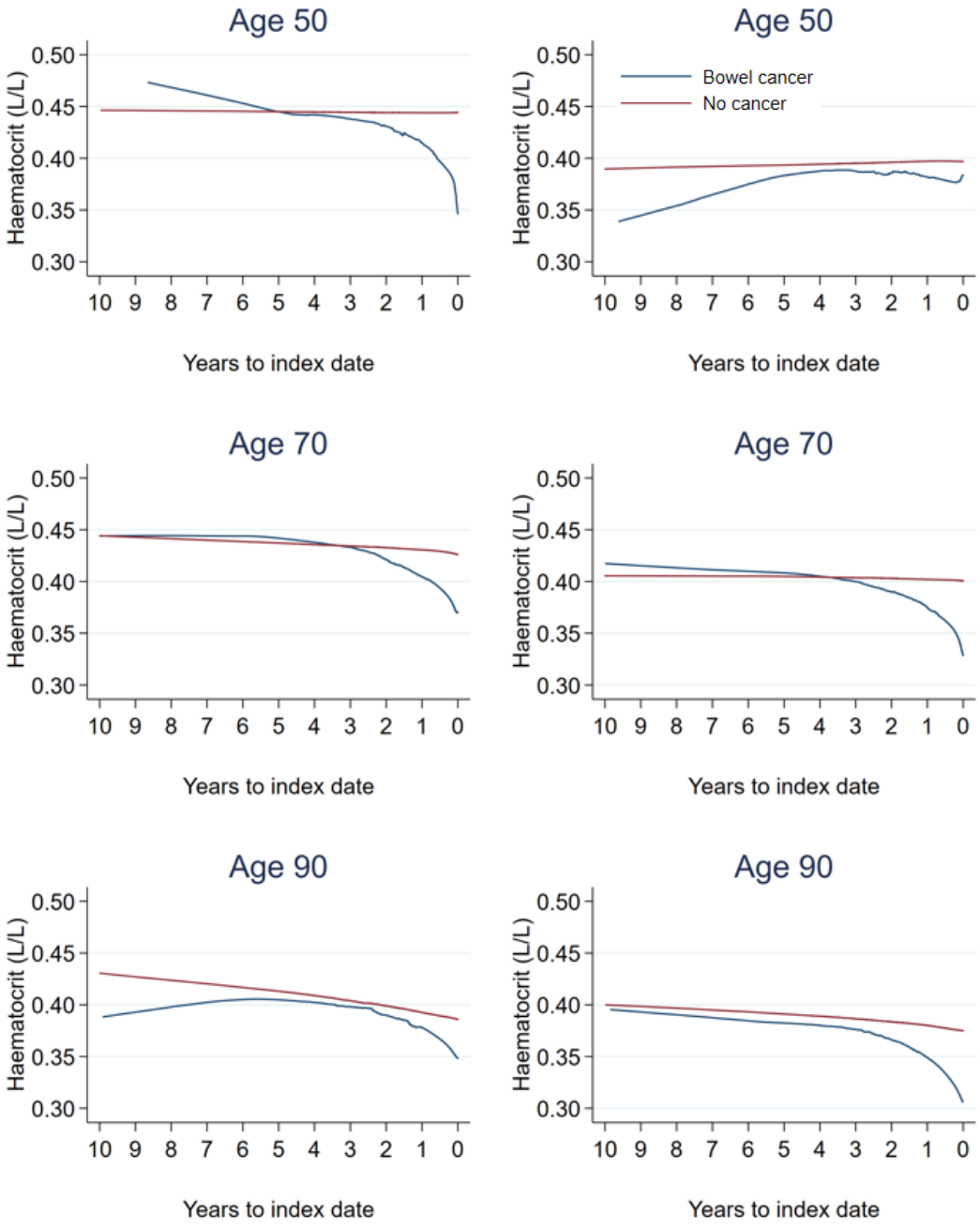
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C17: Trajectories (LOWESS) in haemoglobin for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



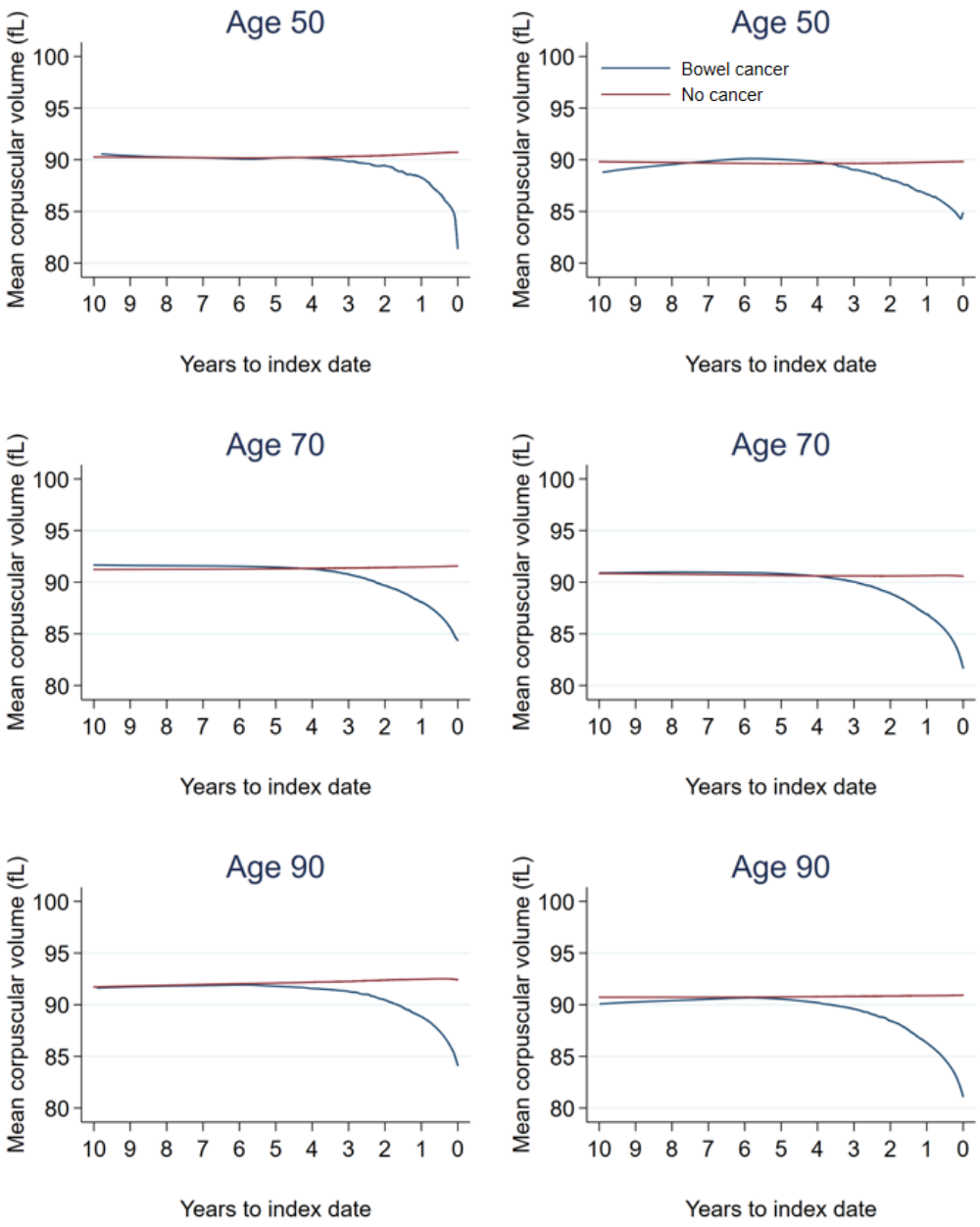
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C18: Trajectories (LOWESS) in haematocrit for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



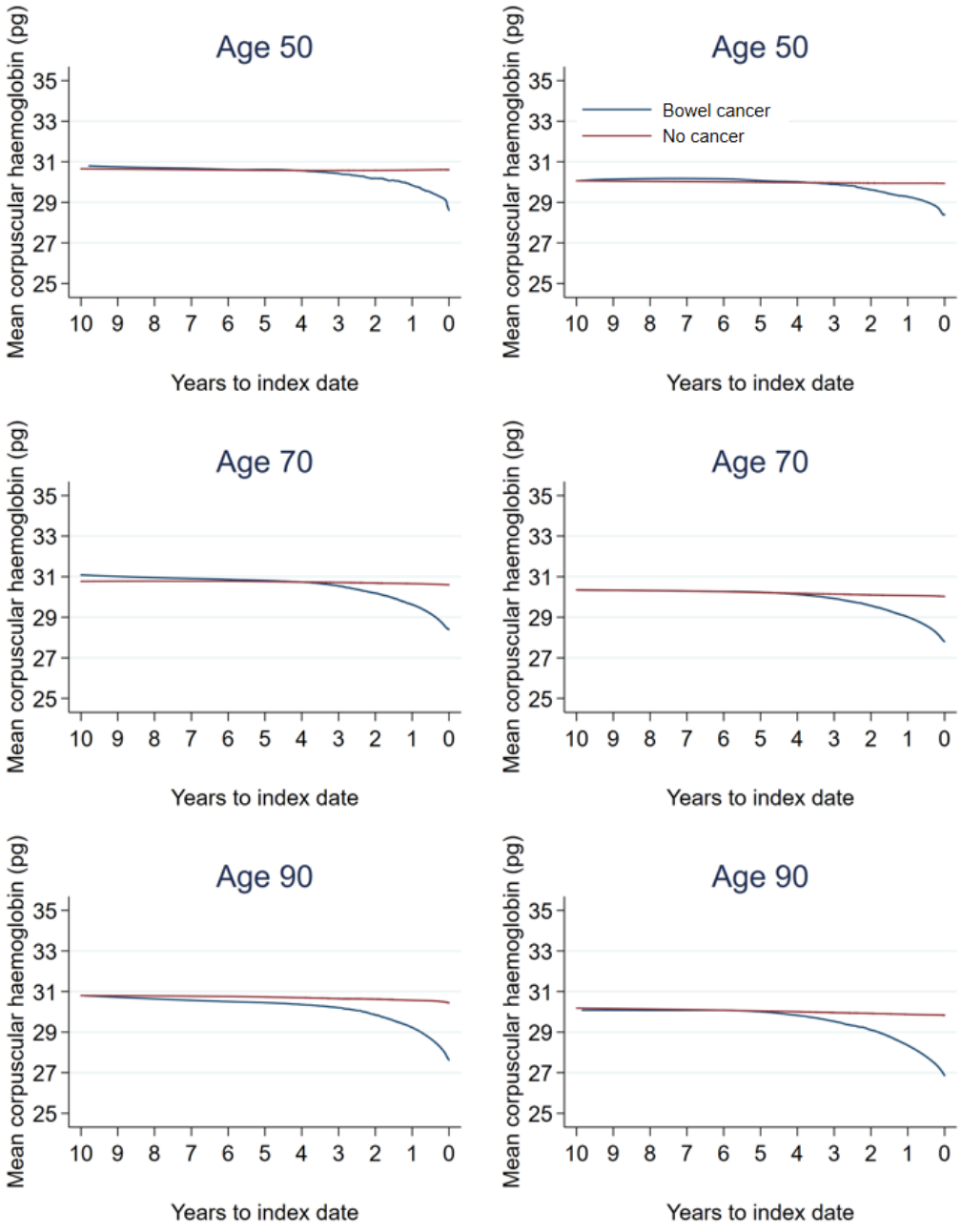
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C19: Trajectories (LOWESS) in mean corpuscular volume for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



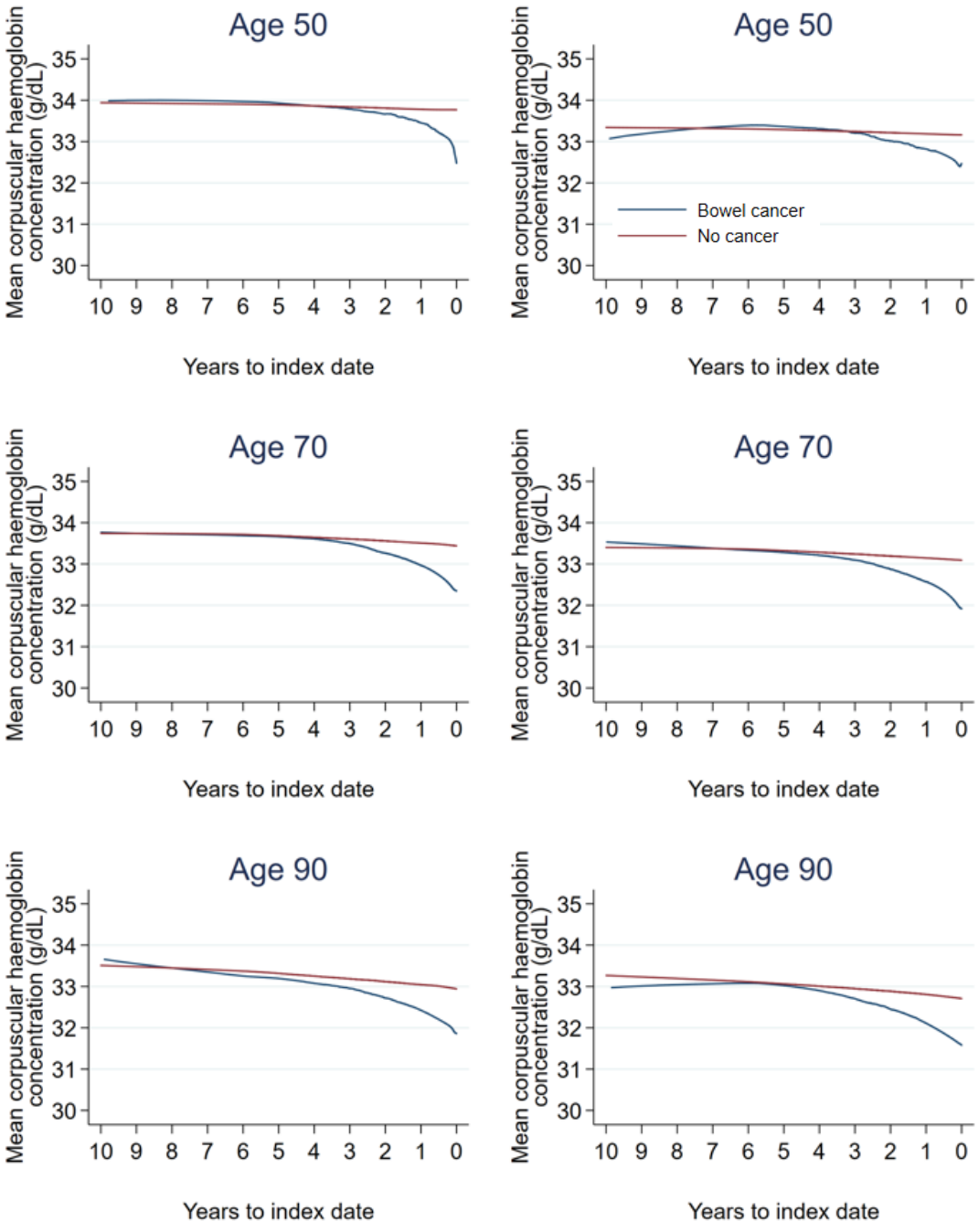
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C20: Trajectories (LOWESS) in mean corpuscular haemoglobin for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



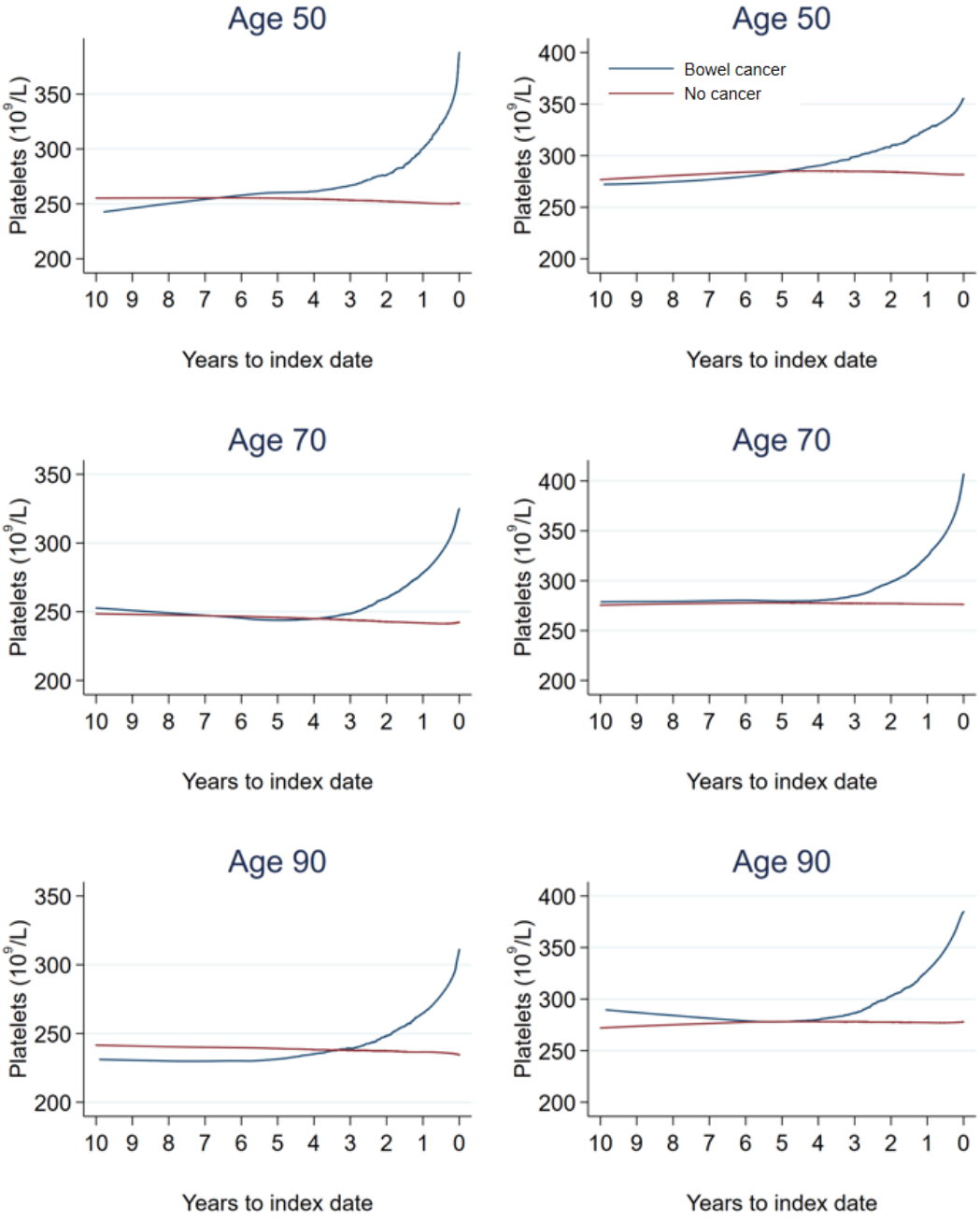
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C21: Trajectories (LOWESS) in mean corpuscular haemoglobin concentration for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



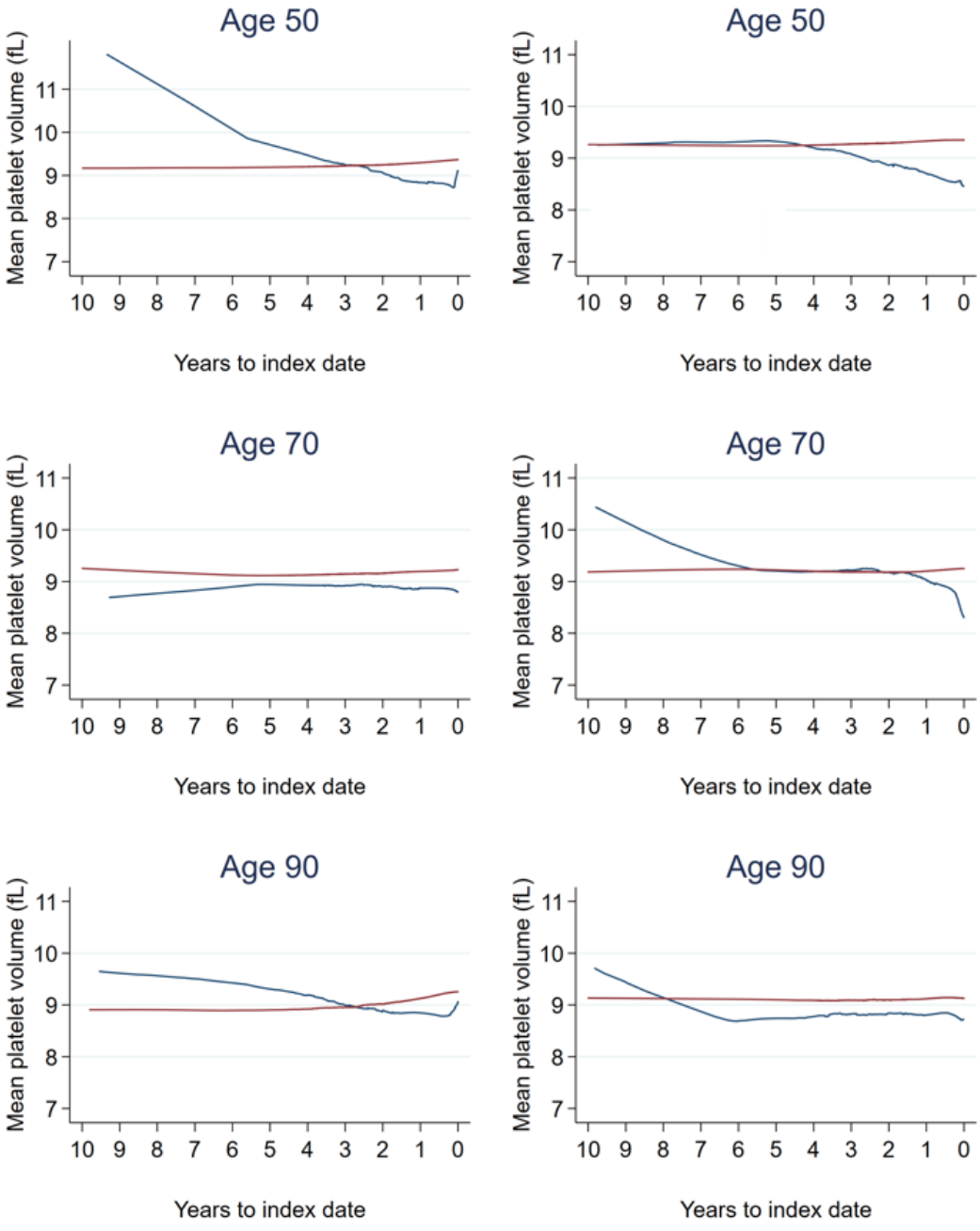
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C22: Trajectories (LOWESS) in platelet count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



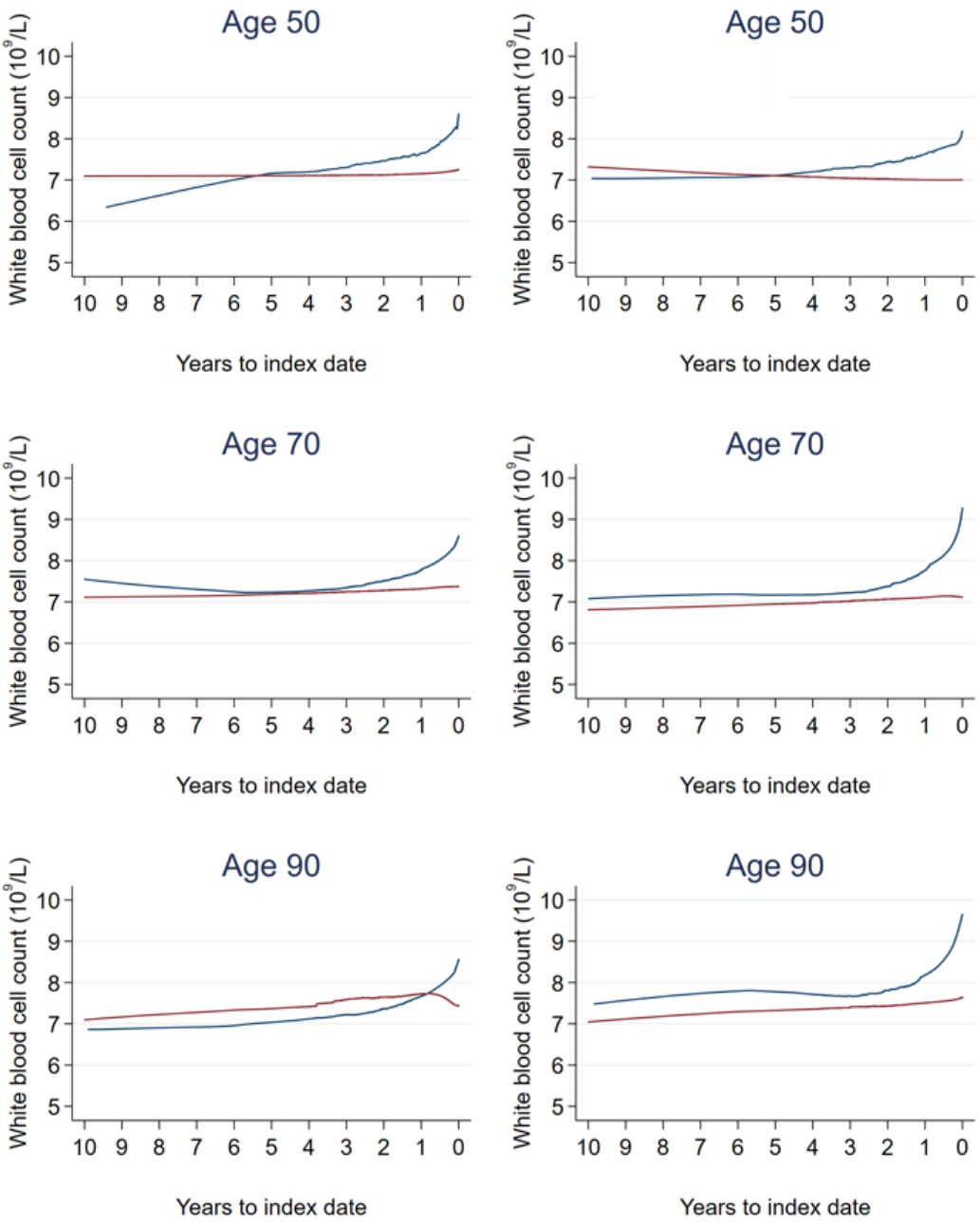
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C23: Trajectories (LOWESS) in mean platelet volume for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



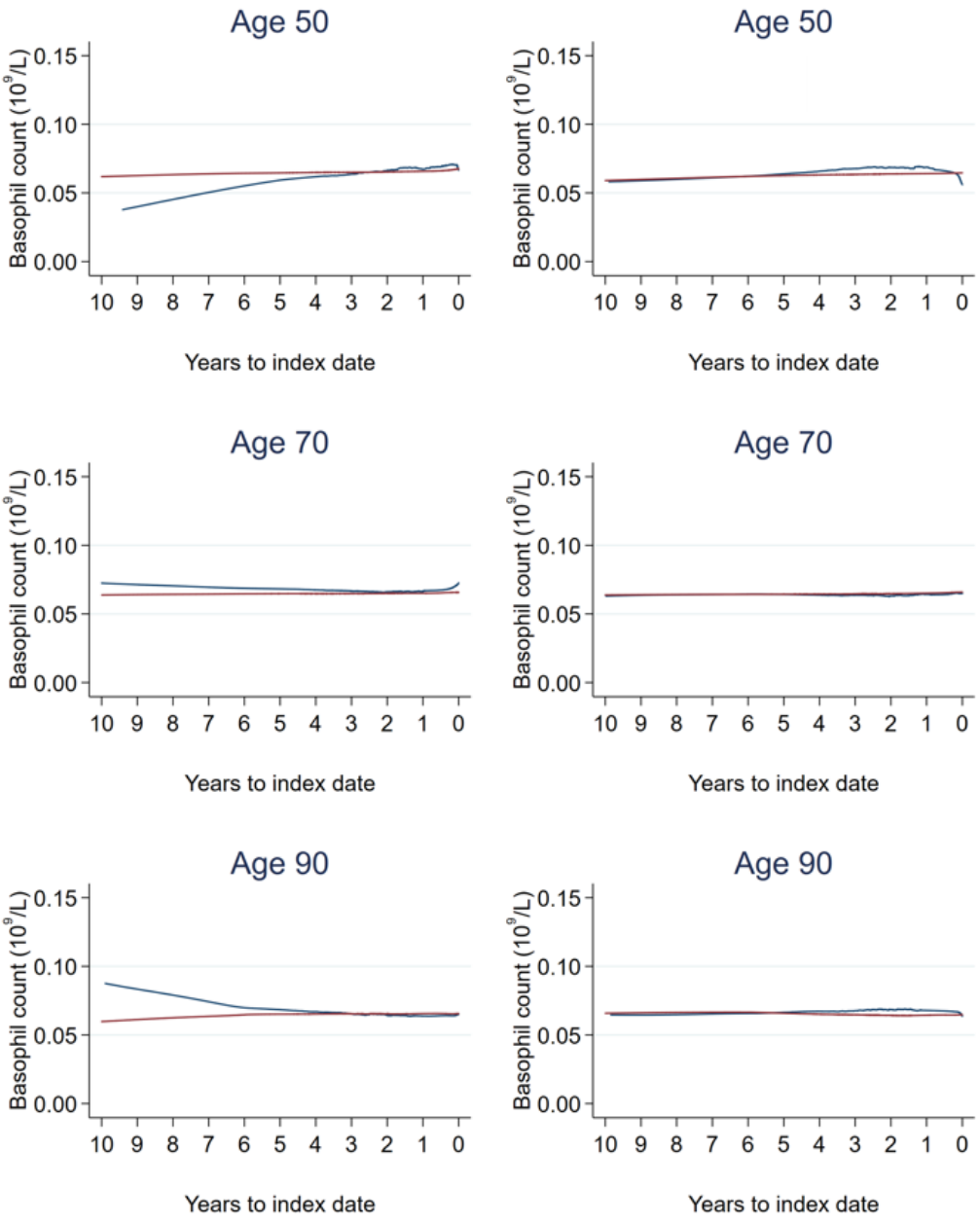
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C24: Trajectories (LOWESS) in white blood cell count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



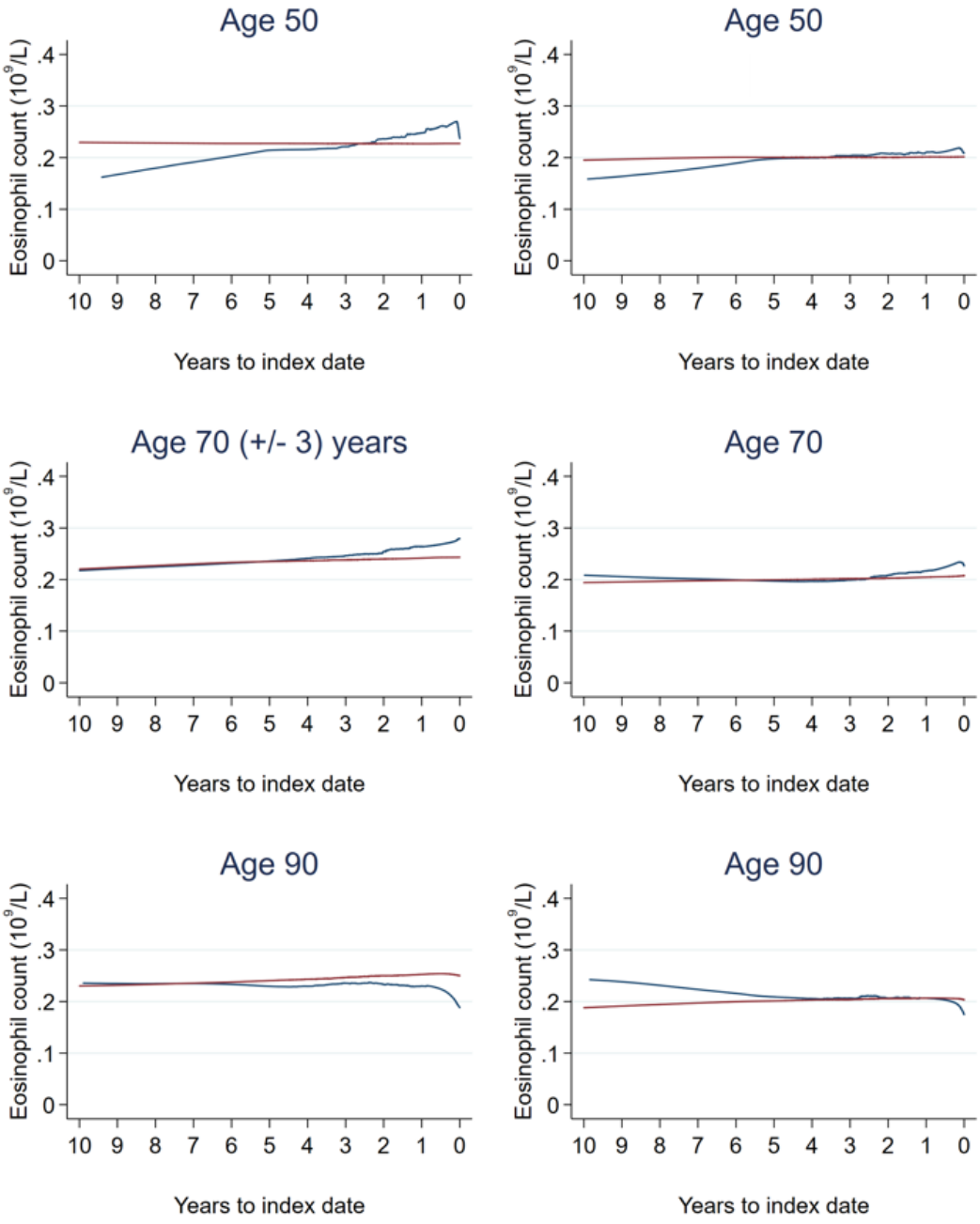
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C25: Trajectories (LOWESS) in basophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



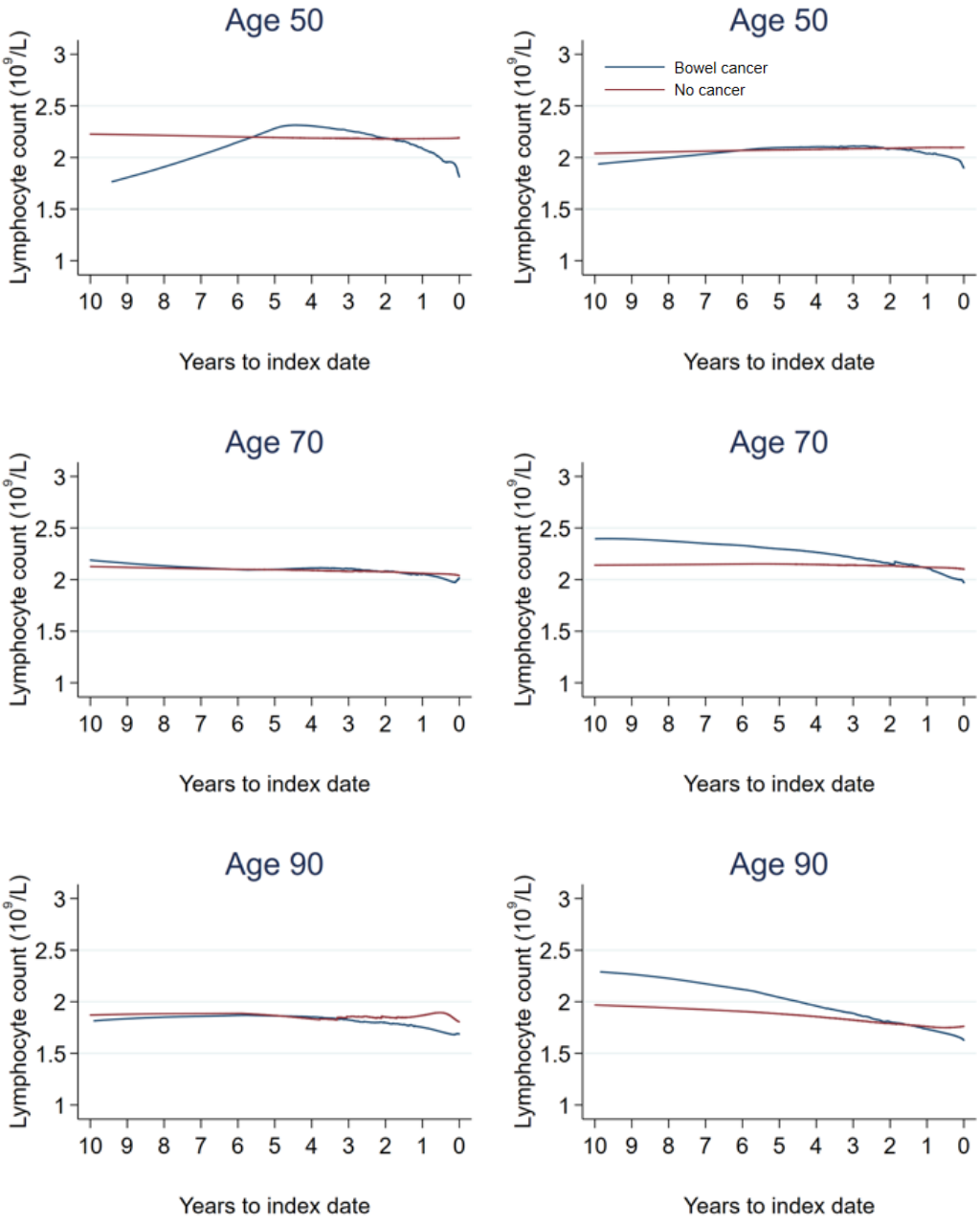
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C26: Trajectories (LOWESS) in eosinophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



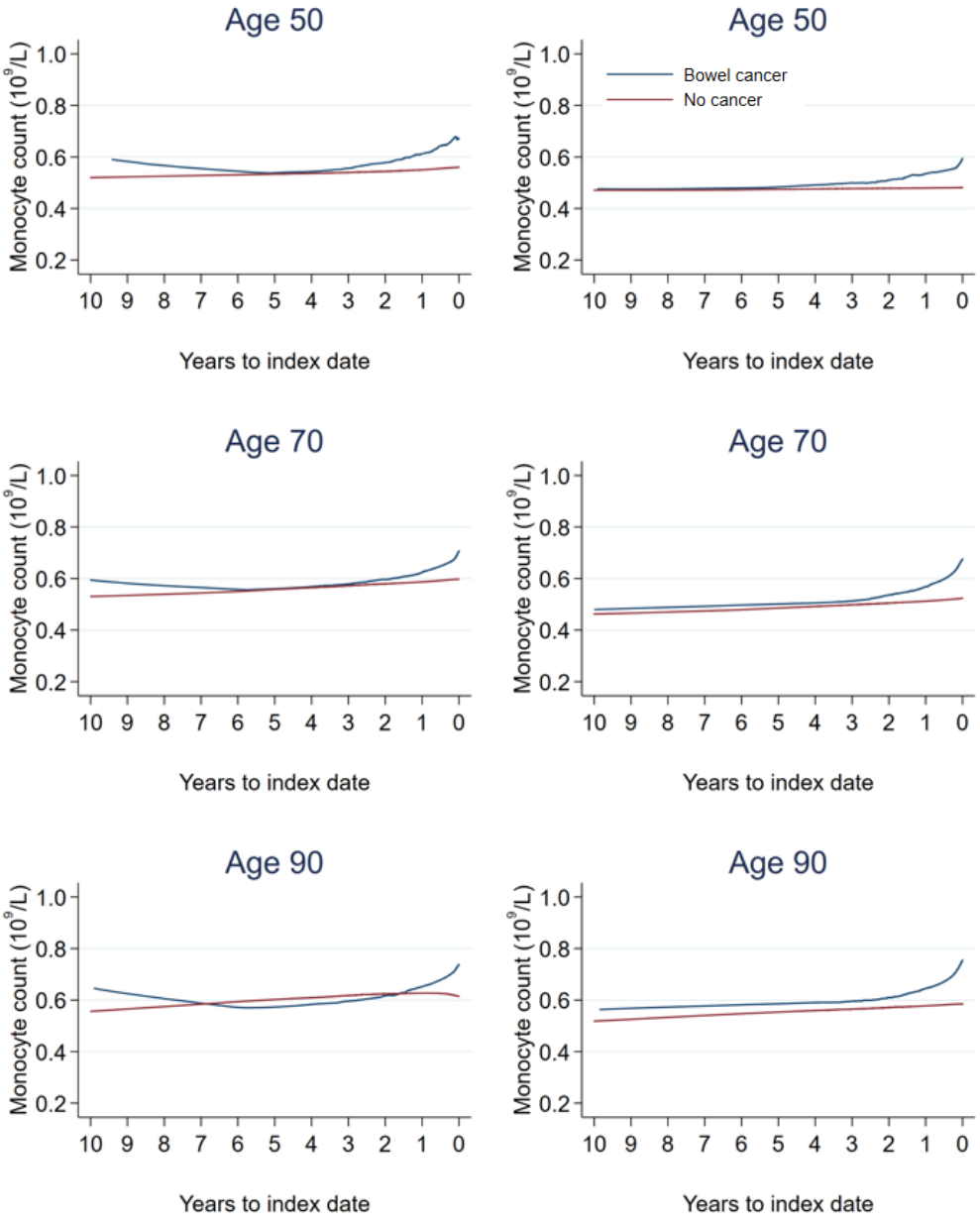
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C27: Trajectories (LOWESS) in lymphocyte count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



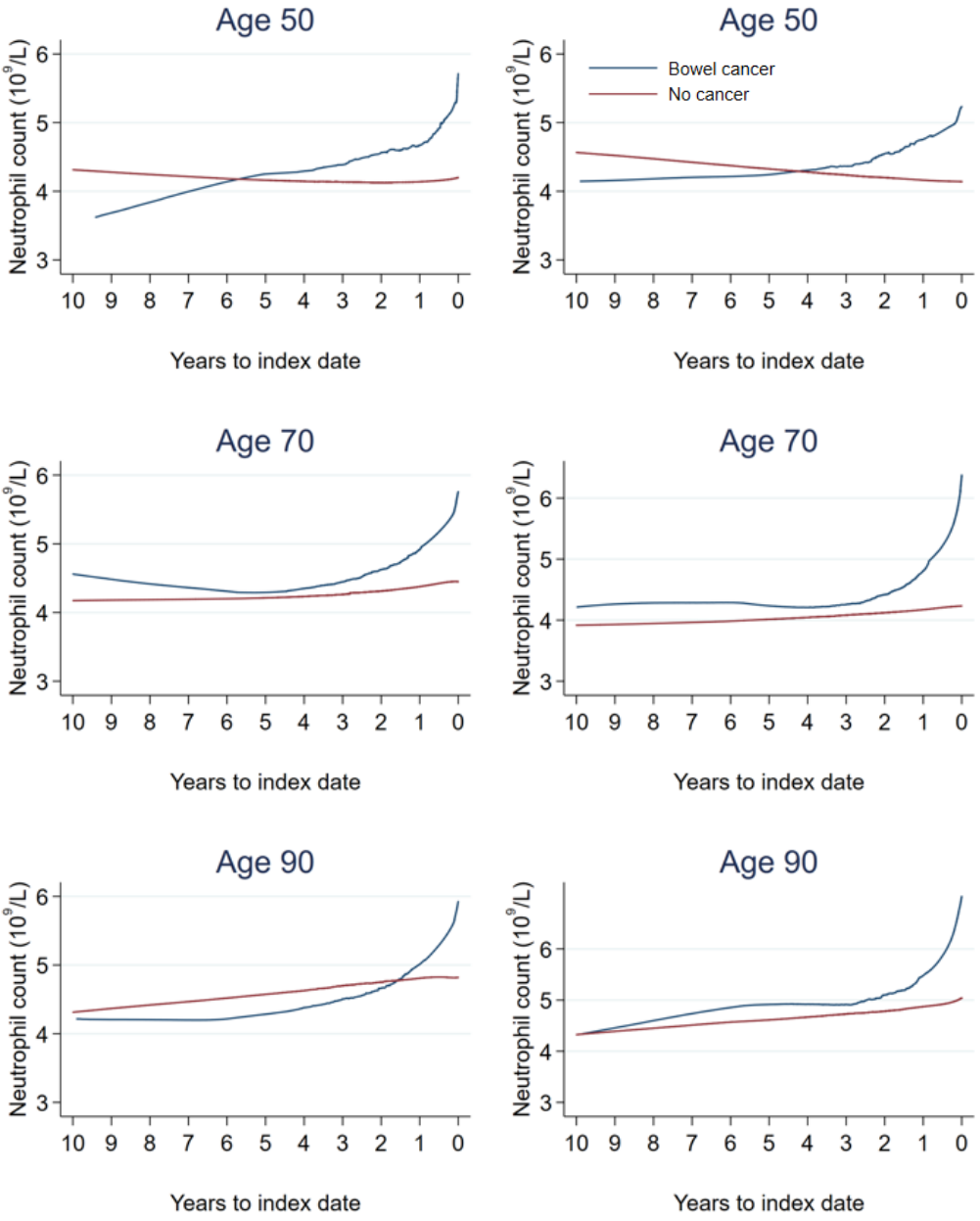
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C28: Trajectories (LOWESS) in monocyte count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Figure C29: Trajectories (LOWESS) in neutrophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

Table C3: Akaike and Bayesian information criteria to compare methods for non-linearity of time to index date[1, 2, 3]

FBC parameter	Males		Females	
	AIC	BIC	AIC	BIC
Red blood cell count (10^9/L)				
LS: knots at 2, 4 years	636462.3	636652.8	645210.2	645407.8
LS: knots at 1, 2, 4 years	636306.6	636532.7	645119.0	645353.7
RCS: knots at 1, 2, 4 years	636553.1	636707.8	645402.8	645563.4
RCS: knots at 1, 2, 3 years	636514.3	636669.0	645419.3	645579.8
RCS: knots at 1, 2, 5 years	636602.9	636757.6	645397.4	645558.0
RCS: knots at 0.1, 2, 5 years	636536.4	636691.1	645357.1	645517.7
RCS: knots at 1, 2, 4, 5 years	636464.9	636655.3	645211.6	645409.2
RCS: knots at 1, 2, 5, 7 years	636503.9	636694.4	645182.2	645379.8
RCS: knots at 1, 2, 5, 8 years	636503.1	636693.6	645173.4	645371.0
RCS: knots at 1, 2, 5, 9 years	636423.2	636649.4		
RCS: knots at 1, 2, 5, 7, 9 years			645141.9	645376.5
FP[4] powers: 0.5, 0.5	636331.5	636486.2		
FP[4] powers: 1, 2			645447.9	645608.4
Haemoglobin (g/dL)				
LS: knots at 2, 4 years	3481581	3481773	5280569	5280767
LS: knots at 1, 2, 4 years	3480679	3480906	5279844	5280080
RCS: knots at 1, 2, 4 years	3481827	3481983	5280984	5281146
RCS: knots at 1, 2, 3 years	3481517	3481673	5280739	5280901
RCS: knots at 1, 2, 5 years	3482118	3482273	5281230	5281391
RCS: knots at 0.1, 2, 5 years	3481807	3481963	5280973	5281135
RCS: knots at 1, 2, 4, 5 years	3481629	3481820	5280555	5280753

RCS: knots at 1, 2, 5, 7 years	3481839	3482030	5280610	5280809
RCS: knots at 1, 2, 5, 8 years	3481853	3482045	5280607	5280806
RCS: knots at 1, 2, 5, 9 years	3481582	3481809		
RCS: knots at 1, 2, 5, 7, 9 years			5280385	5280621
FP[4] powers: 0.5, 1	3480837	3480993		
FP[4] powers: 1, 3			5282990	5283152
Haematocrit (L/L)				
LS: knots at 2, 4 years	-932949.6	-932784.4	-1532662	-1532489
LS: knots at 1, 2, 4 years	-933025.5	-932829.3	-1532770	-1532565
RCS: knots at 1, 2, 4 years	-932963.2	-932829.0	-1532653	-1532513
RCS: knots at 1, 2, 3 years	-933008.4	-932874.2	-1532714	-1532574
RCS: knots at 1, 2, 5 years	-932923.6	-932789.4	-1532595	-1532455
RCS: knots at 0.1, 2, 5 years	-932964.9	-932830.7	-1532645	-1532504
RCS: knots at 1, 2, 4, 5 years	-932945.5	-932780.3	-1532674	-1532501
RCS: knots at 1, 2, 5, 7 years	-932903.3	-932738.1	-1532641	-1532468
RCS: knots at 1, 2, 5, 8 years	-932902.5	-932737.2	-1532644	-1532471
RCS: knots at 1, 2, 5, 9 years	-932894.4	-932698.2		
RCS: knots at 1, 2, 5, 7, 9 years			-1532672	-1532467
FP[4] powers: 0, 1	-932915.9	-932781.6		
FP[4] powers: 3, 3			-1531291	-1531151
Mean corpuscular volume (fL)				
LS: knots at 2, 4 years	6185884	6186039	10060425	10060586
LS: knots at 1, 2, 4 years	6184998	6185177	10059642	10059828
RCS: knots at 1, 2, 4 years	6186244	6186376	10060592	10060728
RCS: knots at 1, 2, 3 years	6186010	6186141	10060388	10060524
RCS: knots at 1, 2, 5 years	6186456	6186588	10060799	10060935
RCS: knots at 0.1, 2, 5 years	6186207	6186338	10060575	10060711

RCS: knots at 1, 2, 4, 5 years	6185917	6186073	10060440	10060601
RCS: knots at 1, 2, 5, 7 years	6186203	6186358	10060640	10060802
RCS: knots at 1, 2, 5, 8 years	6186244	6186399	10060660	10060821
RCS: knots at 1, 2, 5, 9 years	6186003	6186182		
RCS: knots at 1, 2, 5, 7, 9 years			10060416	10060602
FP[4] powers: 0, 0.5	6185982	6186114		
FP[4] powers: 1, 1			10060337	10060474
Mean corpuscular haemoglobin (pg)				
LS: knots at 2, 4 years	3746480	3746635	6152778	6152939
LS: knots at 1, 2, 4 years	3745852	3746031	6152186	6152371
RCS: knots at 1, 2, 4 years	3746591	3746722	6152913	6153049
RCS: knots at 1, 2, 3 years	3746404	3746535	6152701	6152837
RCS: knots at 1, 2, 5 years	3746771	3746902	6153143	6153279
RCS: knots at 0.1, 2, 5 years	3746574	3746705	6152942	6153078
RCS: knots at 1, 2, 4, 5 years	3746514	3746669	6152742	6152903
RCS: knots at 1, 2, 5, 7 years	3746637	3746792	6152830	6152991
RCS: knots at 1, 2, 5, 8 years	3746644	3746799	6152834	6152995
RCS: knots at 1, 2, 5, 9 years	3746554	3746733		
RCS: knots at 1, 2, 5, 7, 9 years			6152689	6152874
FP[4] powers: 0, 1	3746595	3746726		
FP[4] powers: 3, 3			6158168	6158304
Mean corpuscular haemoglobin concentration (g/dL)				
LS: knots at 2, 4 years	2810427	2810581	4533542	4533702
LS: knots at 1, 2, 4 years	2810202	2810380	4533436	4533621
RCS: knots at 1, 2, 4 years	2810579	2810709	4533877	4534013
RCS: knots at 1, 2, 3 years	2810546	2810676	4533845	4533980
RCS: knots at 1, 2, 5 years	2810603	2810734	4533899	4534034

RCS: knots at 0.1, 2, 5 years	2810540	2810670	4533859	4533994
RCS: knots at 1, 2, 4, 5 years	2810436	2810590	4533523	4533683
RCS: knots at 1, 2, 5, 7 years	2810411	2810565	4533448	4533608
RCS: knots at 1, 2, 5, 8 years	2810423	2810577	4533471	4533631
RCS: knots at 1, 2, 5, 9 years	2810389	2810566		
RCS: knots at 1, 2, 5, 7, 9 years			4533413	4533597
FP[4] powers: 1, 2	2810742	2810872		
FP[4] powers: 1, 3			4534096	4534231
Platelets (10^{12}/L)				
LS: knots at 2, 4 years	12178716	12178871	19241274	19241434
LS: knots at 1, 2, 4 years	12178142	12178321	19240478	19240663
RCS: knots at 1, 2, 4 years	12179151	12179283	19241814	19241950
RCS: knots at 1, 2, 3 years	12178996	12179127	19241636	19241772
RCS: knots at 1, 2, 5 years	12179287	12179418	19241987	19242124
RCS: knots at 0.1, 2, 5 years	12179125	12179256	19241778	19241914
RCS: knots at 1, 2, 4, 5 years	12178728	12178883	19241270	19241431
RCS: knots at 1, 2, 5, 7 years	12178853	12179008	19241394	19241555
RCS: knots at 1, 2, 5, 8 years	12178875	12179030	19241417	19241578
RCS: knots at 1, 2, 5, 9 years	12178797	12178976		
RCS: knots at 1, 2, 5, 7, 9 years			19241342	19241527
FP[4] powers: 2, 2	12181342	12181473	19244940	19245076
Mean platelet volume (fL)				
Linear	403173.5	403262.9	618668	618761
LS: knots at 5 years	403115.1	403224.4	618638	618753
LS: knots at 2, 4 years	403038.1	403167.3	618428	618563
LS: knots at 1, 2, 4 years	403044.0	403193.1	618424	618580
RCS: knots at 1, 2, 4 years	403017.6	403127.0	618424	618539

RCS: knots at 1, 2, 3 years	403021.0	403130.4	618413	618528
RCS: knots at 1, 2, 5 years	403020.7	403130.1	618446	618561
RCS: knots at 0.1, 2, 5 years	403017.5	403126.8	618437	618551
RCS: knots at 1, 2, 4, 5 years	403031.4	403160.6	618411	618546
RCS: knots at 1, 2, 5, 7 years	403032.6	403161.8	618405	618540
RCS: knots at 1, 2, 5, 8 years	403031.3	403160.5	618406	618541
RCS: knots at 1, 2, 5, 9 years	403023.4	403172.5		
RCS: knots at 1, 2, 5, 7, 9 years			618396	618552
FP[4] powers: 1, 1	403016.7	403126.1		
FP[4] powers: 0.5, 1			618433	618548
White blood cell count (10^{12}/L)				
LS: knots at 2, 4 years	4825395	4825548	8054469	8054628
LS: knots at 1, 2, 4 years	4825379	4825556	8054394	8054577
RCS: knots at 1, 2, 4 years	4825397	4825527	8054500	8054635
RCS: knots at 1, 2, 3 years	4825384	4825513	8054475	8054610
RCS: knots at 1, 2, 5 years	4825408	4825538	8054522	8054656
RCS: knots at 0.1, 2, 5 years	4825399	4825528	8054505	8054640
RCS: knots at 1, 2, 4, 5 years	4825395	4825548	8054461	8054620
RCS: knots at 1, 2, 5, 7 years	4825404	4825558	8054474	8054633
RCS: knots at 1, 2, 5, 8 years	4825404	4825558	8054475	8054634
RCS: knots at 1, 2, 5, 9 years	4825403	4825579		
RCS: knots at 1, 2, 5, 7, 9 years			8054472	8054656
FP[4] powers: -0.5, 0.5	4825300	4825406		
FP[4] powers: 0, 0.5			8054363	8054474
Basophil count (10^{12}/L)				
Linear	-2735192	-2735087	-4363351	-4363242
LS: knots at 5 years	-2735166	-2735037	-4363323	-4363190

LS: knots at 2, 4 years	-2735145	-2734993	-4363294	-4363137
LS: knots at 1, 2, 4 years	-2735117	-2734942	-4363269	-4363088
Eosinophil count (10^{12}/L)				
Linear	-783055.2	-782950.2	-1538590	-1538481
LS: knots at 5 years	-783027.4	-782899.0	-1538563	-1538430
LS: knots at 2, 4 years	-783011.1	-782859.3	-1538535	-1538377
LS: knots at 1, 2, 4 years	-782990.4	-782815.2	-1538522	-1538340
Lymphocyte count (10^{12}/L)				
LS: knots at 2 years	2866663	2866792	4272393	4272527
LS: knots at 1, 2 years			4272404	4272562
LS: knots at 1, 4 years	2866669	2866821		
LS: knots at 1, 2, 4 years	2866687	2866863	4272416	4272598
RCS: knots at 1, 2, 4 years	2866662	2866791	4272392	4272525
RCS: knots at 1, 2, 3 years	2866662	2866791	4272393	4272527
RCS: knots at 1, 2, 5 years	2866664	2866792	4272390	4272524
RCS: knots at 0.1, 2, 5 years	2866662	2866791	4272389	4272523
RCS: knots at 1, 2, 4, 5 years	2866676	2866828	4272408	4272566
RCS: knots at 1, 2, 5, 7 years	2866676	2866829	4272405	4272563
RCS: knots at 1, 2, 5, 8 years	2866675	2866828	4272403	4272561
RCS: knots at 1, 2, 5, 9 years	2866681	2866857		
RCS: knots at 1, 2, 5, 7, 9 years			4272412	4272594
FP^4 powers: 0, 0.5	2866659	2866764		
FP^4 powers: 0.5, 0.5			4272359	4272469
Monocyte count (10^{12}/L)				
LS: knots at 2, 4 years	-105283.4	-105131.2	-647562	-647404
LS: knots at 1, 2, 4 years	-105321.7	-105146.2	-647696	-647514
RCS: knots at 1, 2, 4 years	-105300.7	-105172.0	-647496	-647362

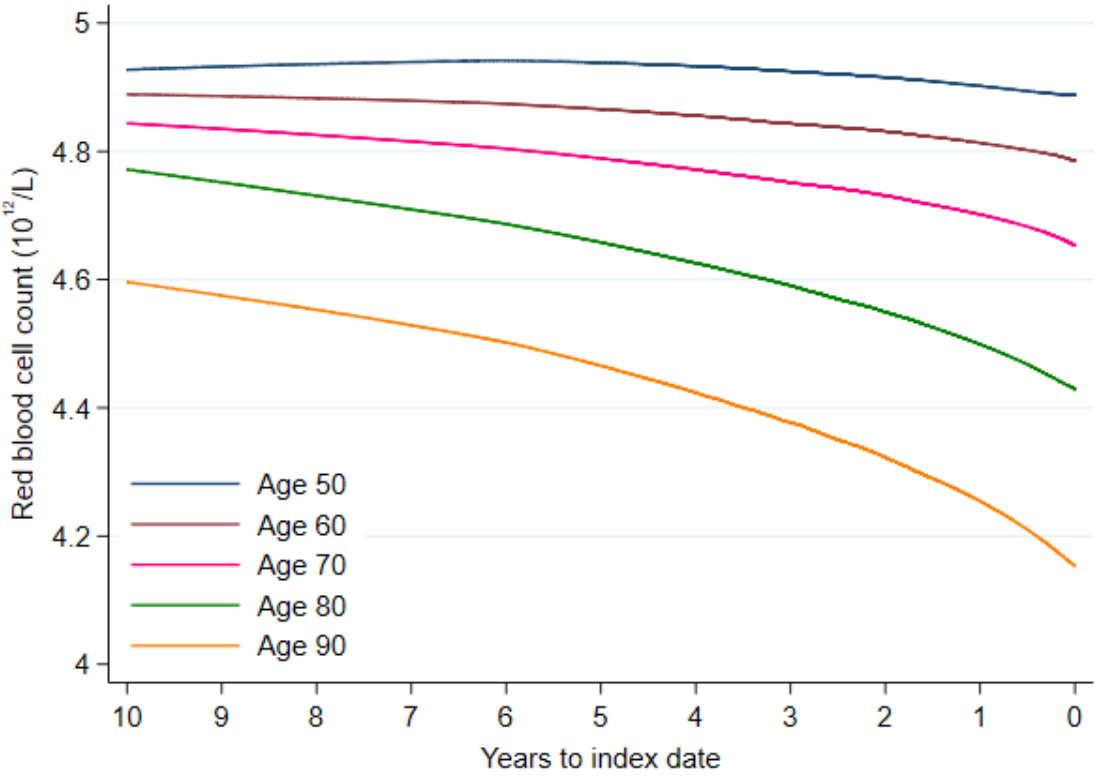
RCS: knots at 1, 2, 3 years	-105311.5	-105182.8	-647542	-647408
RCS: knots at 1, 2, 5 years	-105291.5	-105162.7	-647458	-647324
RCS: knots at 0.1, 2, 5 years	-105304.2	-105175.4	-647491	-647357
RCS: knots at 1, 2, 4, 5 years	-105284.7	-105132.6	-647572	-647414
RCS: knots at 1, 2, 5, 7 years	-105275.8	-105123.6	-647543	-647385
RCS: knots at 1, 2, 5, 8 years	-105278.0	-105125.9	-647543	-647385
RCS: knots at 1, 2, 5, 9 years	-105304.5	-105128.9		
RCS: knots at 1, 2, 5, 7, 9 years			-647581	-647399
FP[4] powers: 0.5, 1	-105392.6	-105263.9		
FP[4] powers: 1, 3			-647292	-647158
Neutrophil count (10^{12}/L)				
LS: knots at 2, 4 years	4307639	4307827	6279759	6279954
LS: knots at 1, 2, 4 years	4307597	4307819	6279579	6279810
RCS: knots at 1, 2, 4 years	4307634	4307787	6279864	6280022
RCS: knots at 1, 2, 3 years	4307614	4307766	6279800	6279958
RCS: knots at 1, 2, 5 years	4307653	4307805	6279920	6280078
RCS: knots at 0.1, 2, 5 years	4307634	4307786	6279874	6280032
RCS: knots at 1, 2, 4, 5 years	4307637	4307824	6279744	6279938
RCS: knots at 1, 2, 5, 7 years	4307650	4307838	6279802	6279996
RCS: knots at 1, 2, 5, 8 years	4307650	4307837	6279812	6280006
RCS: knots at 1, 2, 5, 9 years	4307656	4307879		
RCS: knots at 1, 2, 5, 7, 9 years			6279783	6280014
FP[4] powers: 0, 0.5	4307425	4307577		
FP[4] powers: 0.5, 0.5			6280023	6280145

[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.

[2]Mixed-effects model included time to index (years, in the specified form), age at index (years, modelled as a linear term), bowel cancer status (yes/no), and an interaction between time and bowel cancer status as fixed effects. A random intercept for patient and slope for time were used. For red blood cell count, haemoglobin, haematocrit, and neutrophil count, a second interaction, between time and age, was added as a fixed effect upon visual inspection of LOWESS trends.

[3]Choice and number of knots were based on visual inspection of LOWESS trends. Linear terms were only assessed where a linear relationship seemed plausible upon visual inspection of LOWESS trends.
[4]Fractional polynomial powers were determined using the *fp* command in Stata.
Abbreviations: LS=linear splines, RCS=restricted cubic splines, FP=fractional polynomials.

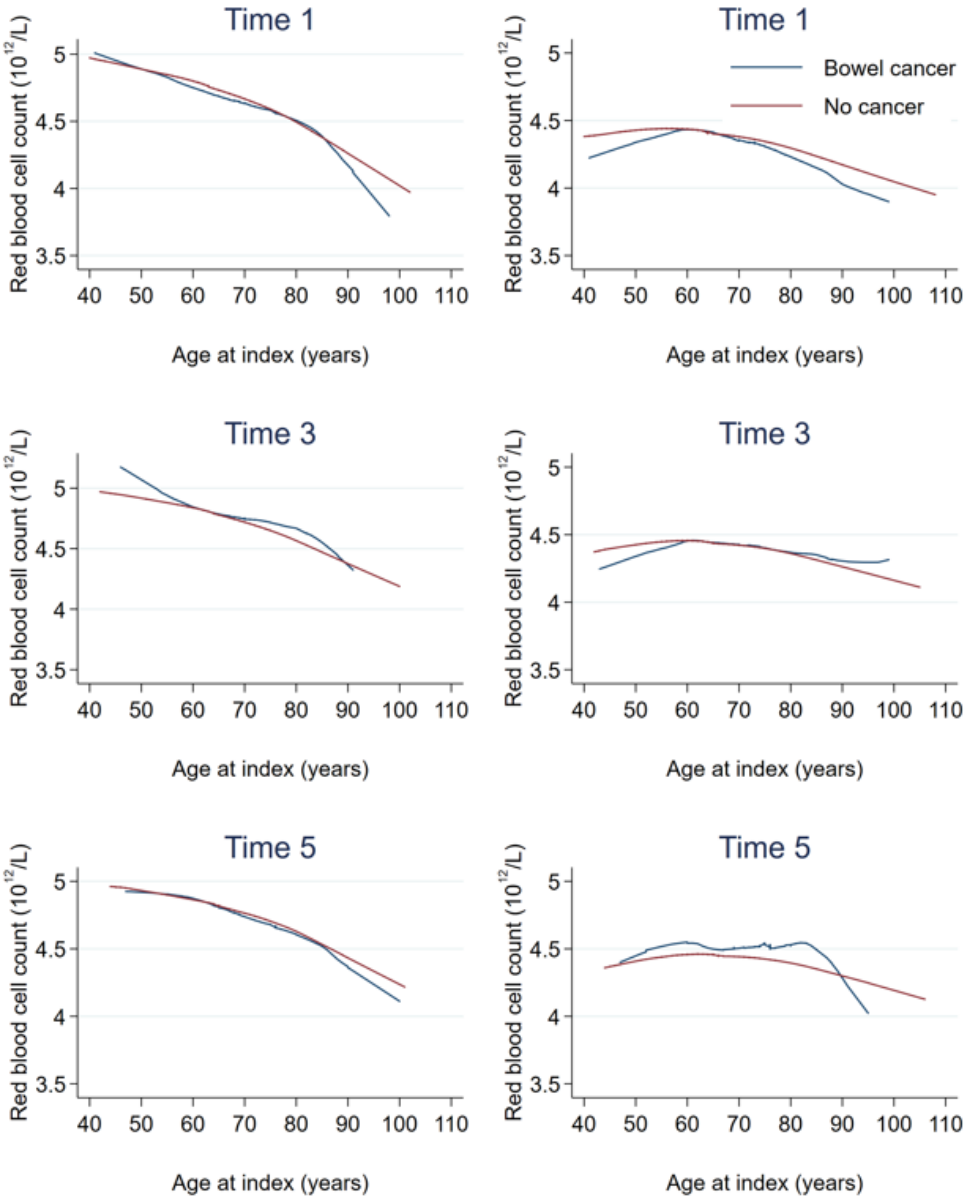
Figure C30: Trajectories (LOWESS) in red blood cell count for male controls by age at index date[1, 2]



[1]Index date was a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years).

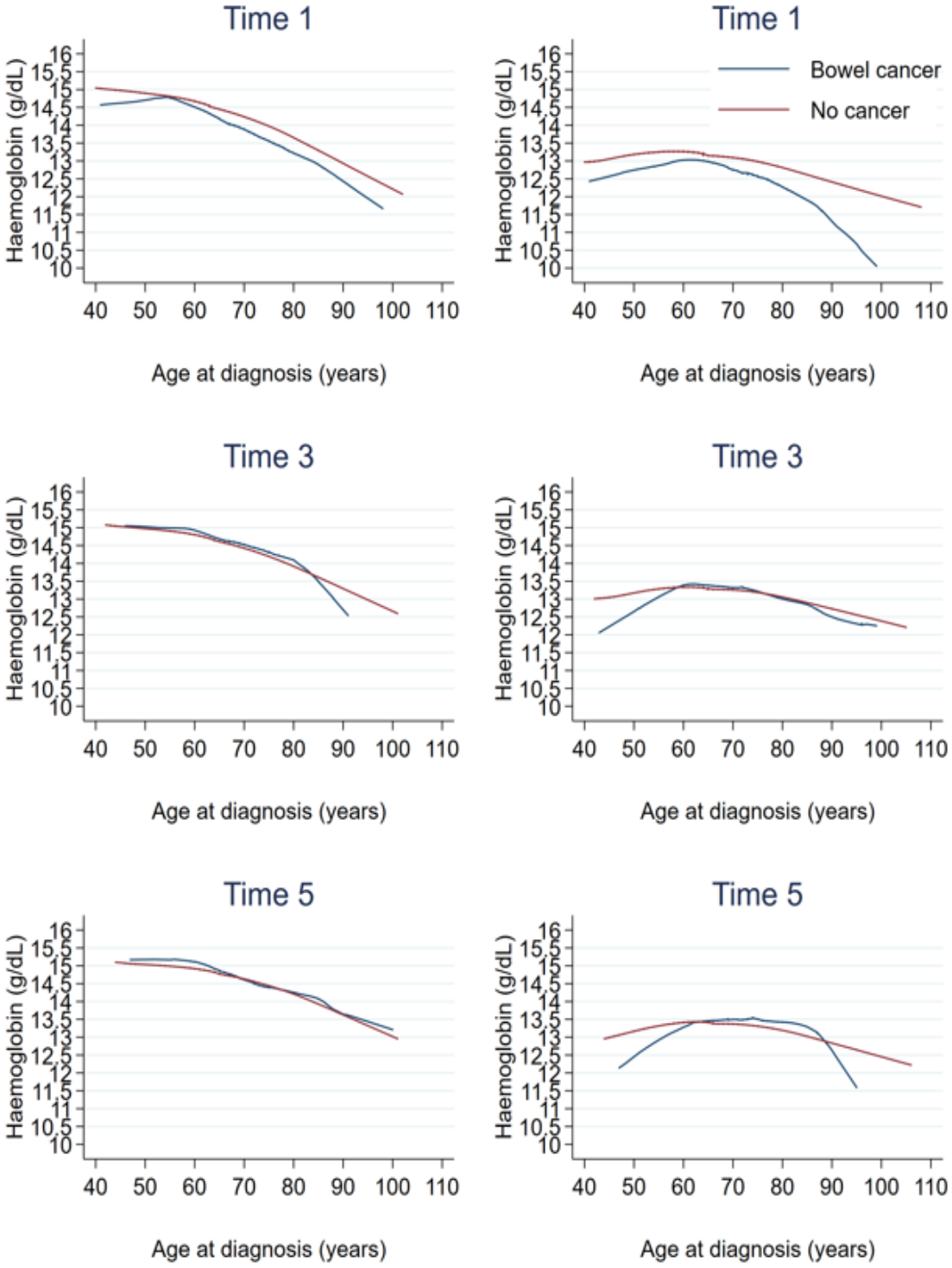
C5. Assessing non-linearity of age to index (years)

Figure C31: Red blood cell count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



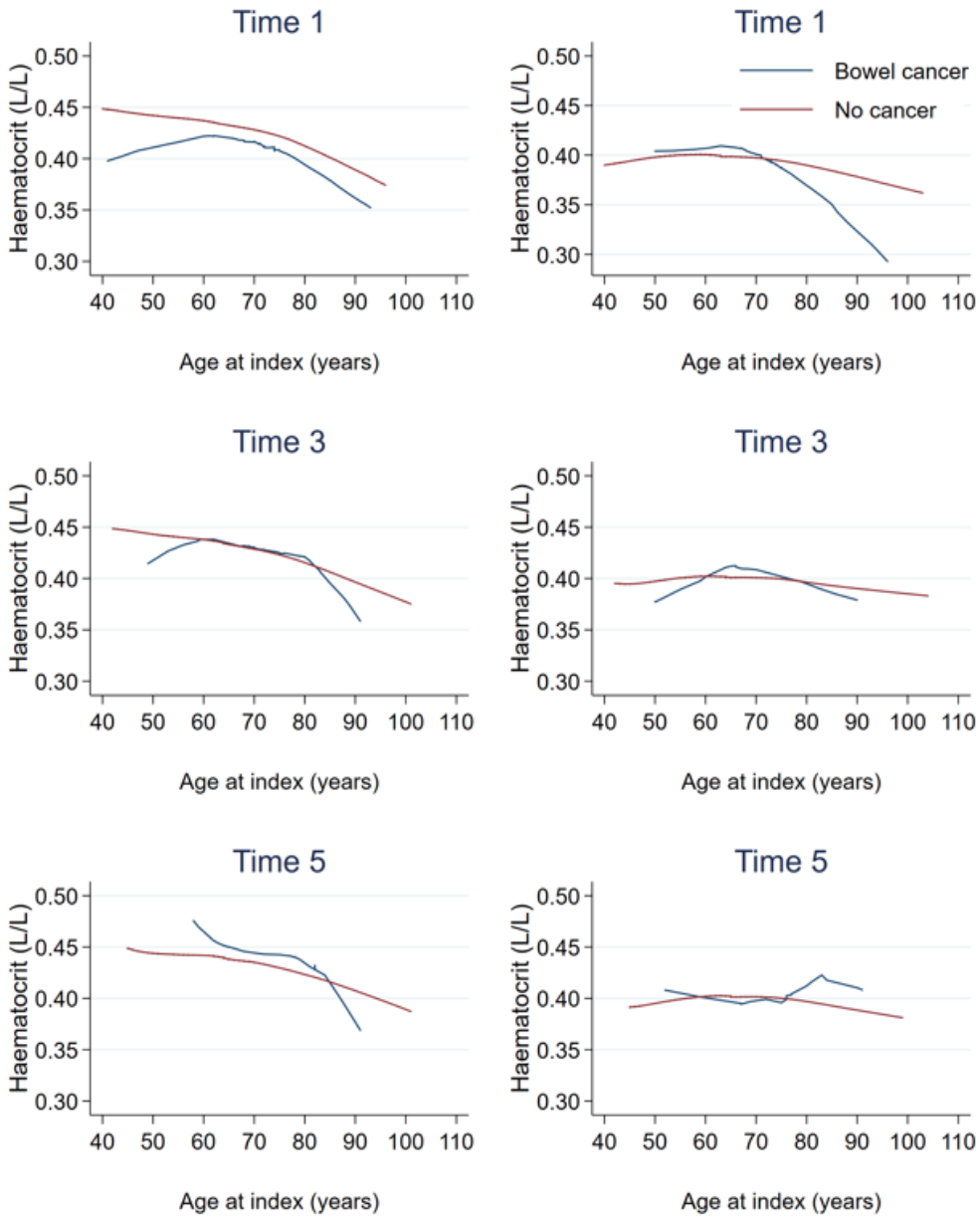
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C32: Haemoglobin by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



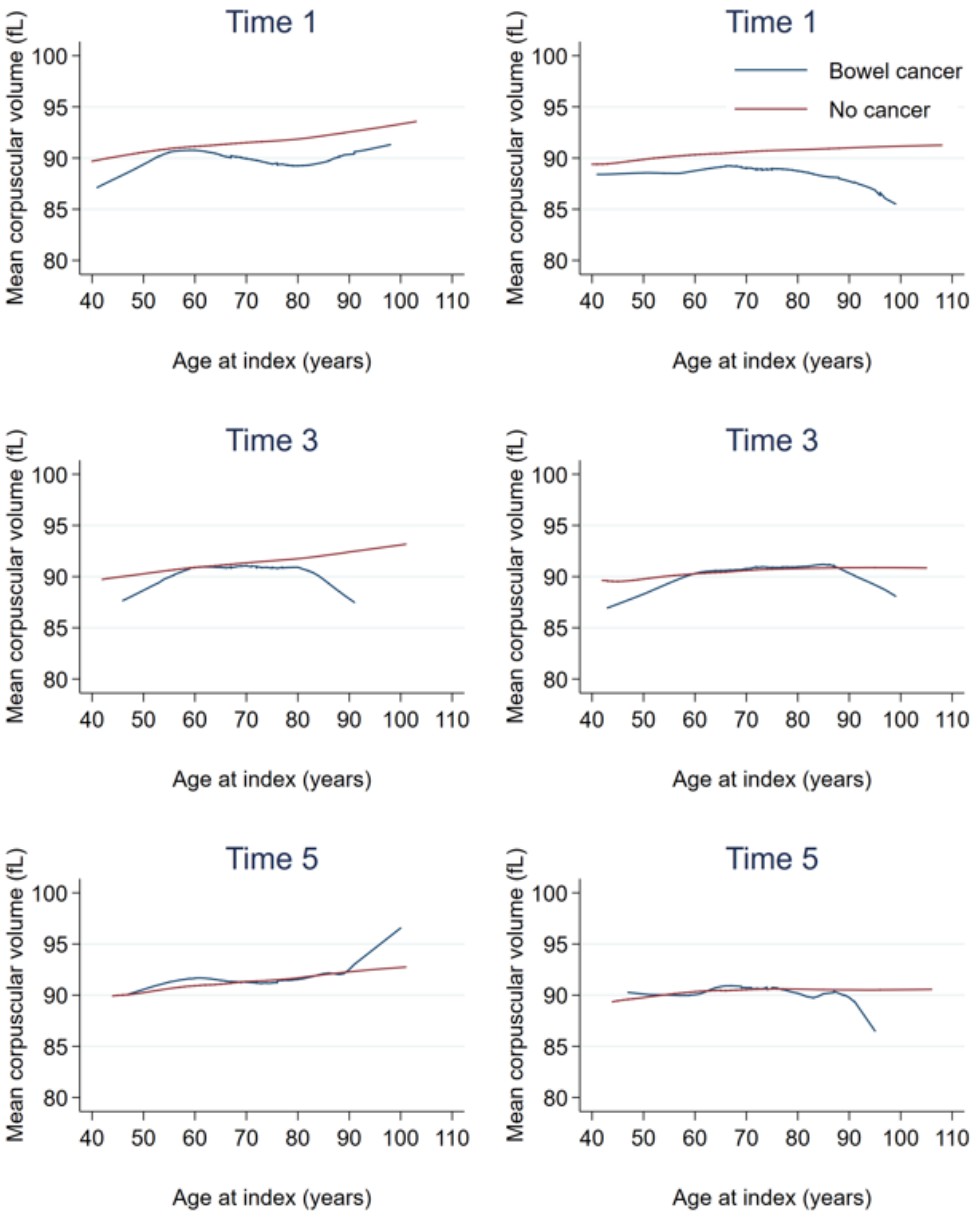
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C33: Haematocrit by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



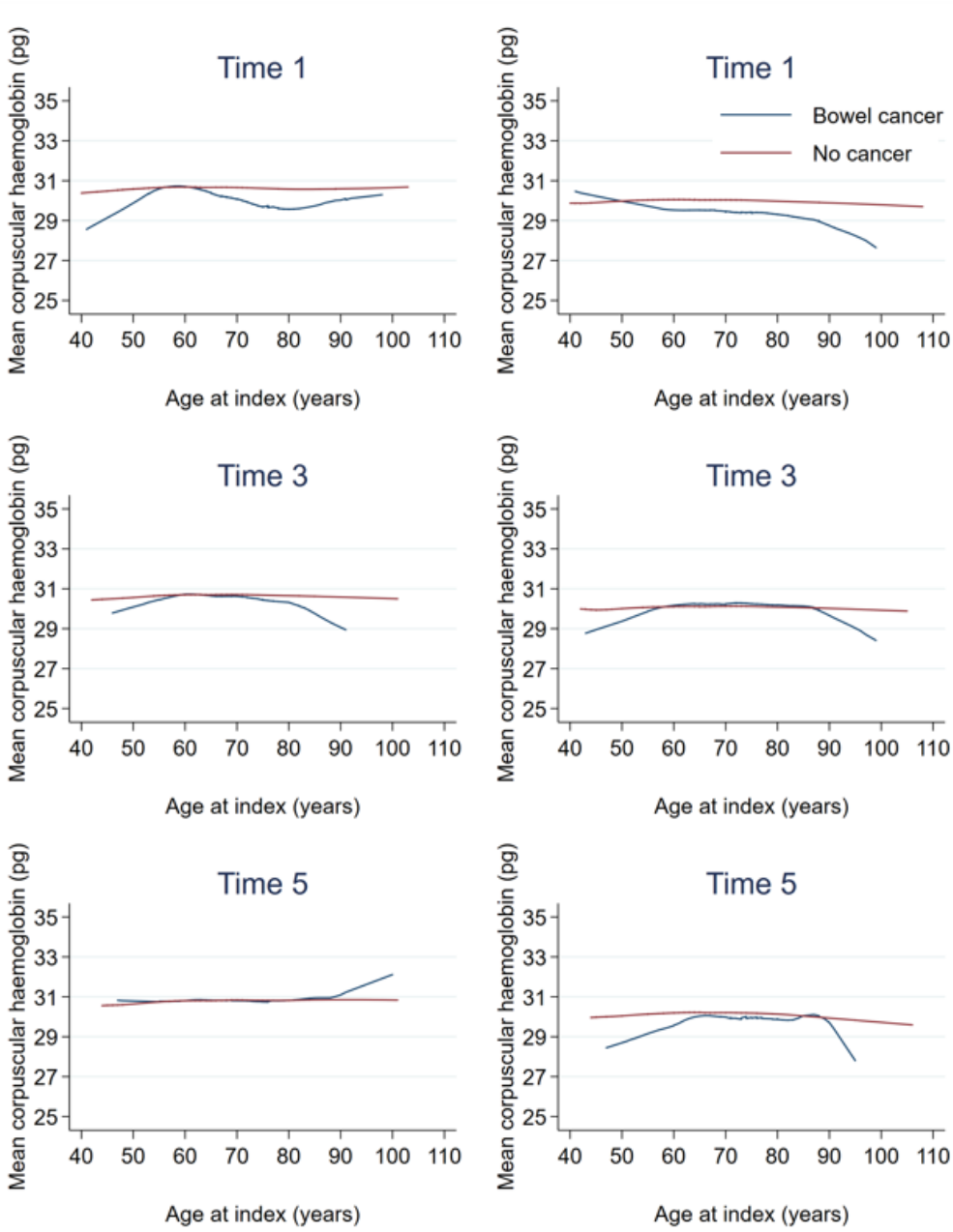
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C34: Mean corpuscular volume by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



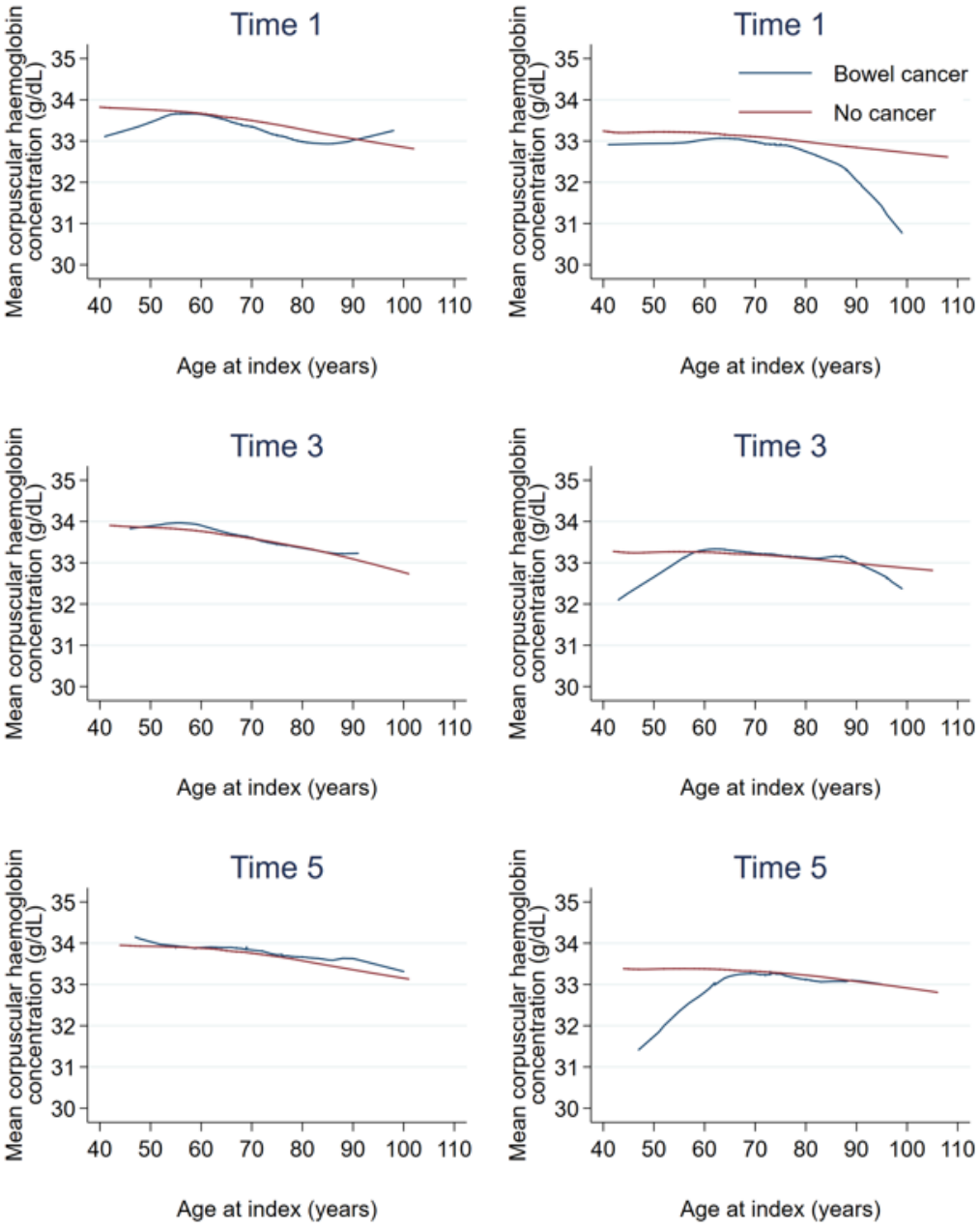
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C35: Mean corpuscular haemoglobin by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



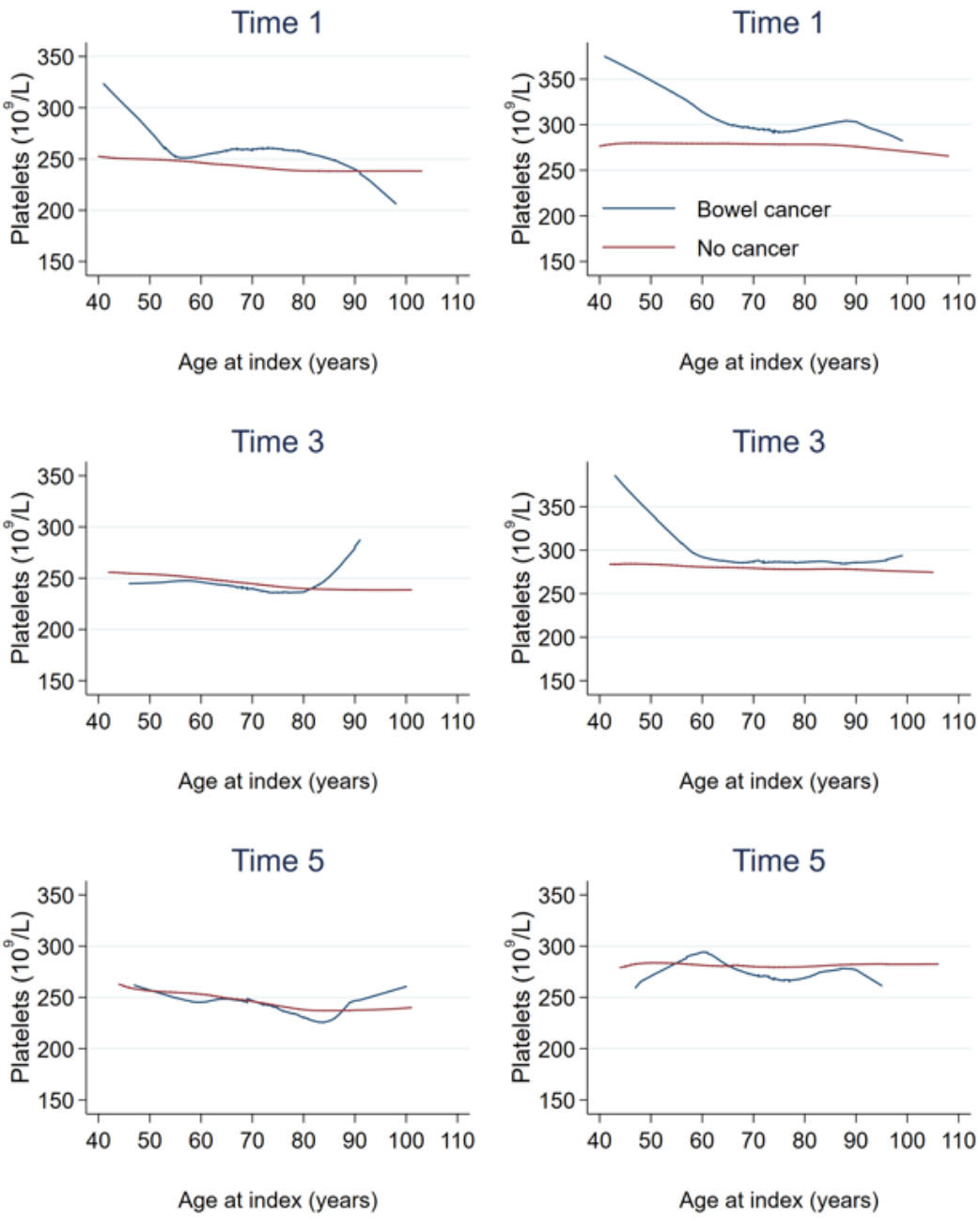
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C36: Mean corpuscular haemoglobin concentration by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



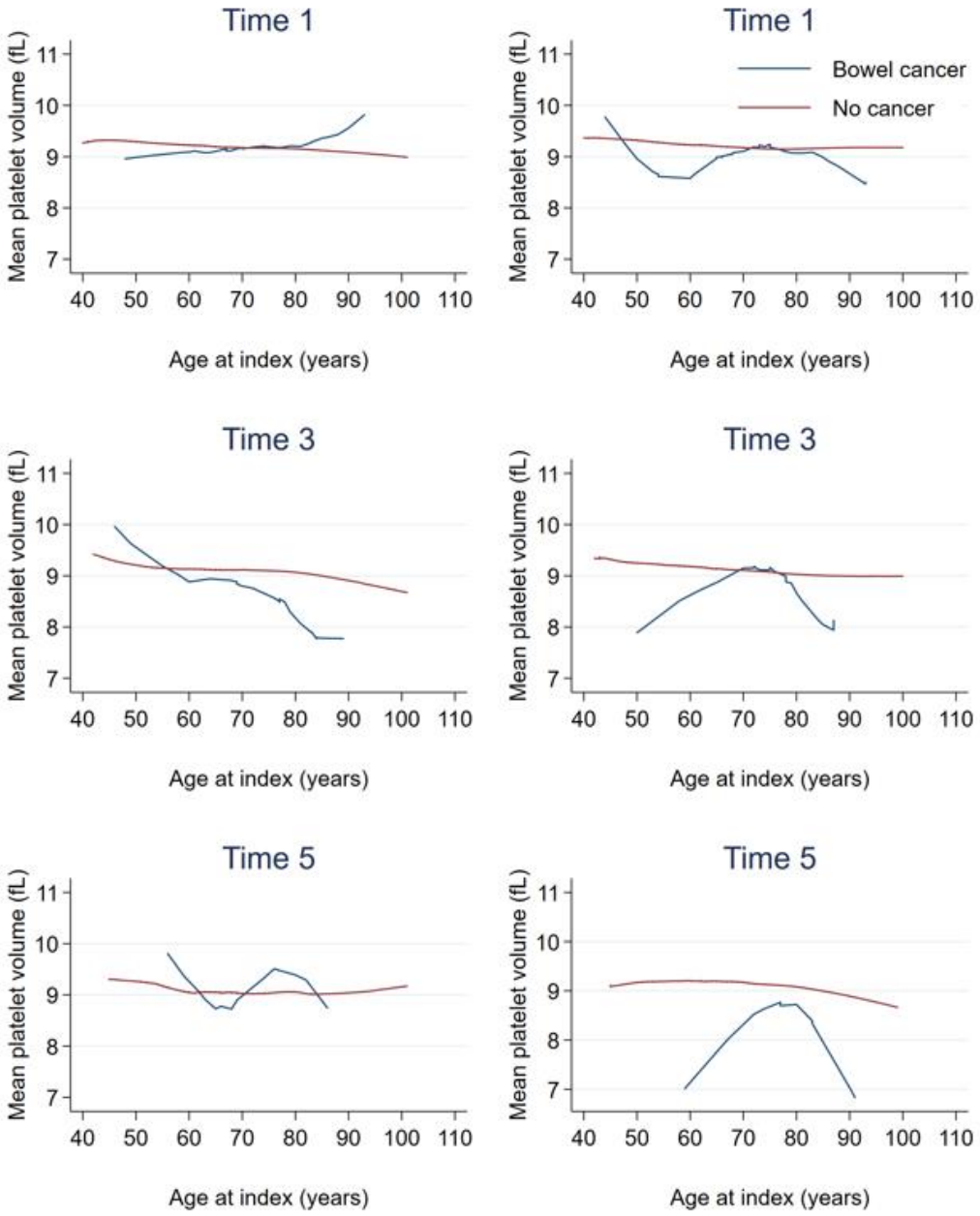
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C37: Platelet count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



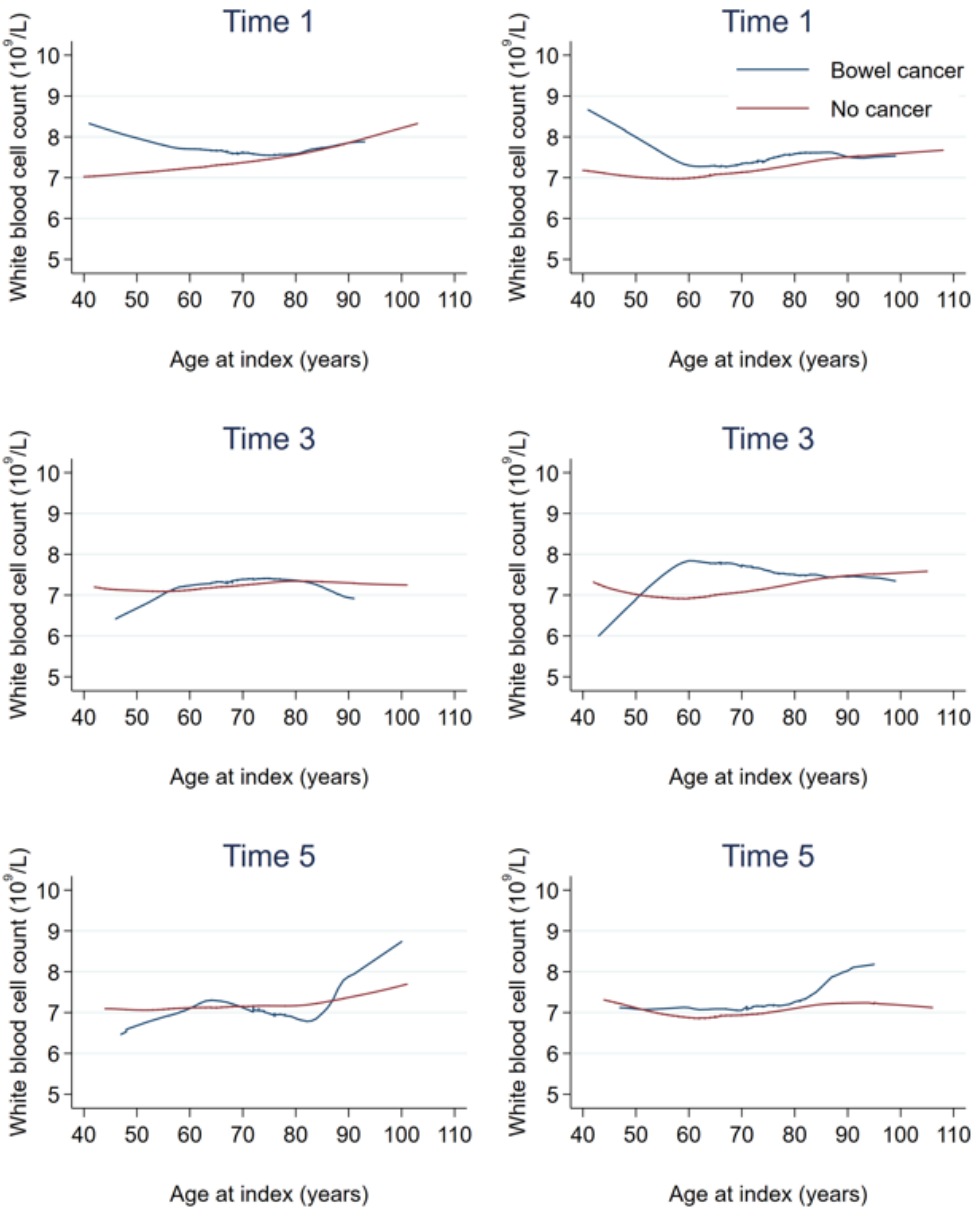
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C38: Mean platelet count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



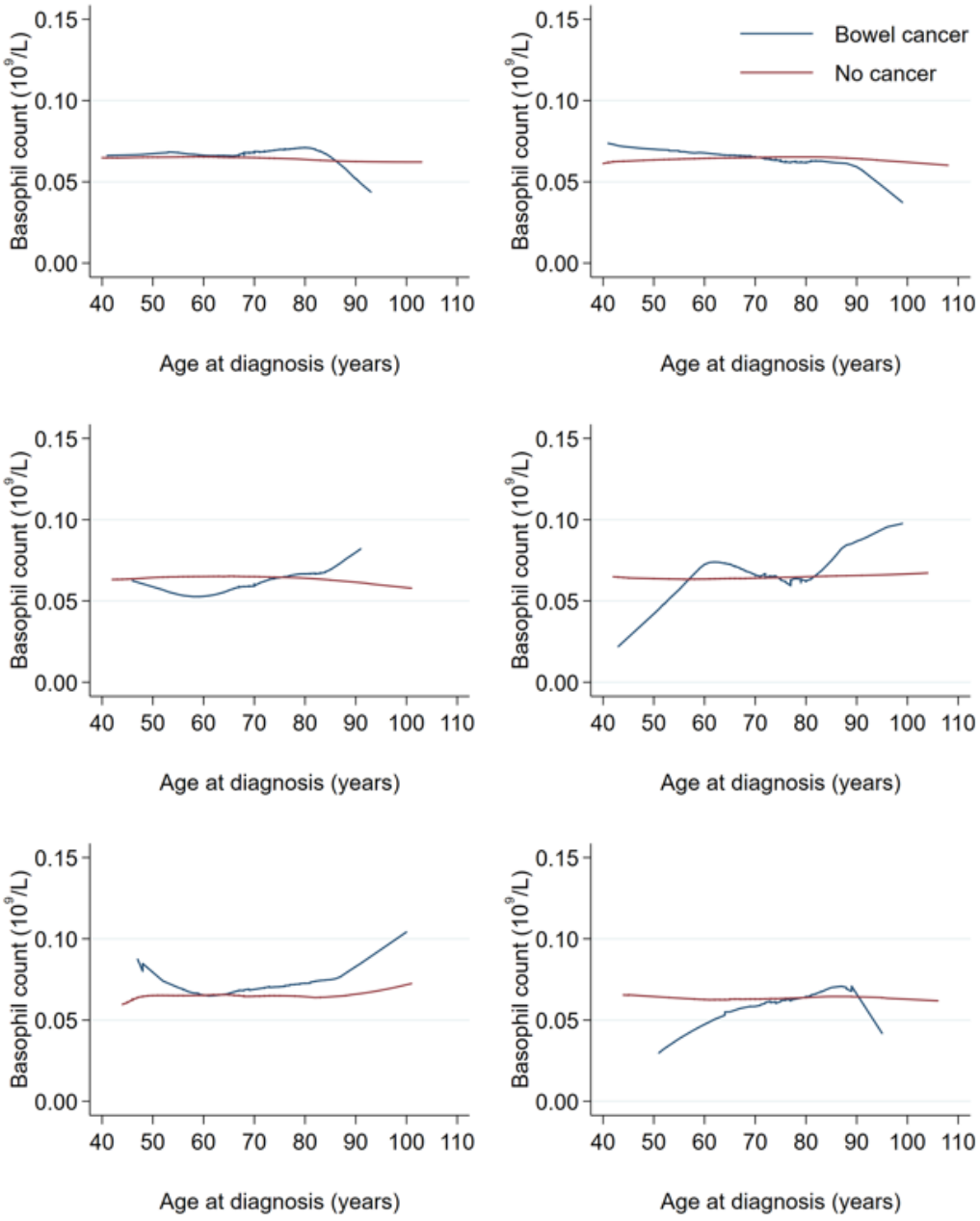
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C39: White blood cell count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



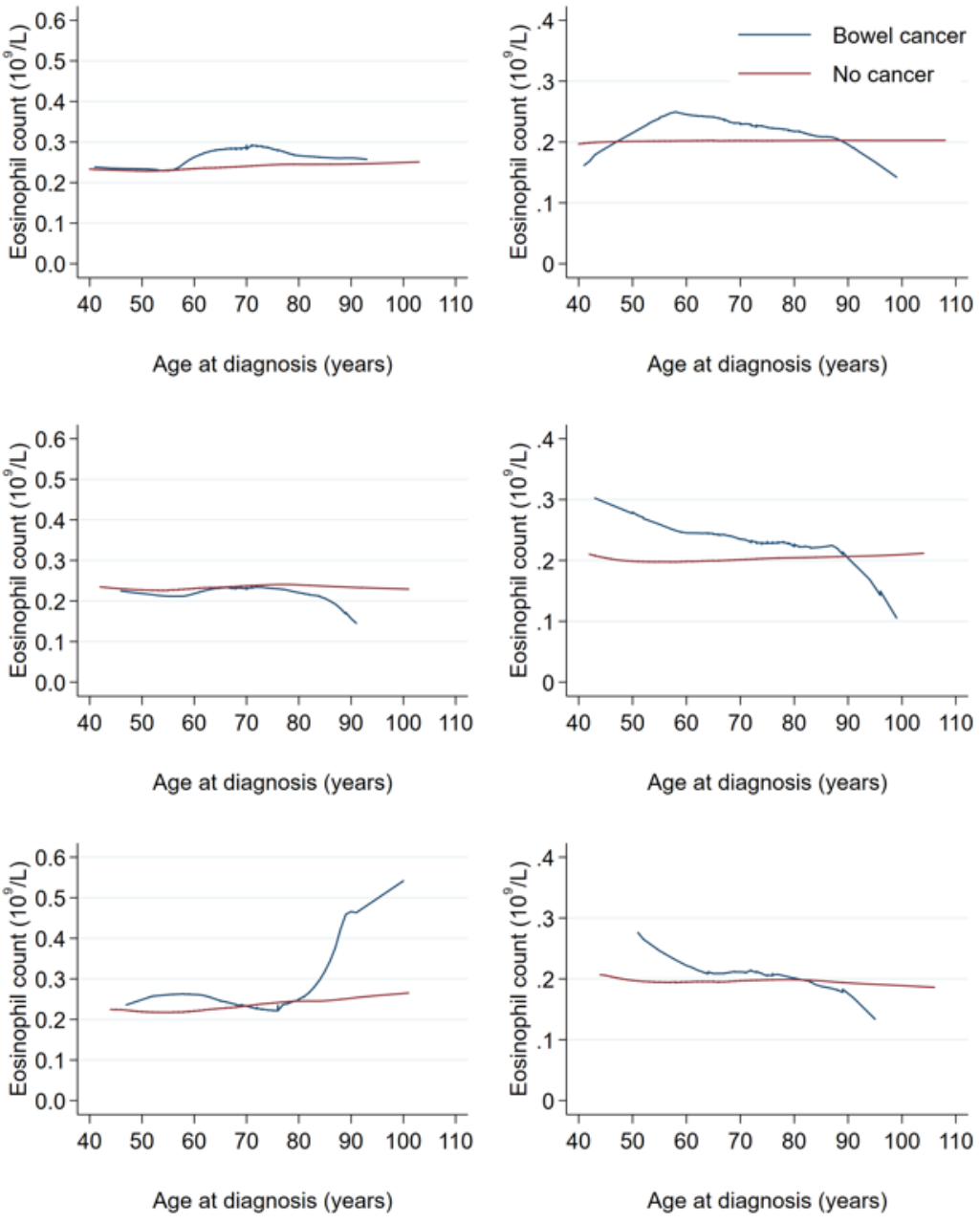
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C40: Basophil count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



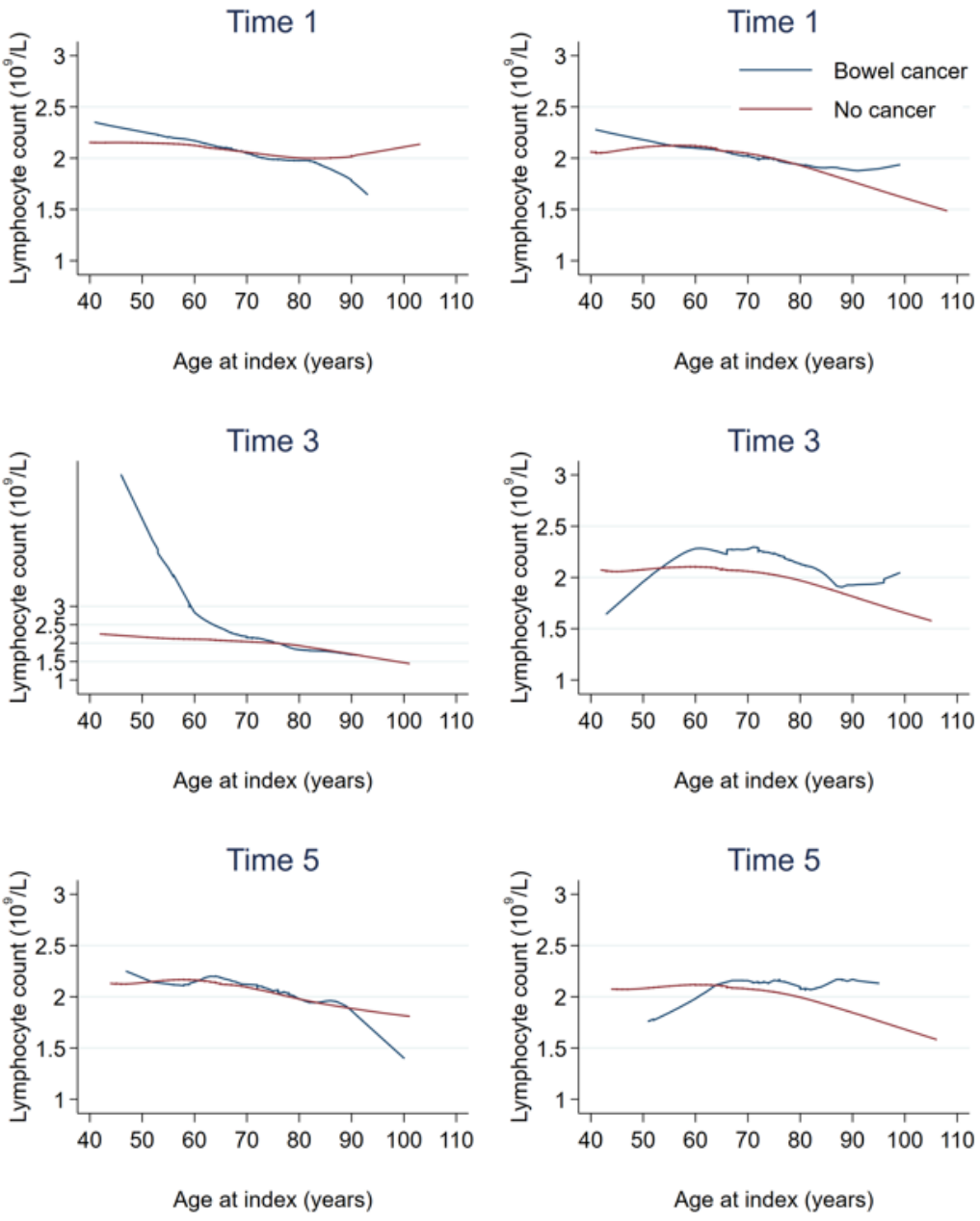
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C41: Eosinophil count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



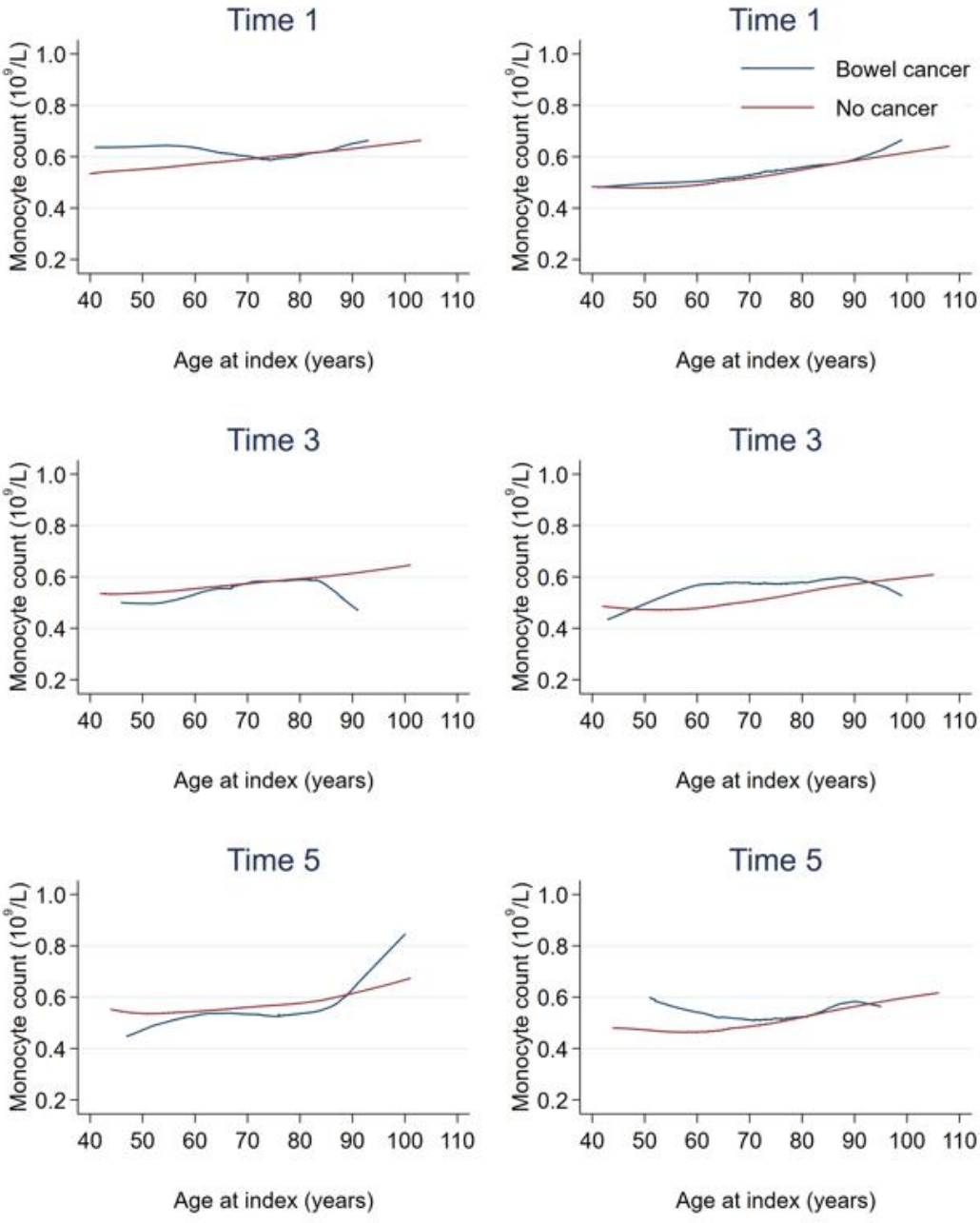
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C42: Lymphocyte count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



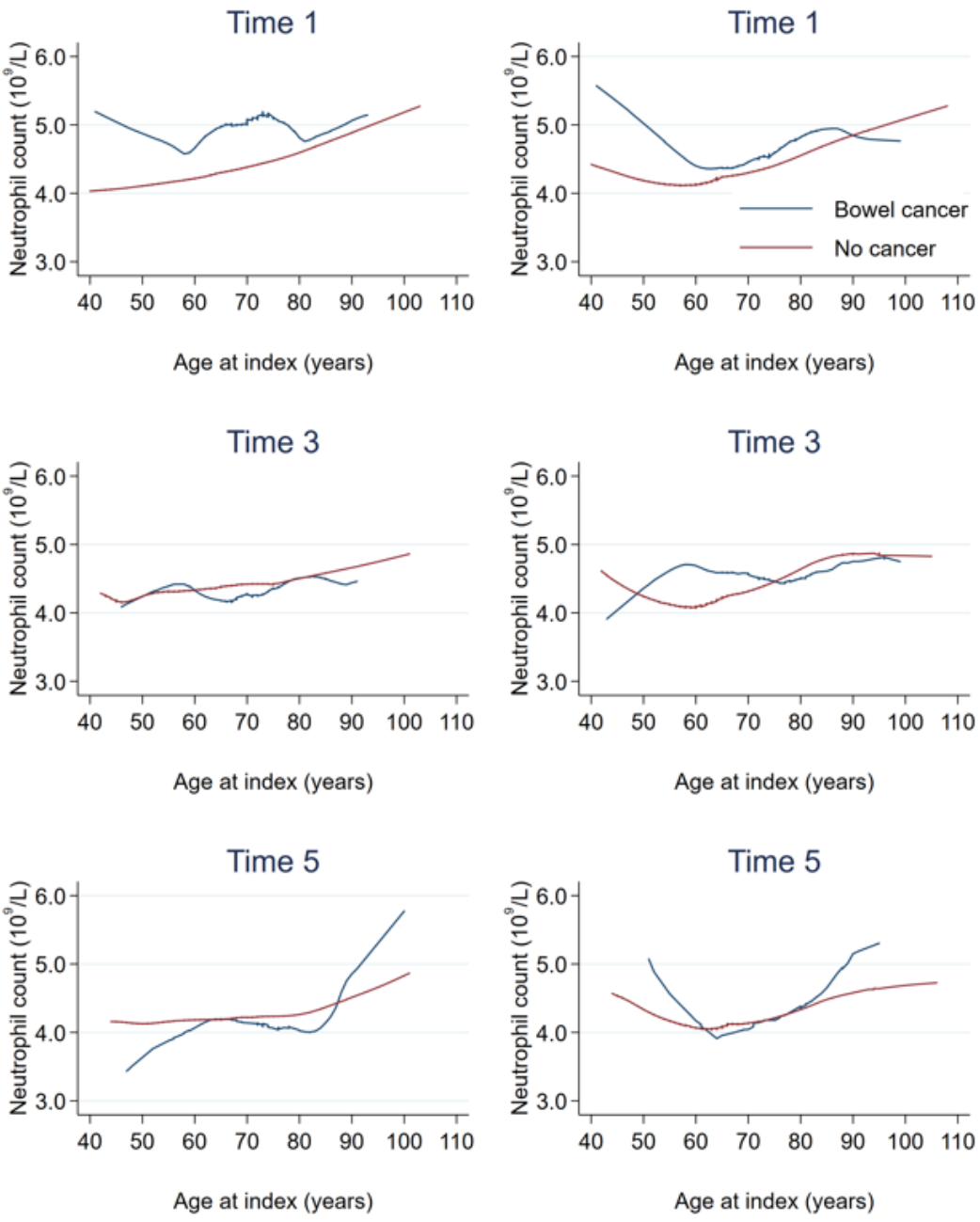
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C43: Monocyte count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Figure C44: Neutrophil count by age at index date (LOWESS) for males (left) and females (right) with and without bowel cancer[1, 2]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends for age are at 1, 3, and 5 years before index date.

Table C4: Akaike and Bayesian information criteria to compare methods for non-linearity of age at index date[1, 2, 3]

FBC parameter	Males		Females	
	AIC	BIC	AIC	BIC
Red blood cell count (10^9/L)				
LS: knots at 65 years	632777	633086	632431	632752
LS: knots at 70 years	632507	632816	633007	633328
LS: knots at 60, 80 years	632745	633126	632402	632797
LS: knots at 50, 70 years	632564	632945	632502	632897
LS: knots at 60, 70, 80 years	632536	632988	632155	632625
RCS: knots at 60, 70, 80 years	632448	632757	632697	633018
RCS: knots at 50, 70, 90 years	632481	632790	632414	632735
FP^4 powers: 3, 3	632820	633106		
FP^4 powers: 2, 2			632743	633064
Haemoglobin (g/dL)				
LS: knots at 65 years	3473369	3473681	5257483	5257806
LS: knots at 70 years	3473617	3473928	5259398	5259721
LS: knots at 60, 80 years	3473422	3473805	5257219	5257616
LS: knots at 50, 70 years	3473371	3473754	5258107	5258504
LS: knots at 60, 70, 80 years	3473232	3473687	5256806	5257278
RCS: knots at 60, 70, 80 years	3473365	3473677	5258750	5259073
RCS: knots at 50, 70, 90 years	3473169	3473480	5258084	5258407
FP^4 powers: 2, 3	3473439	3473750		
FP^4 powers: 1, 2			5258125	5258448
Haematocrit (L/L)				
LS: knots at 65 years	-934266	-933998	-1537267	-1536986

LS: knots at 70 years	-934293	-934024	-1536940	-1536659
LS: knots at 60, 80 years	-934222	-933892	-1537186	-1536840
LS: knots at 50, 70 years	-934204	-933874	-1537055	-1536710
LS: knots at 60, 70, 80 years	-934158	-933766	-1537185	-1536775
RCS: knots at 60, 70, 80 years	-934323	-934054	-1537066	-1536785
RCS: knots at 50, 70, 90 years	-934349	-934081	-1537185	-1536905
FP^4 powers: 3, 3	-934134	-933907		
FP^4 powers: 2, 2			-1536983	-1536702
Mean corpuscular volume (fL)				
LS: knots at 55 years	6184600	6184815	10059014	10059237
LS: knots at 65 years	6184684	6184899	10058987	10059210
LS: knots at 70 years	6184792	6185007	10059167	10059390
LS: knots at 60, 80 years	6184569	6184808	10058906	10059154
LS: knots at 50, 70 years	6184684	6184922	10059127	10059375
LS: knots at 60, 70, 80 years	6184584	6184847	10058922	10059195
RCS: knots at 60, 70, 80 years	6184776	6184991	10059138	10059361
RCS: knots at 50, 60, 80 years	6184660	6184875	10058971	10059195
FP^4 powers: -2, -2	6184554	6184769		
FP^4 powers: -0.5, 0			10059004	10059227
Mean corpuscular haemoglobin (pg)				
LS: knots at 55 years	3744838	3745052	6150970	6151193
LS: knots at 65 years	3744921	3745135	6150835	6151057
LS: knots at 70 years	3745164	3745378	6151102	6151325
LS: knots at 60, 80 years	3744780	3745019	6150752	6150999
LS: knots at 50, 70 years	3744961	3745200	6151046	6151293
LS: knots at 60, 70, 80 years	3744799	3745062	6150769	6151041
RCS: knots at 60, 70, 80 years	3745113	3745327	6151052	6151275

RCS: knots at 50, 60, 90 years	3745032	3745246	6150991	6151214
FP[4] powers: -0.5, -0.5	3744808	3745023		
FP[4] powers: 0.5, 0.5			6150923	6151146
Mean corpuscular haemoglobin concentration (g/dL)				
LS: knots at 55 years	2809342	2809555	4531921	4532142
LS: knots at 65 years	2809254	2809467	4531338	4531560
LS: knots at 70 years	2809419	2809632	4531533	4531754
LS: knots at 60, 80 years	2809221	2809458	4531377	4531623
LS: knots at 50, 70 years	2809307	2809544	4531547	4531793
LS: knots at 60, 70, 80 years	2809233	2809494	4531351	4531621
RCS: knots at 60, 70, 80 years	2809371	2809584	4531482	4531704
RCS: knots at 50, 60, 80 years	2809206	2809419	4531350	4531571
FP[4] powers: 0.5, 0.5	2809226	2809439		
FP[4] powers: 0.5, 2			4531565	4531787
Platelets (10^{12}/L)				
Linear	12178142	12178333	19240535	19240733
LS: knots at 65 years	12178124	12178338	19240536	19240759
LS: knots at 70 years	12178144	12178359	19240531	19240754
LS: knots at 60, 80 years	12178052	12178291	19240547	19240794
LS: knots at 50, 70 years	12178084	12178322	19240328	19240575
LS: knots at 60, 70, 80 years	12178026	12178289	19240436	19240708
RCS: knots at 60, 70, 80 years	12178144	12178359	19240536	19240759
RCS: knots at 50, 70, 90 years	12178141	12178356	19240542	19240765
FP[4] powers: -2, -2	12177984	12178198	19240381	19240603
Mean platelet volume (fL)				
Linear	Relationship seems linear - non-linearity not assessed			
White blood cell count (10^{12}/L)				

LS: knots at 80 years	4825500	4825712	8054303	8054523
LS: knots at 85 years	4825498	4825711	8054438	8054658
LS: knots at 70, 85 years	4825494	4825730	8053396	8053640
LS: knots at 60, 70, 85 years	4825498	4825758	8053026	8053295
RCS: knots at 60, 70, 85 years	4825488	4825700	8053682	8053902
FP^4 powers: -2, -2	4825399	4825611		
FP^4 powers: -1, -0.5			8053253	8053473
Basophil count (10^{12}/L)				
Linear	Relationship seems linear - non-linearity not assessed			
Eosinophil count (10^{12}/L)				
LS: knots at 70 years	-783101	-782961	-1538528	-1538383
LS: knots at 80 years	-783072	-782932	-1538526	-1538380
LS: knots at 70, 80 years	-783085	-782921	-1538497	-1538328
LS: knots at 70, 90, 90 years	-783063	-782876	-1538496	-1538302
RCS: knots at 70, 90, 90 years	-783077	-782937	-1538526	-1538381
RCS: knots at 60, 75, 85 years	-783098	-782958	-1538526	-1538381
FP^4 powers: -2, -2	-783260	-783119	-1538622	-1538477
Lymphocyte count (10^{12}/L)				
LS: knots at 80 years	2867010	2867221	4271532	4271751
LS: knots at 85 years	2867059	2867270	4272038	4272257
LS: knots at 90 years	2867071	2867282	4272404	4272622
LS: knots at 80, 90 years	2867017	2867252	4271393	4271637
RCS: knots at 60, 70, 80 years	2866902	2867113	4270841	4271060
RCS: knots at 50, 70, 90 years	2866902	2867113	4270856	4271075
FP^4 powers: 0, 0.5	2866887	2867098		
FP^4 powers: 1, 2			4271002	4271221
Monocyte count (10^{12}/L)				

LS: knots at 60, 70 years	-105387	-105153	-650259	-650016
LS: knots at 60, 70, 80 years	-105369	-105111	-650372	-650104
LS: knots at 60, 70, 90 years	-105371	-105113	-650343	-650075
RCS: knots at 60, 70, 85 years	-105399	-105188	-649417	-649198
FP[4] powers: -1, -0.5	-105473	-105262		
FP[4] powers: -0.5, -0.5			-649874	-649655
Neutrophil count (10^{12}/L)				
LS: knots at 60 years	4307554	4307765	6274477	6274697
LS: knots at 60, 80 years	4307573	4307808	6274430	6274673
RCS: knots at 60, 70, 85 years	4307544	4307755	6276005	6276224
RCS: knots at 50, 75, 90 years	4307551	4307762	6275894	6276113
FP[4] powers: -0.5, 0.5	4307529	4307740		
FP[4] powers: 0, 0			6275159	6275378

[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.

[2]Mixed-effects models included time to index (years, linear for basophil count and eosinophil count and piecewise linear splines with knots at 1, 2, and 4 years before index for all others), age at index (years, in the specified form), bowel cancer status (yes/no), and an interaction between time and bowel cancer status as fixed effects. A random intercept for patient and slope for time were used. For red blood cell count, haemoglobin, haematocrit, and neutrophil count, a second interaction, between time and age, was added as a fixed effect upon visual inspection of LOWESS trends. For all models except for mean platelet volume, basophil count, lymphocyte count, and monocyte count, a third interaction between age and bowel cancer presence was added as a fixed effect upon visual inspection of LOWESS trends.

[3]Choice and number of knots were based on visual inspection of LOWESS trends. Linear terms were only assessed where a linear relationship seemed plausible upon visual inspection of LOWESS trends.

[4]Fractional polynomial powers were determined using the *fp* command in Stata.

Abbreviations: LS=linear splines, RCS=restricted cubic splines, FP=fractional polynomials.

C6. Mixed-effects models

Table C5: Mixed-effect model coefficients (95% CI) for red blood cell count parameters

Variable	Males						Females					
	RBC	**Hb**	**Hc**	**MCV**	**MCH**	**MCHC**	**RBC**	**Hb**	**Hc**	**MCV**	**MCH**	**MCHC**
N	*376516*	*394601*	*93128*	*388464*	*380868*	*360071*	509513	534243	129276	526287	514876	488199
n cases	*9024*	*9161*	*2256*	*8792*	*8631*	*8363*	7976	8070	2052	7745	7567	7396
Fixed effects:												
Constant	5.2721 (5.2458, 5.2984)	15.6416 (15.5582, 15.7249)	0.4627 (0.4577, 0.4677)	86.8880 (86.7174, 87.0586)	29.6513 (29.5910, 29.7115)	34.0711 (34.0369, 34.1054)	4.1865 (4.1675, 4.2056)	12.1175 (12.0546, 12.1803)	0.3706 (0.3668, 0.3743)	87.1667 (87.0087, 87.3246)	29.1787 (29.1224, 29.2350)	33.0309 (33.0000, 33.0618)
Age at index date[1] (years)	-0.0077 (-0.0082, -0.0072)	-0.0142 (-0.0158, -0.0126)	-0.0004 (-0.0005, -0.0003)	0.0777 (0.0744, 0.0810)	0.0194 (0.0183, 0.0206)	-0.0067 (-0.0074, -0.0061)	0.0050 (0.0047, 0.0054)	0.0238 (0.0225, 0.0250)	0.0006 (0.0005, 0.0007)	0.0649 (0.0617, 0.0680)	0.0182 (0.0171, 0.0193)	0.0033 (0.0027, 0.0039)
Age at index date[1] – knot at 60 (years)	-0.0037 (-0.0051, -0.0024)	-0.0307 (-0.0350, -0.0264)	-0.0008 (-0.0010, -0.0005)	-0.0551 (-0.0642, -0.0461)	-0.0281 (-0.0313, -0.0250)	-0.0117 (-0.0134, -0.0100)	-0.0107 (-0.0117, -0.0096)	-0.0503 (-0.0538, -0.0467)	-0.0012 (-0.0014, -0.0009)	-0.0621 (-0.0710, -0.0532)	-0.0264 (-0.0295, -0.0232)	-0.0113 (-0.0130, -0.0096)
Age at index date[1] – knot at 70 (years)	-0.0127 (-0.0147,	-0.0326 (-0.0388,	-0.0007 (-0.0011,	0.0094	0.0010	-0.0043 (-0.0069,	-0.0079 (-0.0094,	-0.0188 (-0.0238,	-0.0007 (-0.0010,	-0.0023	-0.0047 (-0.0093,	-0.0088 (-0.0112,

	-0.0107)	-0.0263)	-0.0003)	(-0.0044, 0.0233)	(-0.0039, 0.0059)	-0.0016)	-0.0063)	-0.0138)	-0.0004)	(-0.0154, 0.0108)	-0.0000)	-0.0063)
Age at index date[1] – knot at 80 (years)	-0.0027 (-0.0050, -0.0003)	-0.0040 (-0.0115, 0.0034)	-0.0004 (-0.0009, 0.0000)	0.0289 (0.0112, 0.0465)	0.0080 (0.0019, 0.0143)	0.0021 (-0.0012, 0.0054)	-0.0009 (-0.0024, 0.0006)	0.0014 (-0.0036, 0.0064)	0.0001 (-0.0002, 0.0004)	0.0190 (0.0056, 0.0324)	0.0060 (0.0012, 0.0108)	0.0014 (-0.0011, 0.0039)
Time to index date[1] (years)	0.0148 (-0.0197, 0.0493)	-0.0125 (-0.1250, 0.1000)	-0.0014 (-0.0083, 0.0056)	-0.1748 (-0.2074, -0.1422)	-0.0137 (-0.0258, -0.0017)	0.0421 (0.0324, 0.0519)	-0.0161 (-0.0404, 0.0082)	-0.2697 (-0.3529, -0.1865)	-0.0111 (-0.0161, -0.0060)	-0.3001 (-0.3289, -0.2712)	-0.0532 (-0.0637, -0.0427)	0.0267 (0.0186, 0.0347)
Time to index date[1] date – knot at 1 (years)	-0.0107 (-0.0673, 0.0460)	-0.0014 (-0.1867, 0.1839)	0.0026 (-0.0089, 0.0142)	0.0625 (0.0108, 0.1141)	-0.0018 (-0.0209, 0.0173)	-0.0205 (-0.0361, -0.0049)	-0.0063 (-0.0461, 0.0335)	0.2064 (0.0697, 0.3430)	0.0101 (0.0017, 0.0185)	0.1804 (0.1348, 0.2261)	0.0352 (0.0186, 0.0518)	-0.0103 (-0.0231, 0.0025)
Time to index date[1] date – knot at 2 (years)	0.0054 (-0.0348, 0.0456)	-0.0063 (-0.1378, 0.1252)	-0.0047 (-0.0131, 0.0037)	-0.0148 (-0.0481, 0.0185)	0.0257 (0.0135, 0.0381)	0.0298 (0.0197, 0.0400)	-0.0215 (-0.0496, 0.0066)	-0.1740 (-0.2709, -0.0772)	-0.0072 (-0.0133, -0.0011)	0.0340 (0.0045, 0.0635)	0.0399 (0.0291, 0.0506)	0.0349 (0.0266, 0.0432)
Time to index date[1] date – knot at 4 (years)	-0.0395 (-0.0646, -0.0144)	-0.0437 (-0.1257, 0.0383)	0.0038 (-0.0017, 0.0094)	0.1234 (0.1068, 0.1400)	0.0037 (-0.0024, 0.0098)	-0.0324 (-0.0374, -0.0274)	0.0420 (0.0248, 0.0593)	0.2626 (0.2031, 0.3221)	0.0085 (0.0046, 0.0125)	0.0727 (0.0580, 0.0873)	-0.0073 (-0.0126, -0.0019)	-0.0373 (-0.0414, -0.0332)
Bowel cancer present	-0.2529 (-0.4629,	-2.5579 (-3.1632,	-0.0740 (-0.1091,	-8.0698 (-10.5710,	-2.1303 (-3.0312,	-1.7476 (-2.2370,	-0.2530 (-0.4467,	-2.0165 (-2.6006,	-0.0508 (-0.0843,	-6.9632 (-9.7568,	-2.3591 (-3.3783,	-1.1348 (-1.6733,

	-0.0428)	-1.9526)	-0.0389)	-5.5686)	-1.2293)	-1.2582)	-0.0593)	-1.4324)	-0.0173)	-4.1697)	-1.3399)	-0.5962)
Bowel cancer present by time to index date[1] interaction	0.1869 (0.1732, 0.2005)	1.4226 (1.3771, 1.4682)	0.0321 (0.0293, 0.0350)	4.5147 (4.3692, 4.6603)	1.4979 (1.4426, 1.5532)	0.7227 (0.6773, 0.7681)	0.1643 (0.1517, 0.1769)	1.3311 (1.2874, 1.3748)	0.0366 (0.0337, 0.0394)	4.8123 (4.6482, 4.9764)	1.6663 (1.6043, 1.7283)	0.6327 (0.5858, 0.6796)
Bowel cancer present by time to index date[1] (knot at 1) interaction	-0.1625 (-0.1868, -0.1381)	-1.2030 (-1.2841, -1.1220)	-0.0246 (-0.0298, -0.0195)	-3.9394 (-4.1983, -3.6805)	-1.2412 (-1.3386, -1.1438)	-0.6063 (-0.6876, -0.5250)	-0.1247 (-0.1468, -0.1025)	-1.0582 (-1.1352, -0.9813)	-0.0303 (-0.0354, -0.0253)	-4.1124 (-4.4004, -3.8244)	-1.3530 (-1.4605, -1.2454)	-0.4455 (-0.5282, -0.3628)
Bowel cancer present by time to index date[1] (knot at 2) interaction	-0.0208 (-0.0384, -0.0033)	-0.1037 (-0.1622, -0.0451)	-0.0050 (-0.0088, -0.0011)	-0.1012 (-0.2886, 0.0862)	-0.0784 (-0.1483, -0.0086)	-0.0287 (-0.0876, 0.0302)	-0.0336 (-0.0493, -0.0178)	-0.1575 (-0.2123, -0.1026)	-0.0047 (-0.0083, -0.0011)	-0.1472 (-0.3527, 0.0583)	-0.1389 (-0.2148, -0.0630)	-0.1358 (-0.1951, -0.0766)
Bowel cancer present by time to index date[1] (knot at 4) interaction	-0.0035 (-0.0120, 0.0051)	-0.1046 (-0.1332, -0.0761)	-0.0020 (-0.0041, 0.0000)	-0.4093 (-0.5016, -0.3170)	-0.1545 (-0.1888, -0.1201)	-0.0828 (-0.1114, -0.0542)	-0.0053 (-0.0130, 0.0023)	-0.0888 (-0.1152, -0.0624)	-0.0012 (-0.0031, 0.0007)	-0.4000 (-0.5002, -0.2999)	-0.1338 (-0.1708, -0.0969)	-0.0112 (-0.0400, 0.0176)
Time by age interaction	-0.0000 (-0.0007, 0.0006)	0.0007 (-0.0015, 0.0028)	0.0000 (-0.0001, 0.0002)				0.0003 (-0.0002, 0.0007)	0.0043 (0.0027, 0.0059)	0.0002 (0.0001, 0.0003)			
Time (knot at 1) by age interaction	0.0003 (-0.0008, 0.0013)	0.0003 (-0.0032, 0.0038)	-0.0001 (-0.0003, 0.0002)				0.0001 (-0.0007, 0.0009)	-0.0037 (-0.0064, -0.0010)	-0.0002 (-0.0003, -0.0000)			

Time (knot at 2) by age interaction	-0.0002 (-0.0009, 0.0006)	0.0000 (-0.0025, 0.0025)	0.0001 (-0.0001, 0.0003)				0.0004 (-0.0001, 0.0010)	0.0038 (0.0019, 0.0056)	0.0001 (0.0000, 0.0003)			
Time (knot at 4) by age interaction	0.0006 (0.0001, 0.0010)	0.0006 (-0.0009, 0.0021)	-0.0001 (-0.0002, 0.0000)				-0.0010 (-0.0013, -0.0006)	-0.0056 (-0.0067, -0.0044)	-0.0002 (-0.0002, -0.0001)			
Time by age (knot at 60) interaction	0.0016 (-0.0000, 0.0033)	0.0065 (0.0011, 0.0120)	0.0003 (-0.0001, 0.0006)				0.0010 (-0.0003, 0.0023)	0.0015 (-0.0029, 0.0060)	0.0001 (-0.0002, 0.0003)			
Time (knot at 1) by age (knot at 60) interaction	-0.0019 (-0.0046, 0.0007)	-0.0067 (-0.0155, 0.0021)	-0.0002 (-0.0007, 0.0004)				-0.0002 (-0.0023, 0.0018)	0.0039 (-0.0032, 0.0110)	0.0001 (-0.0004, 0.0005)			
Time (knot at 2) by age (knot at 60) interaction	0.0011 (-0.0006, 0.0029)	0.0032 (-0.0027, 0.0090)	-0.0001 (-0.0005, 0.0003)				-0.0005 (-0.0019, 0.0009)	-0.0061 (-0.0108, -0.0014)	-0.0002 (-0.0005, 0.0001)			
Time (knot at 4) by age (knot at 60) interaction	-0.0008 (-0.0018, 0.0001)	-0.0019 (-0.0050, 0.0012)	0.0001 (-0.0001, 0.0003)				0.0015 (0.0008, 0.0022)	0.0100 (0.0076, 0.0124)	0.0003 (0.0001, 0.0005)			

Time by age (knot at 70) interaction	0.0009 (-0.0014, 0.0031)	0.0003 (-0.0073, 0.0078)	-0.0002 (-0.0007, 0.0003)				0.0000 (-0.0017, 0.0017)	-0.0025 (-0.0085, 0.0035)	-0.0002 (-0.0006, 0.0002)			
Time (knot at 1) by age (knot at 70) interaction	0.0029 (-0.0007, 0.0065)	0.0095 (-0.0024, 0.0215)	0.0004 (-0.0004, 0.0011)				0.0002 (-0.0025, 0.0030)	-0.0002 (-0.0098, 0.0094)	0.0003 (-0.0003, 0.0009)			
Time (knot at 2) by age (knot at 70) interaction	-0.0035 (-0.0059, -0.0012)	-0.0109 (-0.0186, -0.0033)	-0.0002 (-0.0007, 0.0003)				-0.0006 (-0.0023, 0.0012)	0.0010 (-0.0051, 0.0072)	-0.0000 (-0.0004, 0.0003)			
Time (knot at 4) by age (knot at 70) interaction	0.0006 (-0.0005, 0.0018)	0.0025 (-0.0012, 0.0062)	0.0000 (-0.0003, 0.0003)				-0.0002 (-0.0011, 0.0006)	-0.0041 (-0.0070, -0.0011)	-0.0001 (-0.0003, 0.0001)			
Time by age (knot at 80) interaction	-0.0011 (-0.0037, 0.0015)	-0.0047 (-0.0135, 0.0041)	0.0000 (-0.0006, 0.0006)				-0.0003 (-0.0020, 0.0014)	-0.0044 (-0.0103, 0.0016)	-0.0000 (-0.0004, 0.0004)			
Time (knot at 1) by age (knot at 80) interaction	-0.0028 (-0.0070, 0.0014)	-0.0092 (-0.0232, 0.0048)	-0.0005 (-0.0014, 0.0004)				-0.0004 (-0.0031, 0.0023)	0.0047 (-0.0048, 0.0141)	-0.0001 (-0.0007, 0.0005)			

Time (knot at 2) by age (knot at 80) interaction	0.0048 (0.0021, 0.0075)	0.0199 (0.0108, 0.0290)	0.0006 (0.0000, 0.0012)				0.0006 (-0.0012, 0.0023)	-0.0012 (-0.0073, 0.0049)	-0.0001 (-0.0005, 0.0003)			
Time (knot at 4) by age (knot at 80) interaction	-0.0014 (-0.0027, -0.0000)	-0.0078 (-0.0121, -0.0034)	-0.0001 (-0.0004, 0.0002)				-0.0004 (-0.0013, 0.0004)	0.0002 (-0.0027, 0.0031)	0.0002 (-0.0000, 0.0004)			
Bowel cancer presence by age interaction	0.0004 (-0.0034, 0.0041)	0.0129 (0.0021, 0.0237)	0.0005 (-0.0001, 0.0011)	0.0441 (-0.0004, 0.0887)	0.0059 (-0.0101, 0.0219)	0.0141 (0.0054, 0.0228)	0.0010 (-0.0025, 0.0045)	0.0057 (-0.0048, 0.0163)	0.0002 (-0.0004, 0.0008)	0.0151 (-0.0355, 0.0657)	0.0069 (-0.0115, 0.0254)	0.0051 (-0.0046, 0.0149)
Bowel cancer presence by age (knot at 60) interaction	-0.0003 (-0.0066, 0.0061)	-0.0241 (-0.0422, -0.0059)	-0.0010 (-0.0021, 0.0001)	-0.1196 (-0.1951, -0.0441)	-0.0385 (-0.0656, -0.0115)	-0.0250 (-0.0395, -0.0104)	-0.0030 (-0.0095, 0.0034)	-0.0255 (-0.0449, -0.0061)	-0.0010 (-0.0021, 0.0001)	-0.0800 (-0.1733, 0.0132)	-0.0370 (-0.0710, -0.0031)	-0.0188 (-0.0366, -0.0010)
Bowel cancer presence by age (knot at 70) interaction	0.0053 (-0.0008, 0.0115)	0.0155 (-0.0020, 0.0330)	0.0011 (0.0001, 0.0022)	-0.0011 (-0.0747, 0.0724)	0.0161 (-0.0100, 0.0423)	0.0116 (-0.0022, 0.0255)	0.0027 (-0.0036, 0.0090)	0.0163 (-0.0026, 0.0352)	0.0009 (-0.0002, 0.0019)	0.0025 (-0.0892, 0.0943)	0.0068 (-0.0264, 0.0400)	0.0083 (-0.0088, 0.0255)
Bowel cancer presence by age (knot at 80) interaction	-0.0094 (-0.0160, -0.0027)	-0.0128 (-0.0319, 0.0063)	-0.0013 (-0.0023, -0.0002)	0.0924 (0.0114, 0.1734)	0.0014 (-0.0272, 0.0301)	-0.0124 (-0.0274, 0.0026)	-0.0006 (-0.0060, 0.0049)	-0.0038 (-0.0200, 0.0124)	-0.0001 (-0.0011, 0.0008)	0.0359 (-0.0429, 0.1147)	0.0080 (-0.0205, 0.0364)	-0.0027 (-0.0173, 0.0119)
Random effects:												

Intercept for patient (variance)	0.1740 (0.1729, 0.1751)	1.4811 (1.4713, 1.4909)	0.0012 (0.0011, 0.0012)	25.9019 (25.7479, 26.0569)	3.1215 (3.1022, 3.1409)	0.7548 (0.7481, 0.7615)	0.1372 (0.1365, 0.1380)	1.2528 (1.2456, 1.2601)	0.0010 (0.0010, 0.0010)	29.6896 (29.5368, 29.8432)	3.7164 (3.6968, 3.7362)	0.9040 (0.8976, 0.9104)
Slope for time to index date[1] (variance)	0.0020 (0.0019, 0.0020)	0.0248 (0.0244, 0.0252)	0.0000 (0.0000, 0.0000)	0.3127 (0.3078, 0.3178)	0.0384 (0.0378, 0.0390)	0.0137 (0.0134, 0.0141)	0.0015 (0.0015, 0.0016)	0.0212 (0.0209, 0.0215)	0.0000 (0.0000, 0.0000)	0.4220 (0.4166, 0.4276)	0.0538 (0.0531, 0.0545)	0.0174 (0.0171, 0.0177)
Covariance between Intercept for patient and slope for time to index date (covariance)	-0.0097 (-0.0099, -0.0095)	-0.1099 (-0.1117, -0.1082)	-0.0001 (-0.0001, -0.0001)	-1.3578 (-1.3820, -1.3336)	-0.1680 (-0.1711, -0.1650)	-0.0622 (-0.0635, -0.0609)	-0.0071 (-0.0072, -0.0069)	-0.0829 (-0.0842, -0.0816)	-0.0001 (-0.0001, -0.0001)	-1.4631 (-1.4882, -1.4380)	-0.1961 (-0.1994, -0.1929)	-0.0713 (-0.0725, -0.0701)
Residual (variance)	0.0506 (0.0505, 0.0508)	0.6220 (0.6200, 0.6241)	0.0005 (0.0005, 0.0005)	6.0476 (6.0268, 6.0685)	0.8071 (0.8043, 0.8100)	0.5825 (0.5804, 0.5845)	0.0447 (0.0446, 0.0449)	0.5960 (0.5945, 0.5976)	0.0005 (0.0004, 0.0005)	7.9522 (7.9311, 7.9734)	1.0205 (1.0178, 1.0233)	0.6239 (0.6222, 0.6256)

[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.

Abbreviations: RBC=red blood cells, Hb=haemoglobin, Hc=haematocrit, MCV=mean corpuscular volume, MCH=mean corpuscular haemoglobin, MCHC=mean corpuscular haemoglobin concentration.

Table C6: Mixed-effect model coefficients (95% CI) for platelet parameters

Variable	Males		Females	
	Platelets	**MPV**	**Platelets**	**MPV**
N	*383016*	*59412*	*517999*	*81765*
n cases	*1283*	*8912*	*1175*	*7868*
Fixed effects:				
Constant	257.1629 (254.9297, 259.3962)	9.5231 (9.4712, 9.5749)	276.9223 (274.8745, 278.9700)	9.6562 (9.6165, 9.6959)
Age at index date[1] (years)	-0.2013 (-0.2445, -0.1581)	-0.0031 (-0.0040, -0.0023)	-0.0199 (-0.0603, 0.0205)	-0.0046 (-0.0052, -0.0040)
Age at index date[1] – knot at 60 (years)	-0.5702 (-0.6874, -0.4529)		-0.4970 (-0.6110, -0.3830)	
Age at index date[1] – knot at 70 (years)	0.4389 (0.2598, 0.6180)		0.9261 (0.7584, 1.0938)	
Age at index date[1] – knot at 80 (years)	0.1632 (-0.0635, 0.3899)		-0.7335 (-0.9056, -0.5615)	
Time to index date[1] (years)	2.7155 (2.2004, 3.2306)	-0.0974 (-0.1178, -0.0770)	3.1482 (2.7212, 3.5753)	-0.0902 (-0.1057, -0.0746)
Time to index date[1] date – knot at 1 (years)	-0.7797 (-1.6028, 0.0434)	0.0433 (0.0112, 0.0753)	-1.3100 (-1.9898, -0.6302)	0.0331 (0.0087, 0.0576)
Time to index date[1] date – knot at 2 (years)	0.3853 (-0.1495, 0.9200)	0.0291 (0.0084, 0.0498)	-0.1326 (-0.5733, 0.3080)	0.0453 (0.0295, 0.0612)
Time to index date[1] date – knot at 4 (years)	-2.2486 (-2.5110, -1.9862)	0.0148 (0.0038, 0.0258)	-2.1814 (-2.3969, -1.9658)	-0.0010 (-0.0094, 0.0074)
Bowel cancer present	128.6209 (96.3364, 160.9054)	-0.3766 (-0.4700, -0.2831)	110.9094 (74.8031, 147.0156)	-0.4480 (-0.5434, -0.3525)
Bowel cancer by time to index date[1] interaction	-57.3824 (-59.7265, -55.0383)	0.1735 (0.0770, 0.2699)	-70.0709 (-72.5177, -67.6241)	0.3215 (0.2301, 0.4130)
Bowel cancer by time to index date[1] (knot at 1) interaction	50.8421 (46.6410, 55.0433)	-0.1107 (-0.2790, 0.0576)	61.9149 (57.5957, 66.2340)	-0.2983 (-0.4579, -0.1387)

Bowel cancer by time to index date[1] (knot at 2) interaction	2.9225 (-0.1281, 5.9731)	-0.0245 (-0.1446, 0.0955)	3.9056 (0.8180, 6.9931)	-0.0328 (-0.1470, 0.0813)
Bowel cancer by time to index date[1] (knot at 4) interaction	3.5438 (2.0653, 5.0224)	-0.0356 (-0.0989, 0.0276)	3.0603 (1.5774, 4.5433)	0.0439 (-0.0152, 0.1031)
Bowel cancer by age interaction	-0.9456 (-1.5201, -0.3710)		-0.4272 (-1.0804, 0.2260)	
Bowel cancer by age (knot at 60) interaction	0.7702 (-0.2018, 1.7422)		1.2526 (0.0544, 2.4507)	
Bowel cancer by age (knot at 70) interaction	0.9991 (0.0572, 1.9410)		-0.8446 (-2.0166, 0.3275)	
Bowel cancer by age (knot at 80) interaction	-2.1333 (-3.1669, -1.0998)		-0.0662 (-1.0705, 0.9380)	
Random effects:				
Intercept for patient (variance)	3.8e+03 (3.7e+03, 3.8e+03)	1.8264 (1.7987, 1.8546)	4.5e+03 (4.5e+03, 4.6e+03)	1.8457 (1.8225, 1.8691)
Slope for time to index date[1] (variance)	39.5014 (38.6595, 40.3616)	0.0271 (0.0260, 0.0282)	47.4155 (46.6445, 48.1993)	0.0274 (0.0266, 0.0283)
Covariance between Intercept for patient and slope for time to index date (covariance)	-1.7e+02 (-1.7e+02, -1.6e+02)	-0.1103 (-0.1151, -0.1055)	-1.9e+02 (-2.0e+02, -1.9e+02)	-0.1117 (-0.1155, -0.1079)
Residual (variance)	1.6e+03 (1.6e+03, 1.7e+03)	0.2919 (0.2890, 0.2949)	1.8e+03 (1.8e+03, 1.8e+03)	0.2884 (0.2862, 0.2906)

[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.

Abbreviations: MPV=mean platelet volume.

Table C7: Mixed-effect model coefficients (95% CI) for white blood cell parameters

Variable	Males						Females					
	WBC	**Basophils**	**Eosinophils**	**Lymphocytes**	**Monocytes**	**Neutrophils**	**WBC**	**Basophils**	**Eosinophils**	**Lymphocytes**	**Monocytes**	**Neutrophils**
N	*356191*	*328523*	*329715*	*338453*	*335510*	*339093*	*485552*	*444145*	*445731*	*461204*	*456981*	*461893*
n cases	*8133*	*7869*	*7892*	*8113*	*8055*	*8509*	*7249*	*6941*	*6966*	*7224*	*7155*	*7574*
Fixed effects:												
Constant	6.8860 (6.7850, 6.9871)	0.0653 (0.0645, 0.0660)	0.2148 (0.2110, 0.2186)	2.4595 (2.4323, 2.4867)	0.4892 (0.4808, 0.4977)	3.9542 (3.8711, 4.0372)	8.1097 (8.0197, 8.1997)	0.0614 (0.0608, 0.0619)	0.1950 (0.1922, 0.1978)	2.0705 (2.0519, 2.0891)	0.5118 (0.5057, 0.5179)	5.7413 (5.6847, 5.7978)
Age at index date[1] (years)	0.0050 (0.0031, 0.0070)	0.0000 (-0.0000, 0.0000)	0.0003 (0.0002, 0.0003)	-0.0063 (-0.0067, -0.0058)	0.0013 (0.0012, 0.0015)	0.0048 (0.0032, 0.0063)	-0.0218 (-0.0235, -0.0200)	0.0001 (0.0001, 0.0001)	0.0001 (0.0000, 0.0001)	-0.0001 (-0.0004, 0.0002)	-0.0005 (-0.0006, -0.0004)	-0.0311 (-0.0322, -0.0300)
Age at index date[1] – knot at 60 (years)	0.0025 (-0.0025, 0.0074)				0.0008 (0.0003, 0.0012)	0.0141 (0.0113, 0.0170)	0.0461 (0.0416, 0.0507)				0.0037 (0.0033, 0.0040)	0.0652 (0.0633, 0.0671)
Age at index date[1] – knot at 70 (years)	0.0066 (0.0004, 0.0127)		0.0006 (0.0005, 0.0008)		0.0001 (-0.0003, 0.0006)		0.0041 (-0.0013, 0.0095)		0.0001 (0.0000, 0.0002)		0.0017 (0.0012, 0.0022)	

Age at index date[1] – knot at 80 (years)				-0.0135 (-0.0165, -0.0105)		0.0010 (-0.0049, 0.0070)				-0.0337 (-0.0357, -0.0316)	-0.0029 (-0.0034, -0.0024)	-0.0131 (-0.0161, -0.0100)
Age at index date[1] – knot at 85 (years)	-0.0126 (-0.0254, 0.0002)						-0.0395 (-0.0473, -0.0317)					
Age at index date[1] – knot at 90 (years)										0.0420 (0.0353, 0.0486)		
Time to index date[1] (years)	0.0046 (-0.0261, 0.0354)	-0.0002 (-0.0002, -0.0001)	-0.0009 (-0.0011, -0.0008)	0.0102 (-0.0019, 0.0223)	-0.0040 (-0.0065, -0.0016)	-0.0436 (-0.0706, -0.0166)	0.0004 (-0.0289, 0.0298)	-0.0003 (-0.0003, -0.0003)	-0.0004 (-0.0005, -0.0002)	0.0192 (0.0098, 0.0286)	-0.0042 (-0.0058, -0.0025)	-0.0226 (-0.0417, -0.0035)
Time to index date[1] date – knot at 1 (years)	-0.0248 (-0.0743, 0.0246)			0.0058 (-0.0134, 0.0249)	-0.0031 (-0.0070, 0.0008)	0.0254 (-0.0186, 0.0694)	-0.0040 (-0.0512, 0.0433)			-0.0118 (-0.0268, 0.0032)	0.0004 (-0.0023, 0.0031)	0.0269 (-0.0039, 0.0577)
Time to index date[1] date – knot at 2 (years)	0.0081 (-0.0244, 0.0405)			-0.0029 (-0.0152, 0.0094)	0.0021 (-0.0005, 0.0046)	0.0091 (-0.0206, 0.0388)	-0.0011 (-0.0320, 0.0299)			0.0020 (-0.0077, 0.0117)	-0.0016 (-0.0034, 0.0001)	-0.0066 (-0.0270, 0.0137)

Time to index date[1] date – knot at 4 (years)	0.0105 (-0.0054, 0.0264)			-0.0025 (-0.0088, 0.0037)	0.0004 (-0.0008, 0.0018)	0.0086 (-0.0064, 0.0235)	0.0114 (-0.0036, 0.0265)			-0.0039 (-0.0086, 0.0008)	0.0022 (0.0013, 0.0030)	0.0130 (0.0029, 0.0231)
Bowel cancer present	-0.5156 (-1.9304, 0.8992)	0.0018 (0.0006, 0.0031)	-0.0198 (-0.0636, 0.0241)	-0.1594 (-0.2052, -0.1136)	0.1034 (0.0959, 0.1108)	1.4463 (0.3253, 2.5673)	0.4778 (-1.0104, 1.9660)	0.0016 (0.0003, 0.0029)	0.0020 (-0.0378, 0.0418)	-0.1701 (-0.2112, -0.1290)	0.1282 (0.1214, 0.1350)	1.5446 (1.4734, 1.6159)
Bowel cancer present by time to index date[1] interaction	-0.8386 (-0.9812, -0.6959)	-0.0002 (-0.0005, 0.0001)	-0.0021 (-0.0031, -0.0012)	0.1662 (0.1123, 0.2202)	-0.0823 (-0.0935, -0.0710)	-0.9473 (-1.0745, -0.8201)	-1.4093 (-1.5821, -1.2365)	-0.0003 (-0.0006, -0.0000)	-0.0019 (-0.0028, -0.0011)	0.1314 (0.0773, 0.1855)	-0.1098 (-0.1195, -0.1002)	-1.4270 (-1.5393, -1.3147)
Bowel cancer present by time to index date[1] (knot at 1) interaction	0.6959 (0.4376, 0.9541)			-0.1616 (-0.2574, -0.0658)	0.0806 (0.0602, 0.1011)	0.8945 (0.6613, 1.1278)	1.4600 (1.1503, 1.7696)			-0.1227 (-0.2185, -0.0269)	0.1090 (0.0918, 0.1263)	1.4181 (1.2162, 1.6200)
Bowel cancer present by time to index date[1] (knot at 2) interaction	0.1083 (-0.0814, 0.2980)			0.0037 (-0.0658, 0.0731)	-0.0079 (-0.0230, 0.0072)	0.0181 (-0.1570, 0.1933)	-0.0444 (-0.2687, 0.1799)			0.0302 (-0.0384, 0.0989)	-0.0007 (-0.0132, 0.0118)	0.0331 (-0.1145, 0.1808)
Bowel cancer present by time to index date[1] (knot at 4) interaction	0.0194 (-0.0721, 0.1110)			-0.0147 (-0.0497, 0.0202)	0.0100 (0.0026, 0.0174)	0.0487 (-0.0376, 0.1349)	-0.0650 (-0.1713, 0.0413)			-0.0378 (-0.0706, -0.0049)	-0.0036 (-0.0096, 0.0025)	-0.1051 (-0.1766, -0.0335)

Bowel cancer presence by age interaction	0.0332 (0.0081, 0.0583)		0.0006 (-0.0001, 0.0013)			-0.0053 (-0.0249, 0.0144)	0.0190 (-0.0079, 0.0459)		0.0002 (-0.0004, 0.0008)			
Bowel cancer presence by age (knot at 60) interaction	-0.0825 (-0.1236, -0.0415)					0.0015 (-0.0232, 0.0263)	-0.0325 (-0.0800, 0.0150)					
Bowel cancer presence by age (knot at 70) interaction	0.0546 (0.0221, 0.0872)		-0.0026 (-0.0037, -0.0014)				0.0125 (-0.0251, 0.0501)		-0.0011 (-0.0021, -0.0001)			
Bowel cancer presence by age (knot at 80) interaction						-0.0134 (-0.0403, 0.0135)						
Bowel cancer presence by age (knot at 85) interaction	-0.0695 (-0.1293, -0.0097)						0.0202 (-0.0244, 0.0647)					
Random effects:												
Intercept for patient (variance)	5.3129 (5.2597, 5.3667)	0.0016 (0.0016, 0.0017)	0.0219 (0.0217, 0.0221)	2.3665 (2.3502, 2.3830)	0.0348 (0.0344, 0.0351)	2.1016 (2.0698, 2.1338)	4.9756 (4.9232, 5.0286)	0.0016 (0.0016, 0.0016)	0.0169 (0.0167, 0.0170)	1.3624 (1.3534, 1.3716)	0.0295 (0.0292, 0.0297)	1.9321 (1.9110, 1.9535)

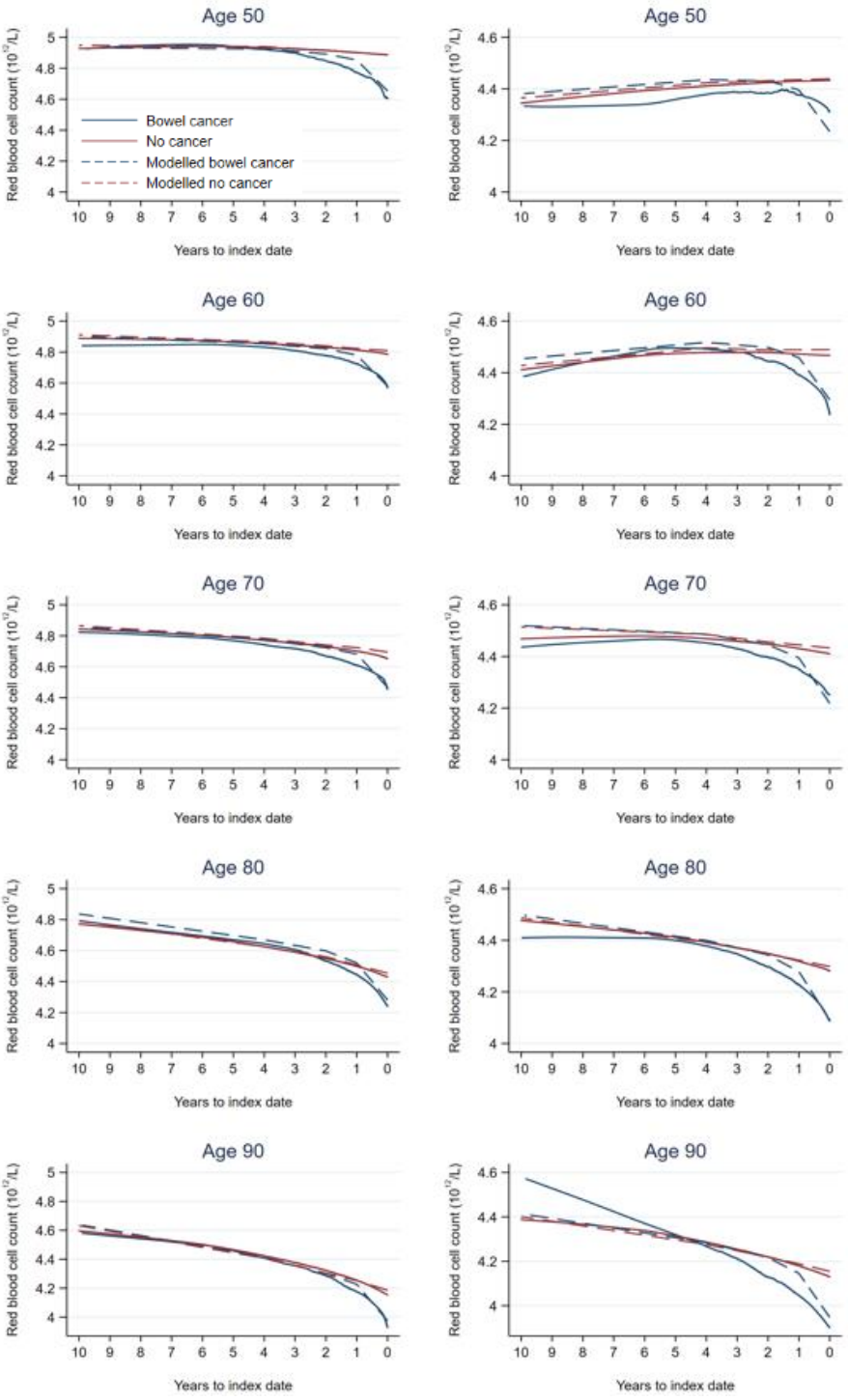
Slope for time to index date[1] (variance)	0.0876 (0.0853, 0.0899)	0.0000 (0.0000, 0.0000)	0.0003 (0.0003, 0.0003)	0.0583 (0.0574, 0.0592)	0.0005 (0.0005, 0.0006)	0.0410 (0.0389, 0.0432)	0.1120 (0.1090, 0.1151)	0.0000 (0.0000, 0.0000)	0.0002 (0.0002, 0.0002)	0.0147 (0.0144, 0.0151)	0.0004 (0.0004, 0.0004)	0.0471 (0.0459, 0.0484)
Covariance between Intercept for patient and slope for time to index date (covariance)	-0.3649 (-0.3756, -0.3542)	-0.0001 (-0.0001, -0.0001)	-0.0011 (-0.0012, -0.0011)	-0.2447 (-0.2483, -0.2410)	-0.0014 (-0.0014, -0.0013)	0.1588 (0.1528, 0.1648)	-0.4378 (-0.4496, -0.4261)	-0.0001 (-0.0001, -0.0001)	-0.0009 (-0.0009, -0.0009)	-0.1000 (-0.1017, -0.0983)	-0.0018 (-0.0019, -0.0018)	-0.1056 (-0.1102, -0.1010)
Residual (variance)	5.8167 (5.7964, 5.8372)	0.0017 (0.0017, 0.0017)	0.0141 (0.0140, 0.0141)	0.6405 (0.6380, 0.6430)	0.0334 (0.0333, 0.0336)	4.7837 (4.7650, 4.8024)	9.0166 (8.9919, 9.0413)	0.0016 (0.0016, 0.0016)	0.0116 (0.0116, 0.0116)	0.7342 (0.7320, 0.7364)	0.0243 (0.0242, 0.0244)	3.6243 (3.6139, 3.6347)

[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.

Abbreviations: WBC=white blood cell count.

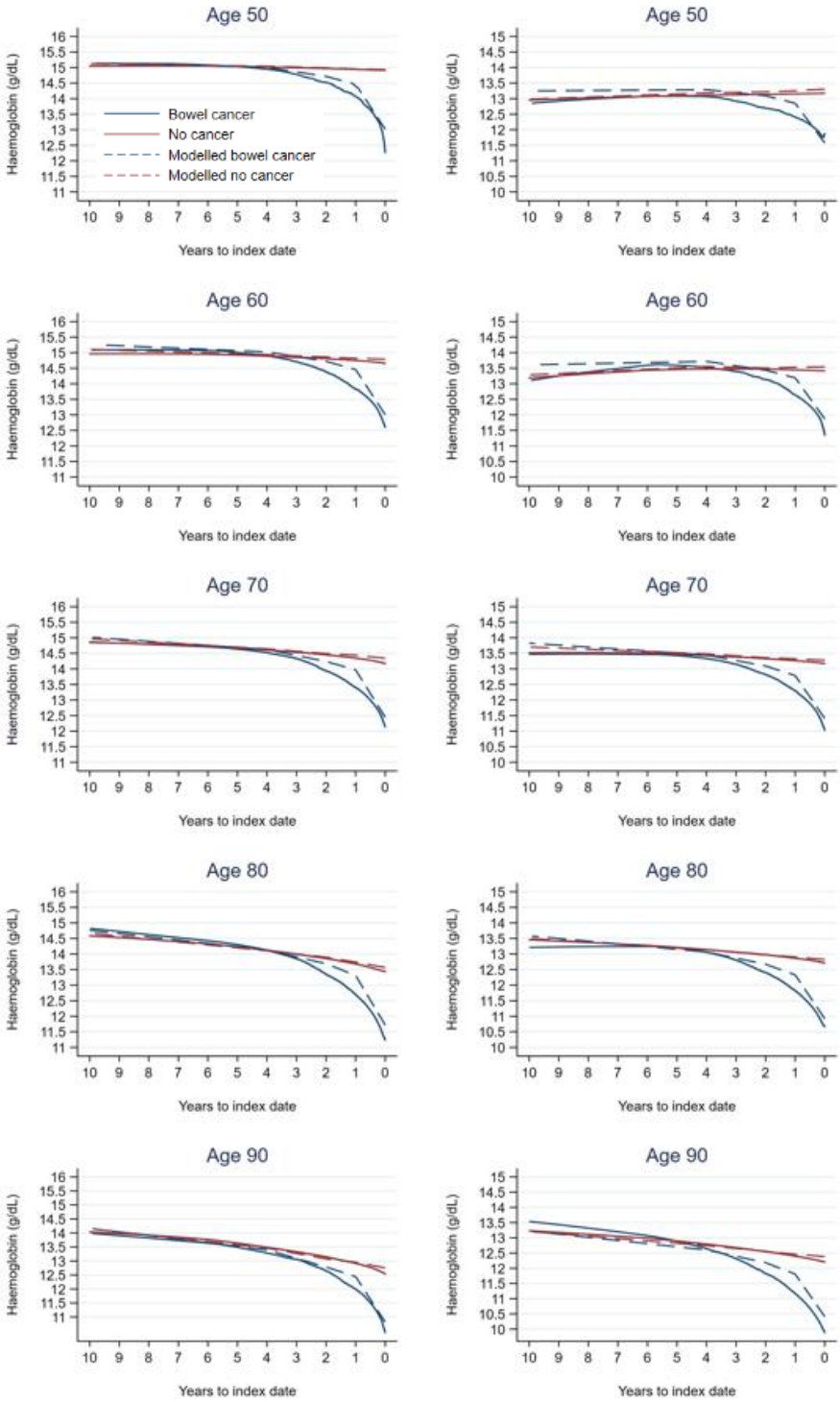
C7. Trajectories (LOWESS and modelled) in each FBC parameter

Figure C45: Trajectories (LOWESS and fixed-effects) in red blood cell count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



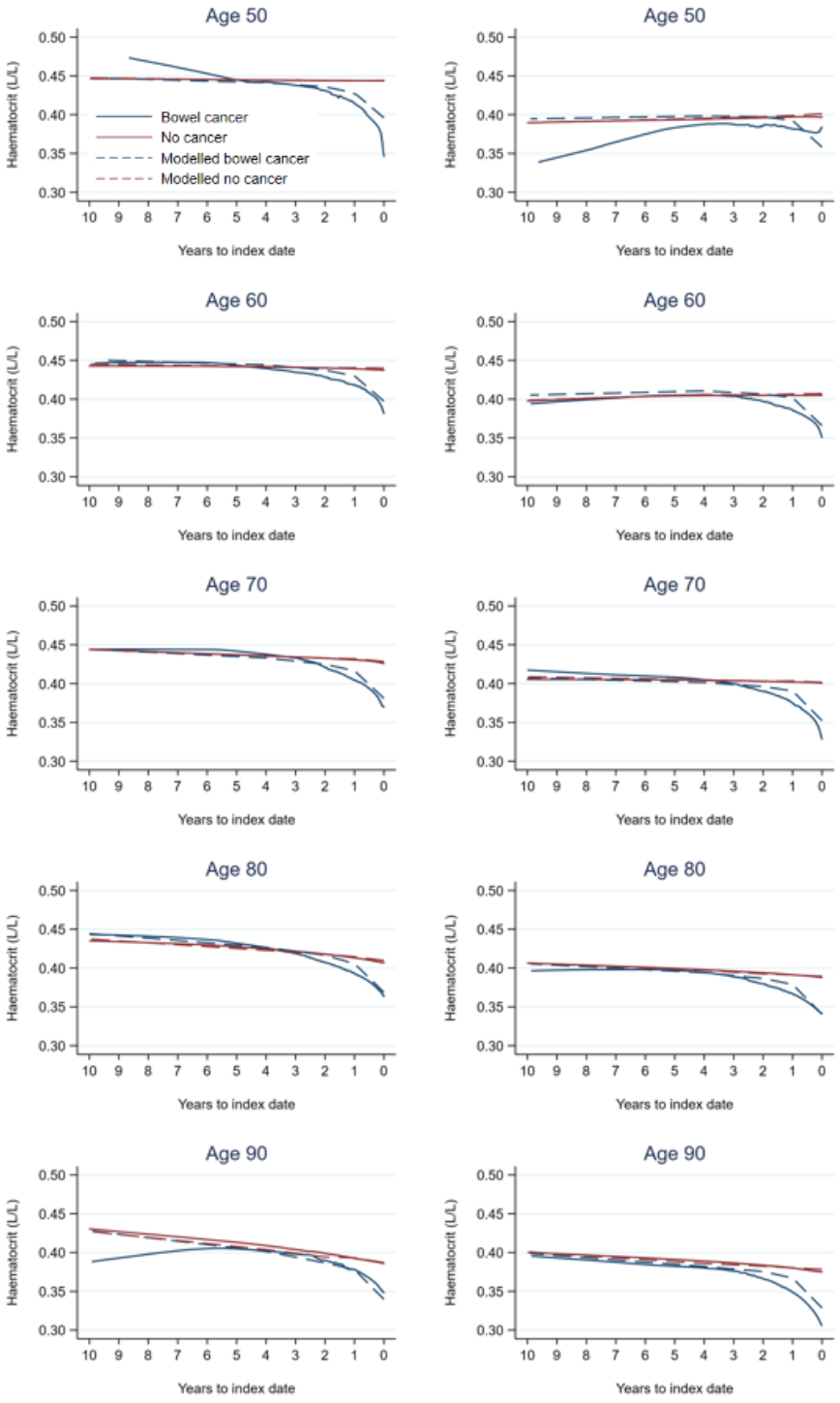
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C46: Trajectories (LOWESS and fixed-effects) in haemoglobin for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



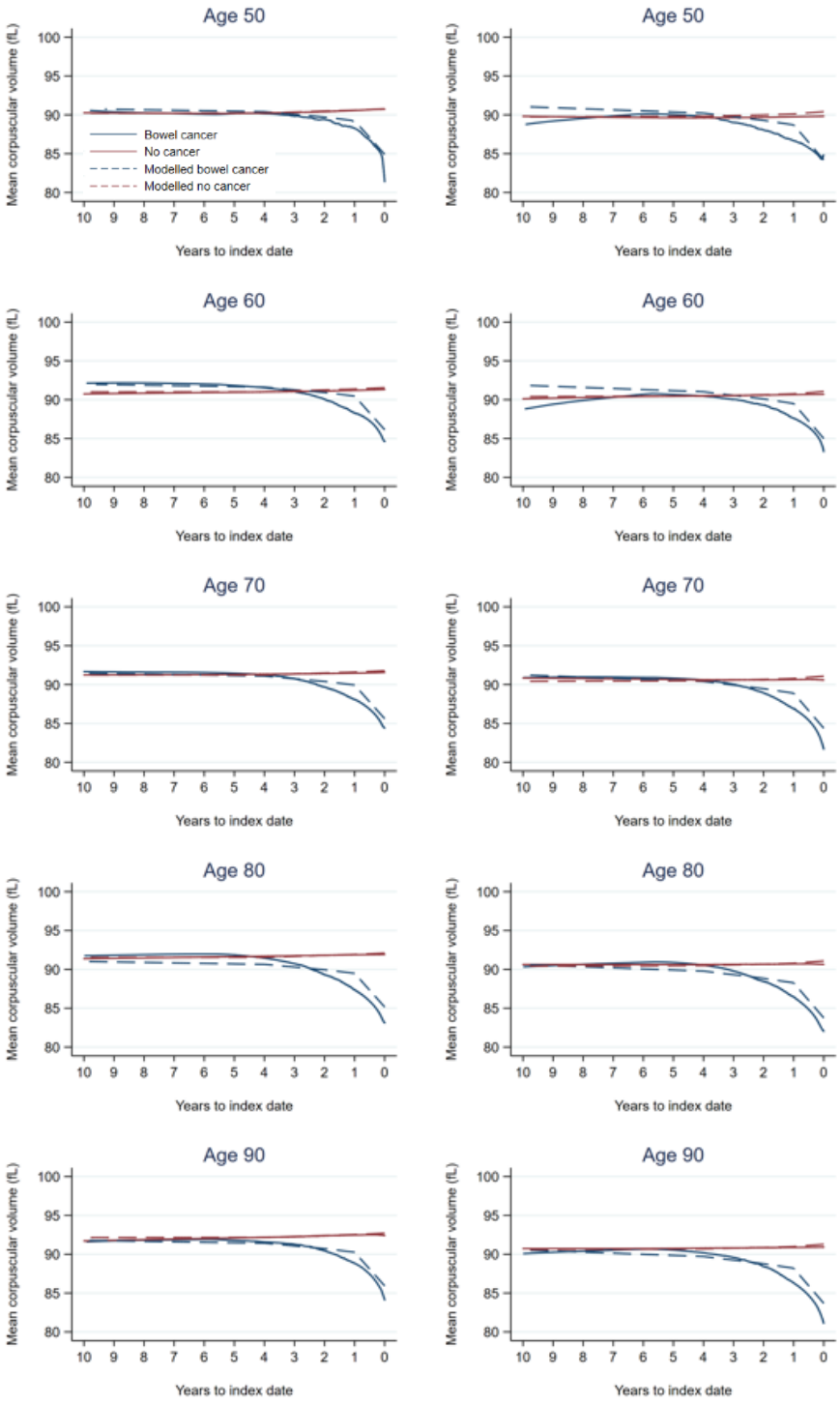
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C47: Trajectories (LOWESS and fixed-effects) in haematocrit for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



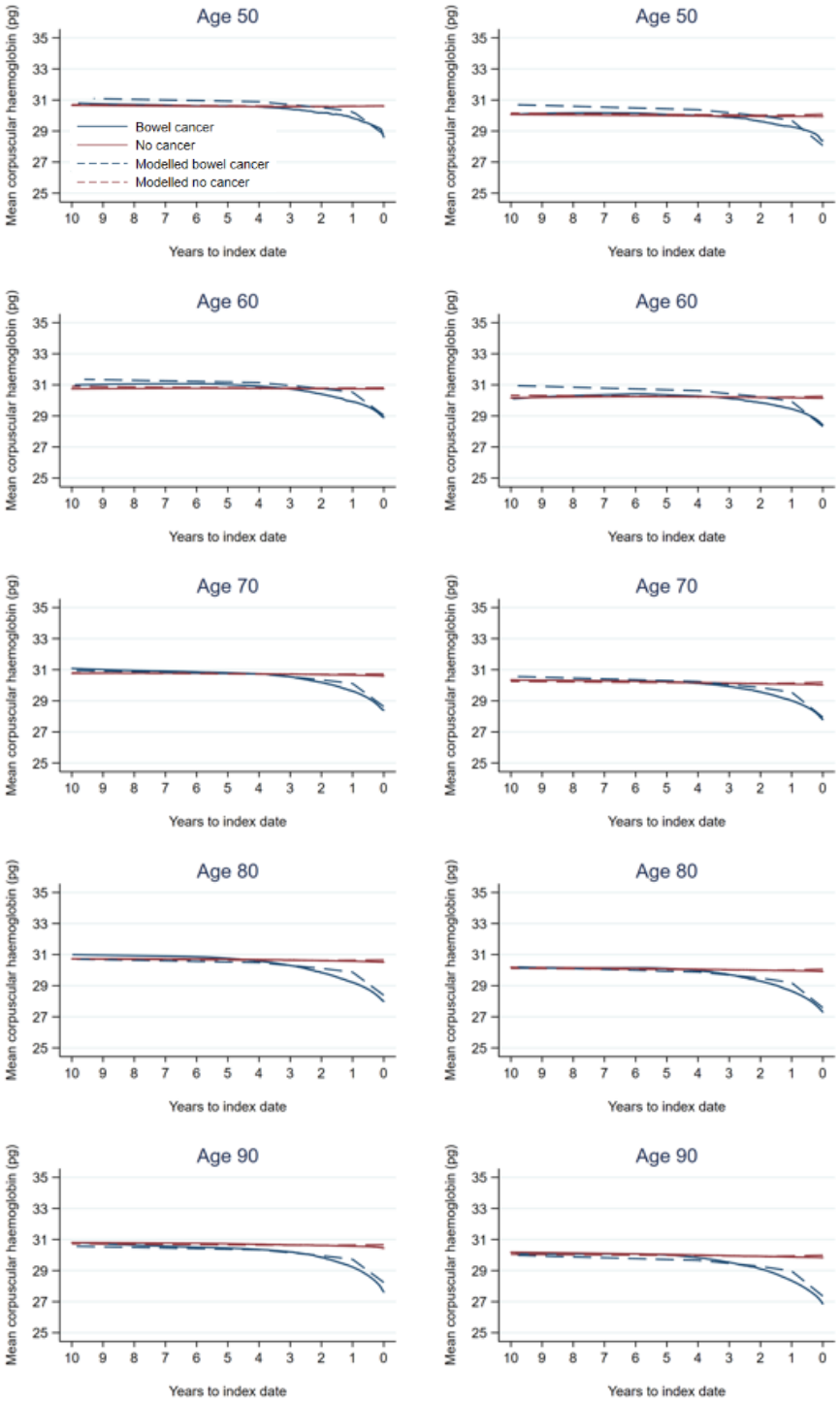
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C48: Trajectories (LOWESS and fixed-effects) in mean corpuscular volume for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



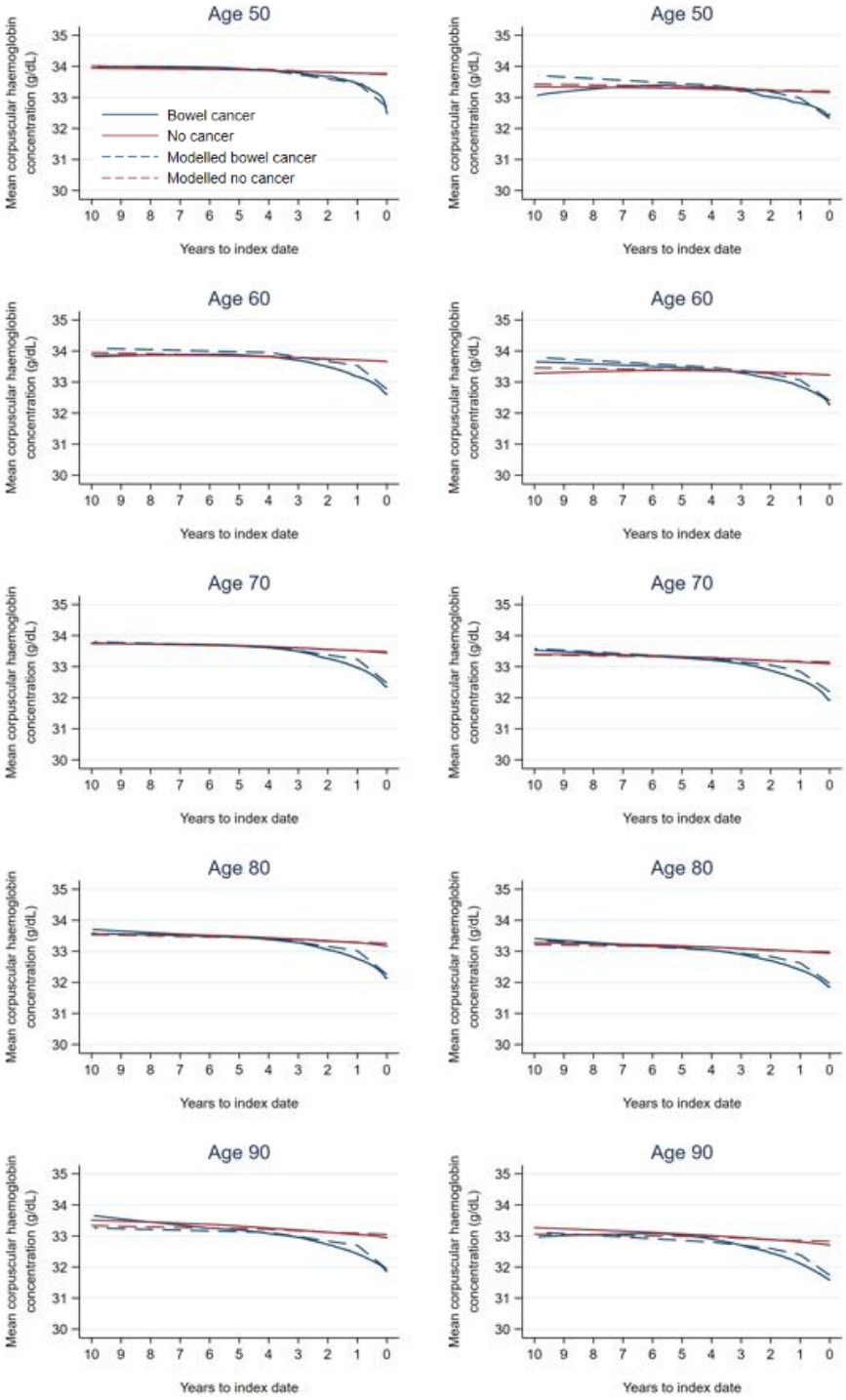
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C49: Trajectories (LOWESS and fixed-effects) in mean corpuscular haemoglobin for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



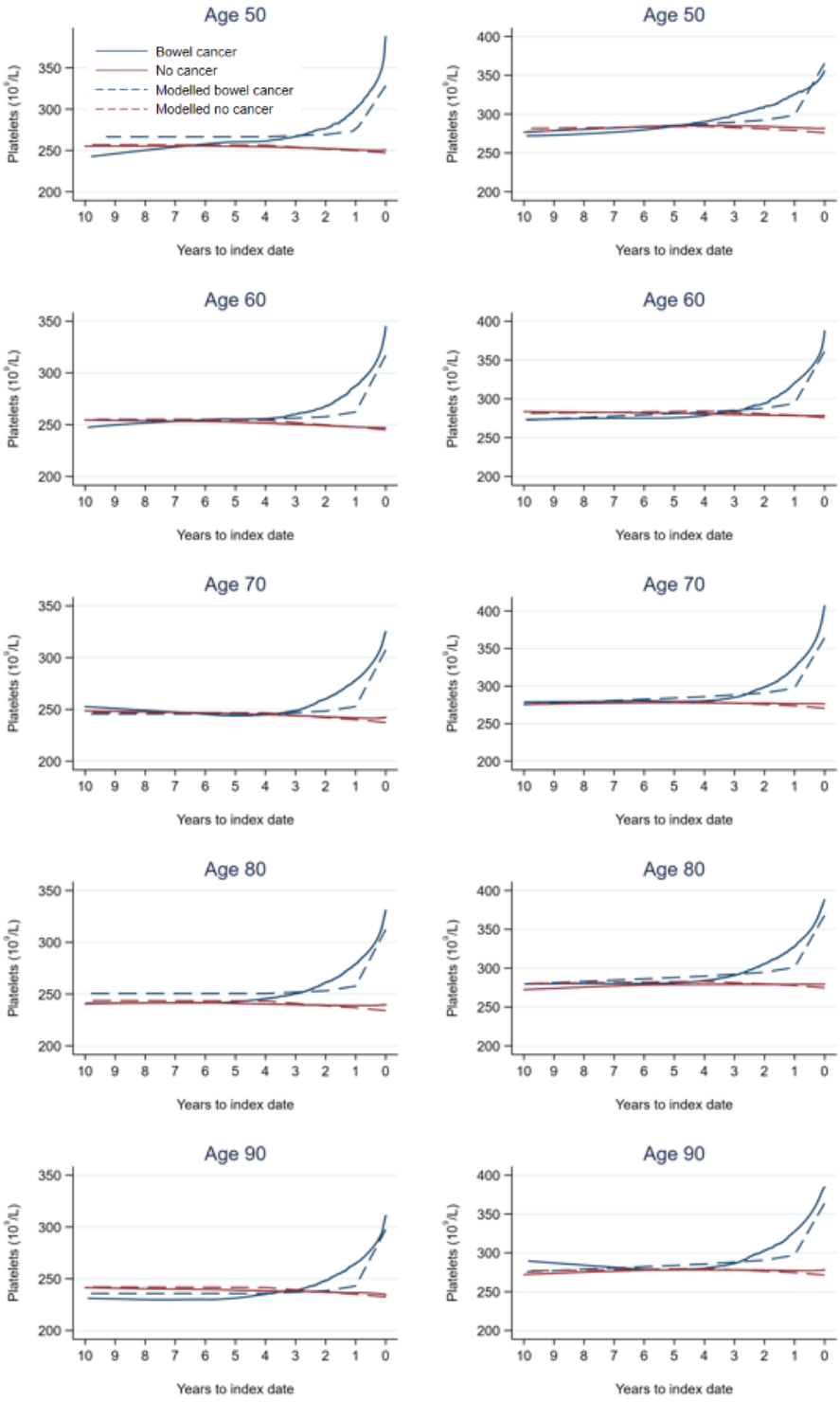
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C50: Trajectories (LOWESS and fixed-effects) in mean corpuscular haemoglobin concentration for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



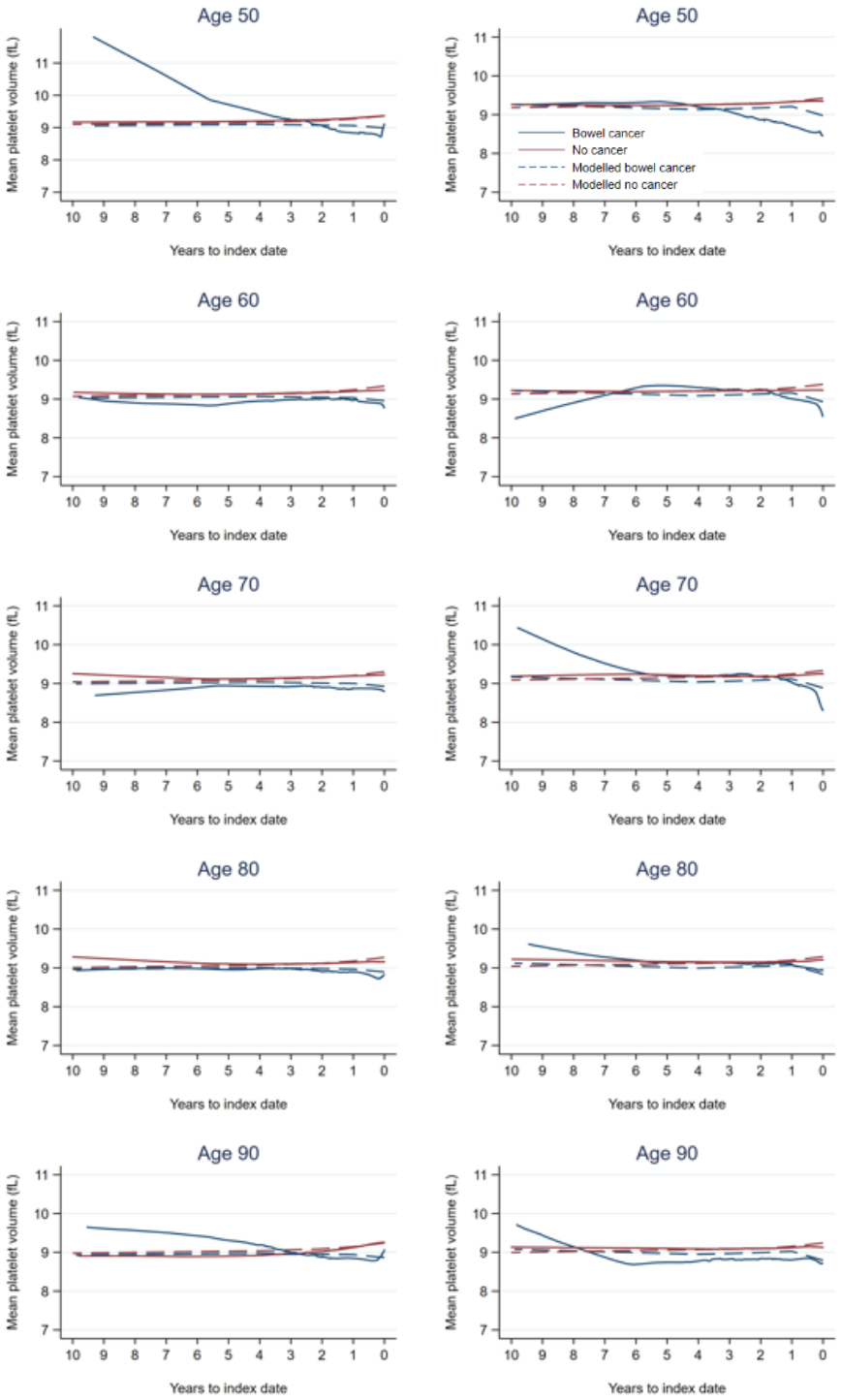
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C51: Trajectories (LOWESS and fixed-effects) in platelets for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



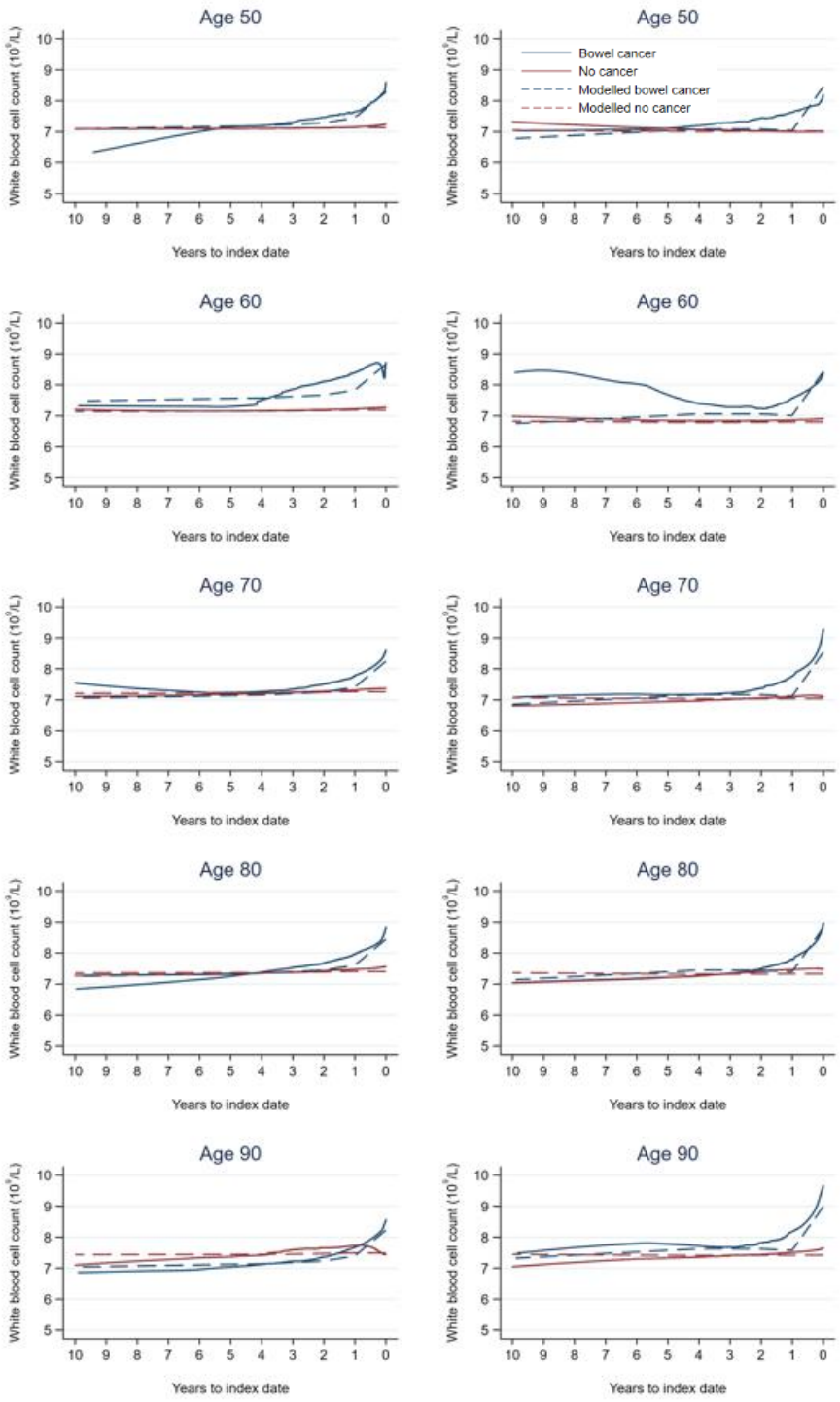
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C52: Trajectories (LOWESS and fixed-effects) in mean platelet volume for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



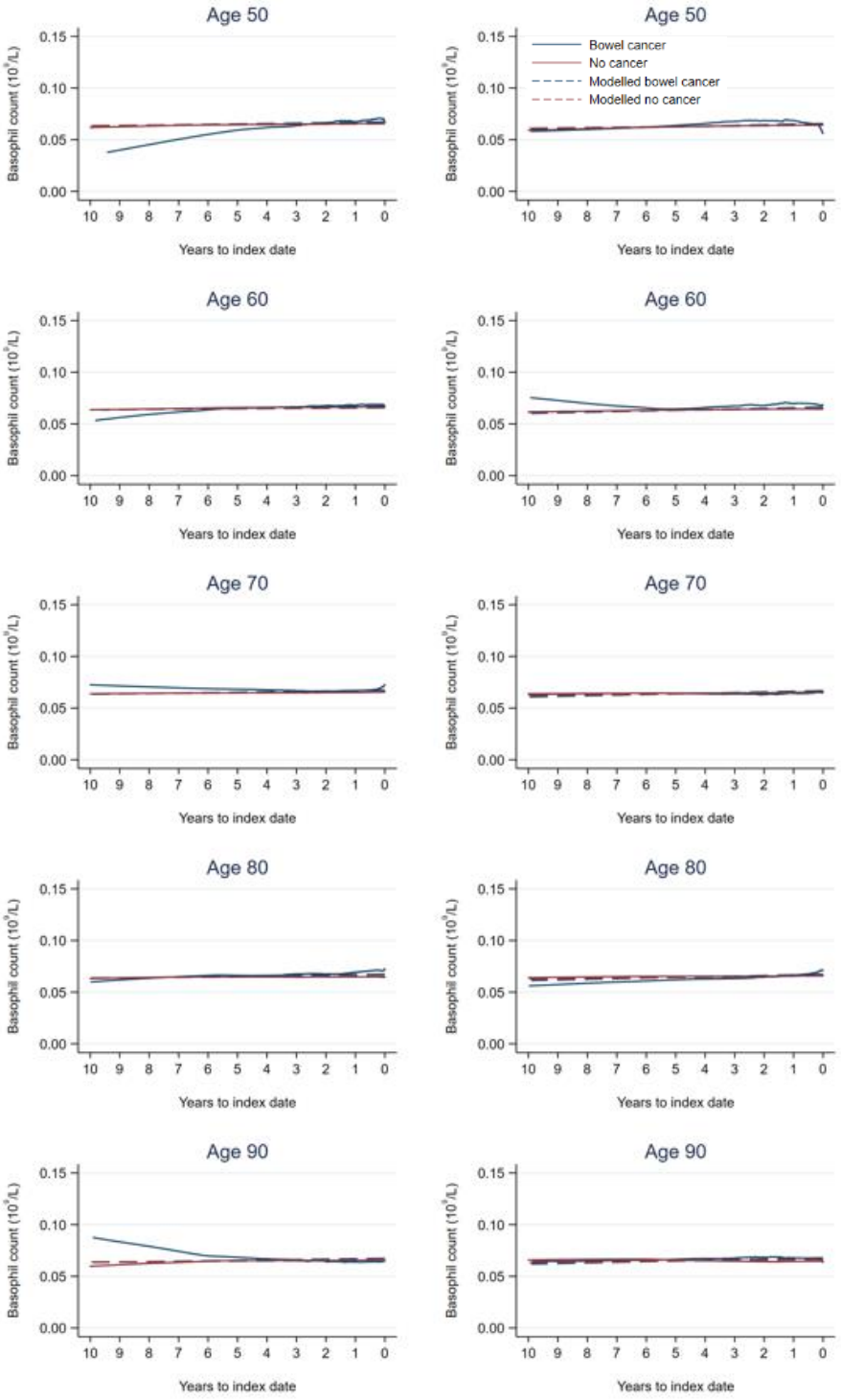
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C53: Trajectories (LOWESS and fixed-effects) in white blood cell count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



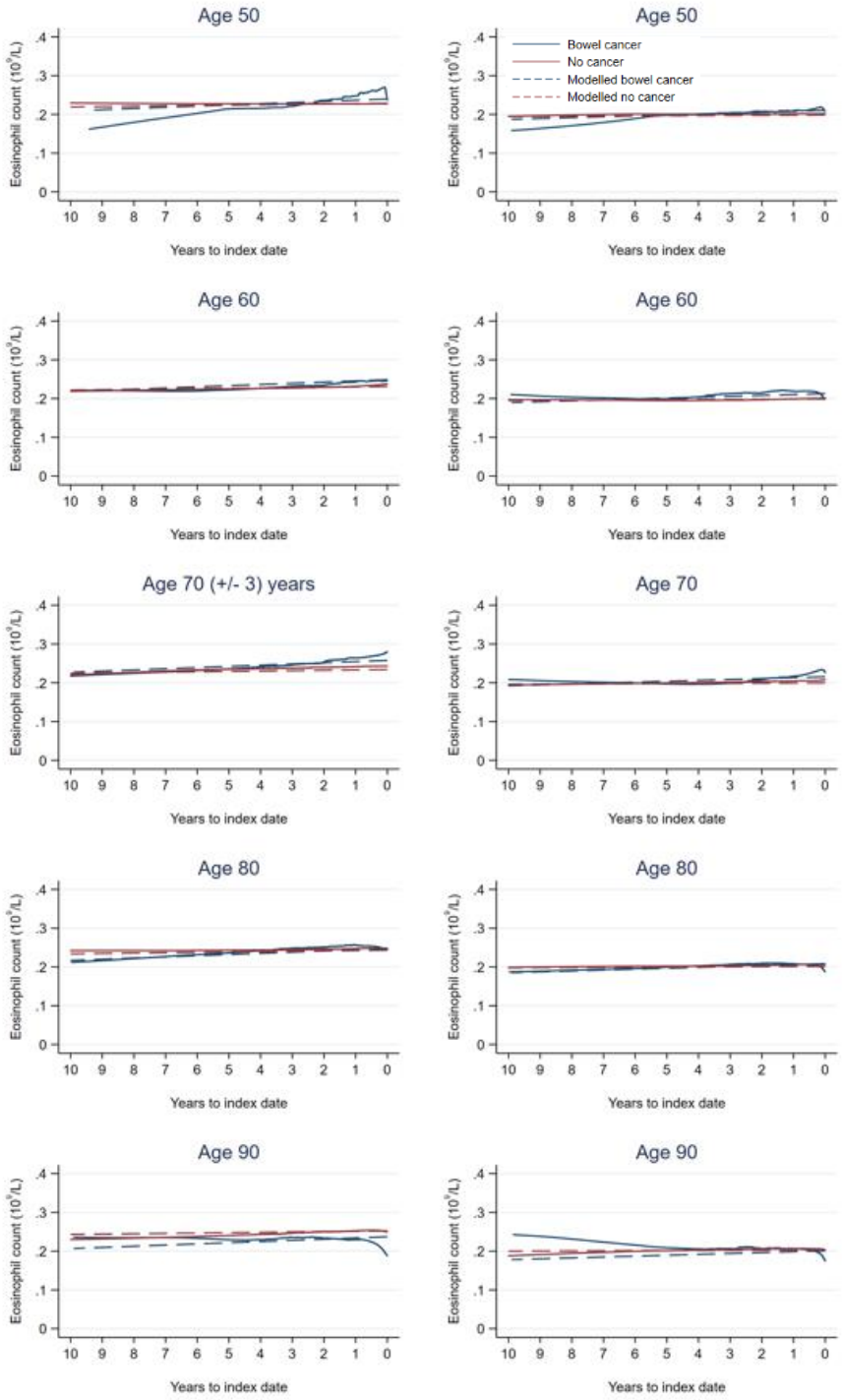
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C54: Trajectories (LOWESS and fixed-effects) in basophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



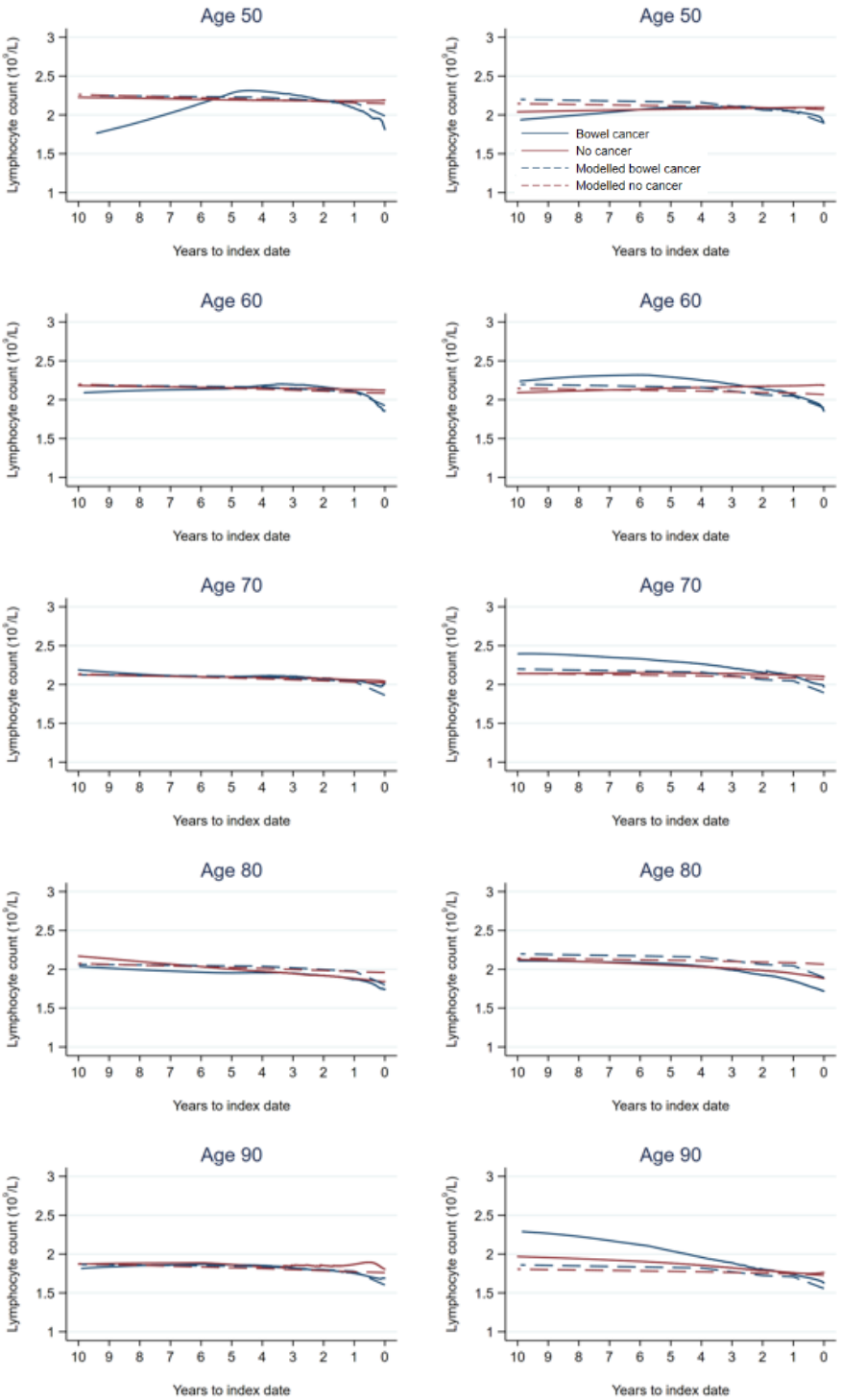
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C55: Trajectories (LOWESS and fixed-effects) in eosinophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



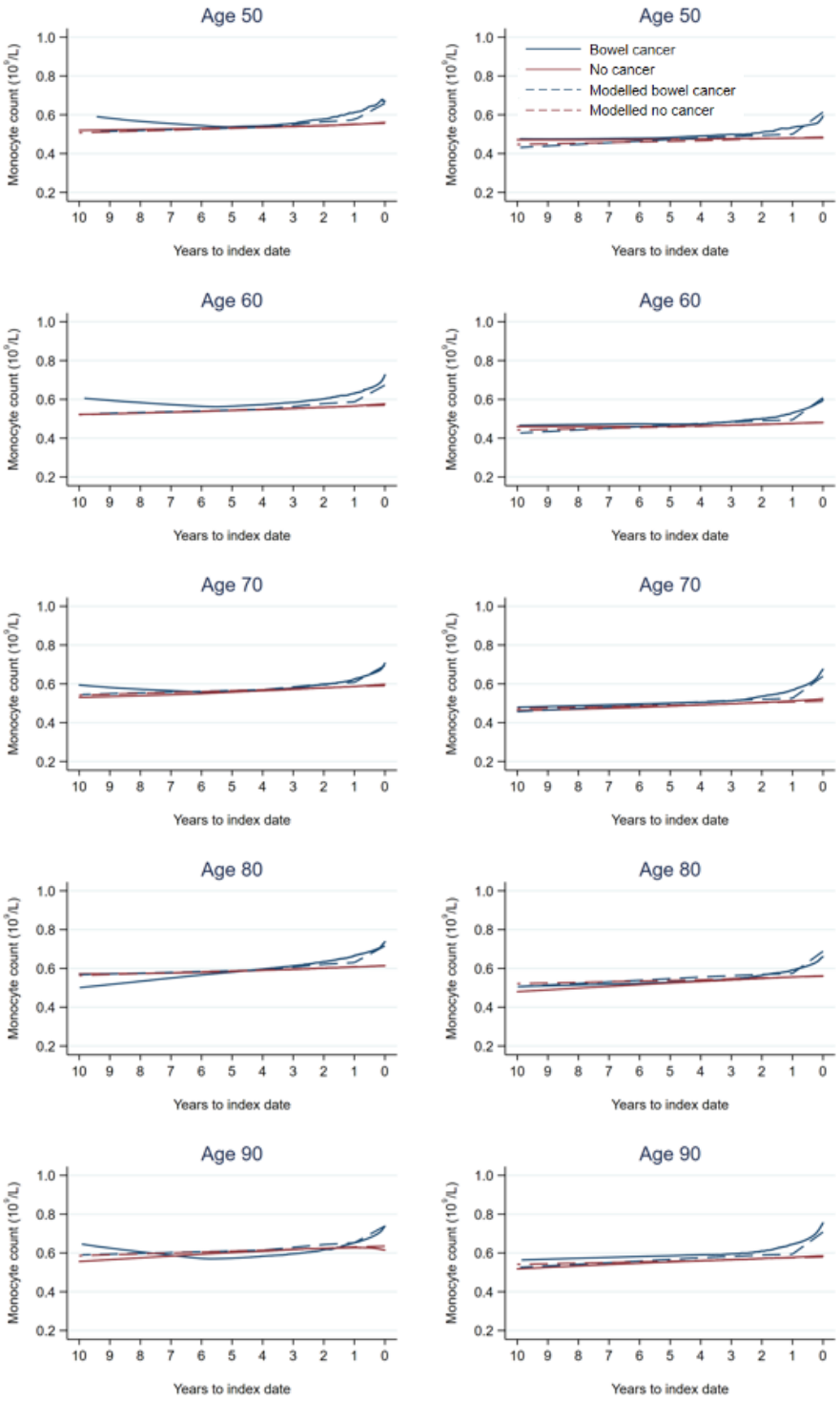
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C56: Trajectories (LOWESS and fixed-effects) in lymphocyte count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



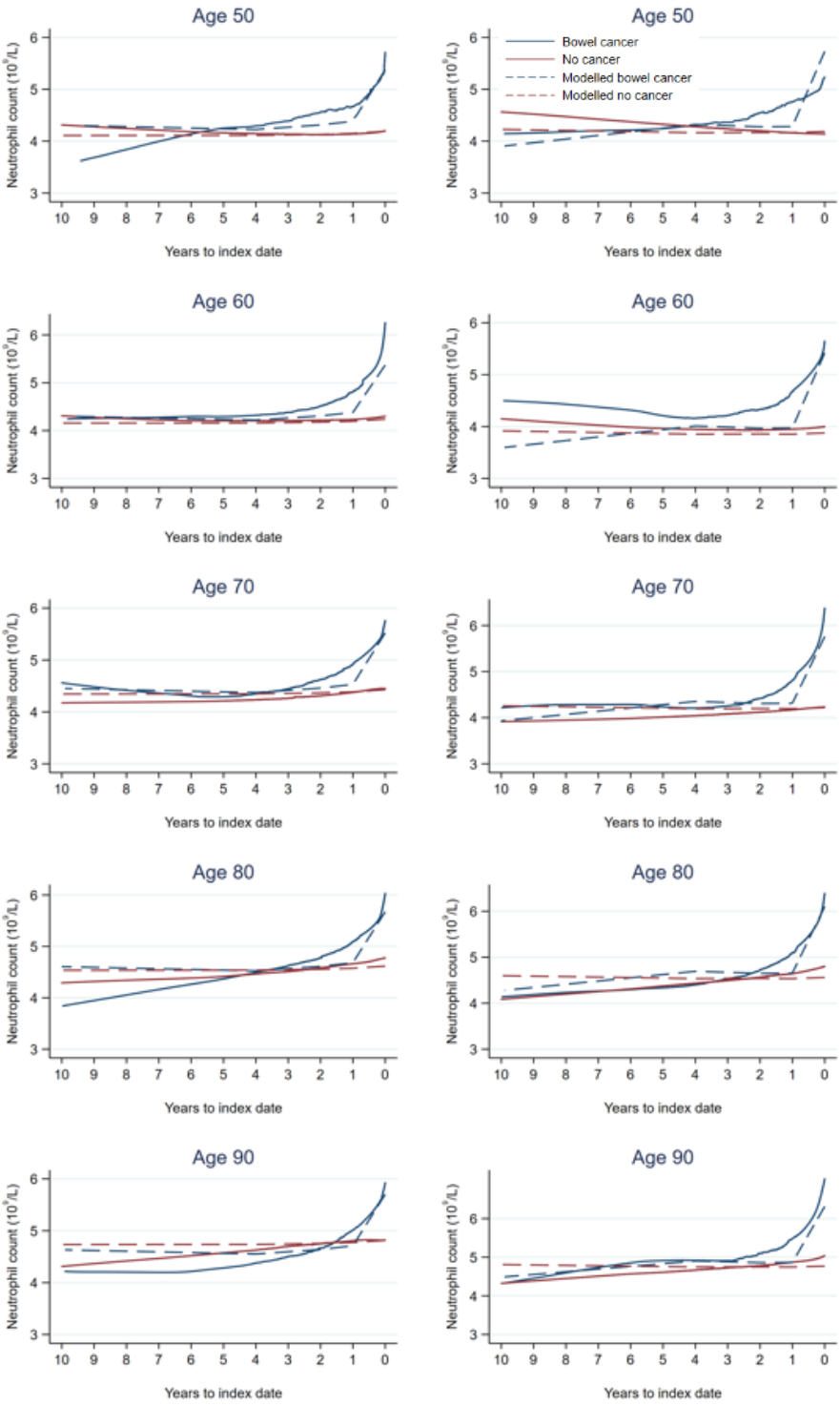
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C57: Trajectories (LOWESS and fixed-effects) in monocyte count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

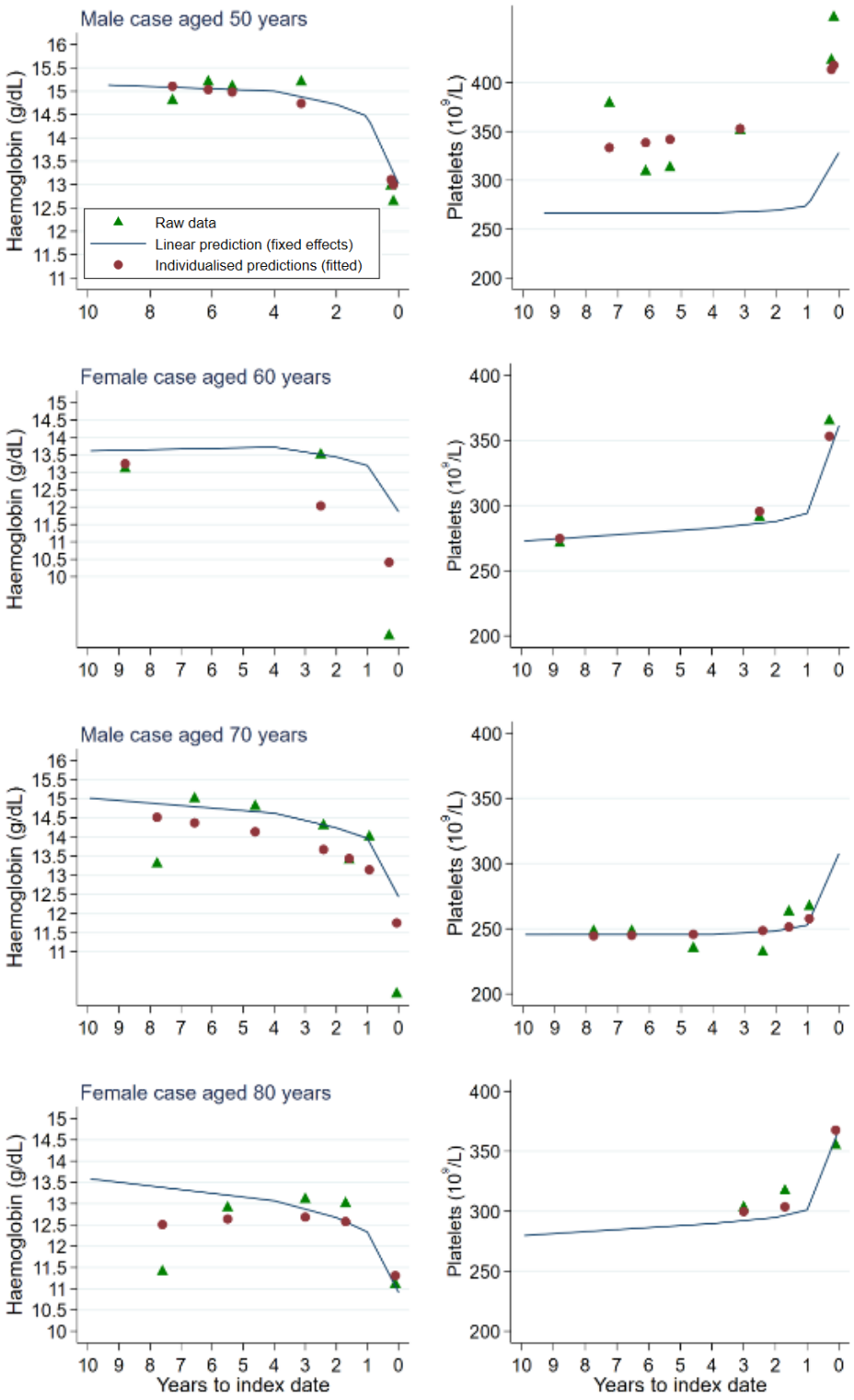
Figure C58: Trajectories (LOWESS and fixed-effects) in neutrophil count for males (left) and females (right) with and without bowel cancer by age at index date[1, 2]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

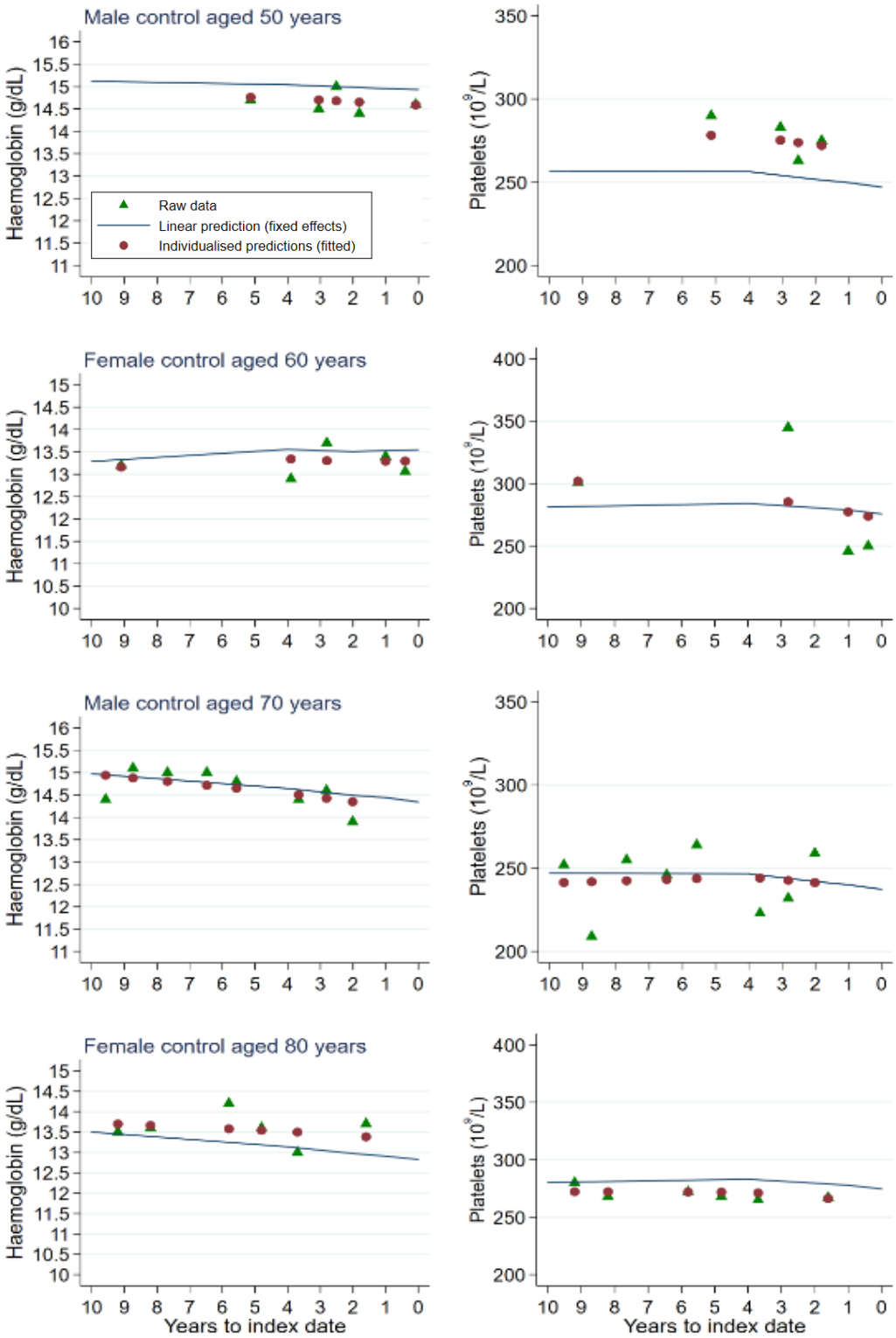
C8. Individualised trajectories

Figure C59: Individualised haemoglobin (left) and platelet (right) predicted levels for four cases at different ages at index date[1]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls

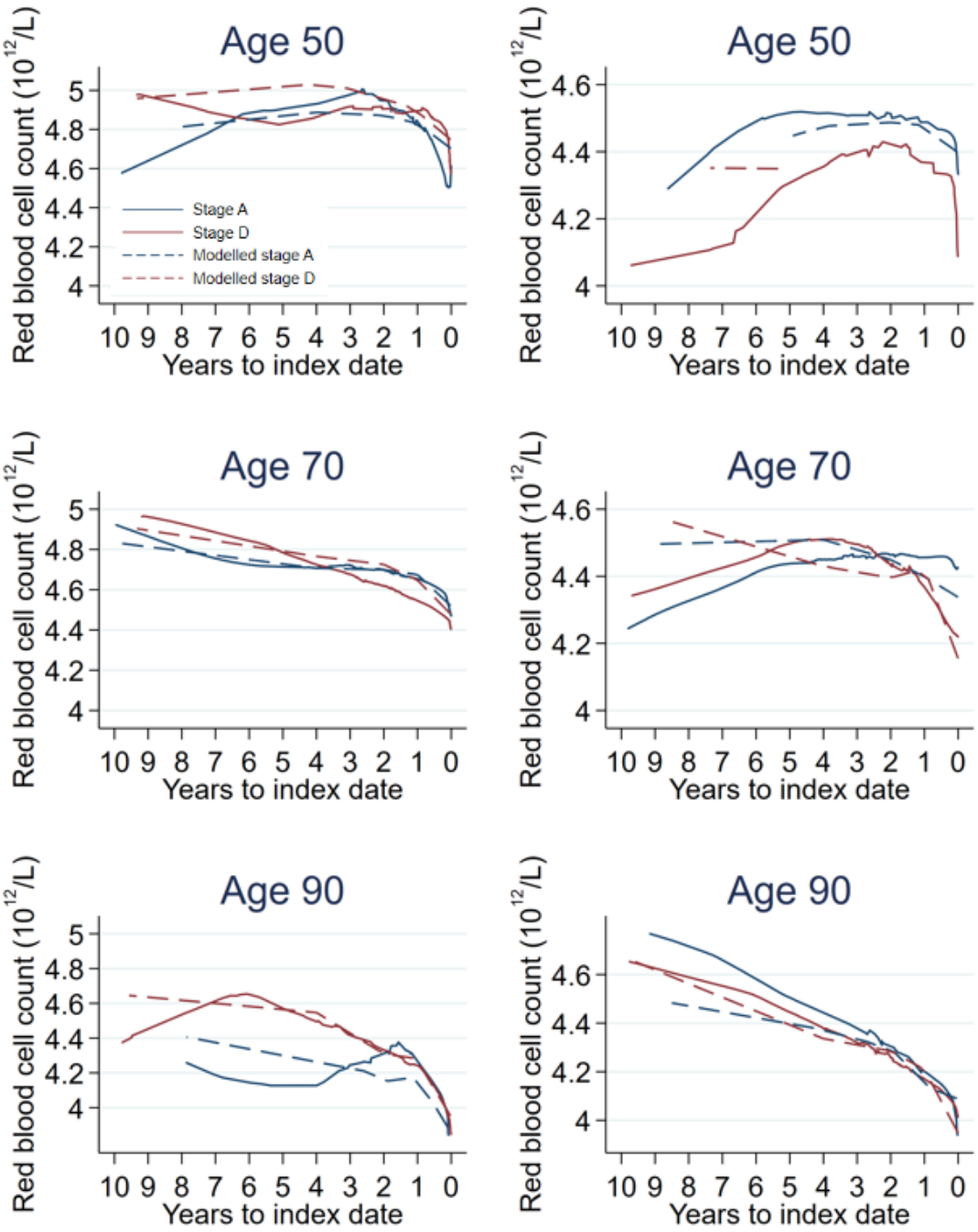
Figure C60: Individualised haemoglobin (left) and platelet (right) predicted levels for four controls at different ages at index date[1]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls

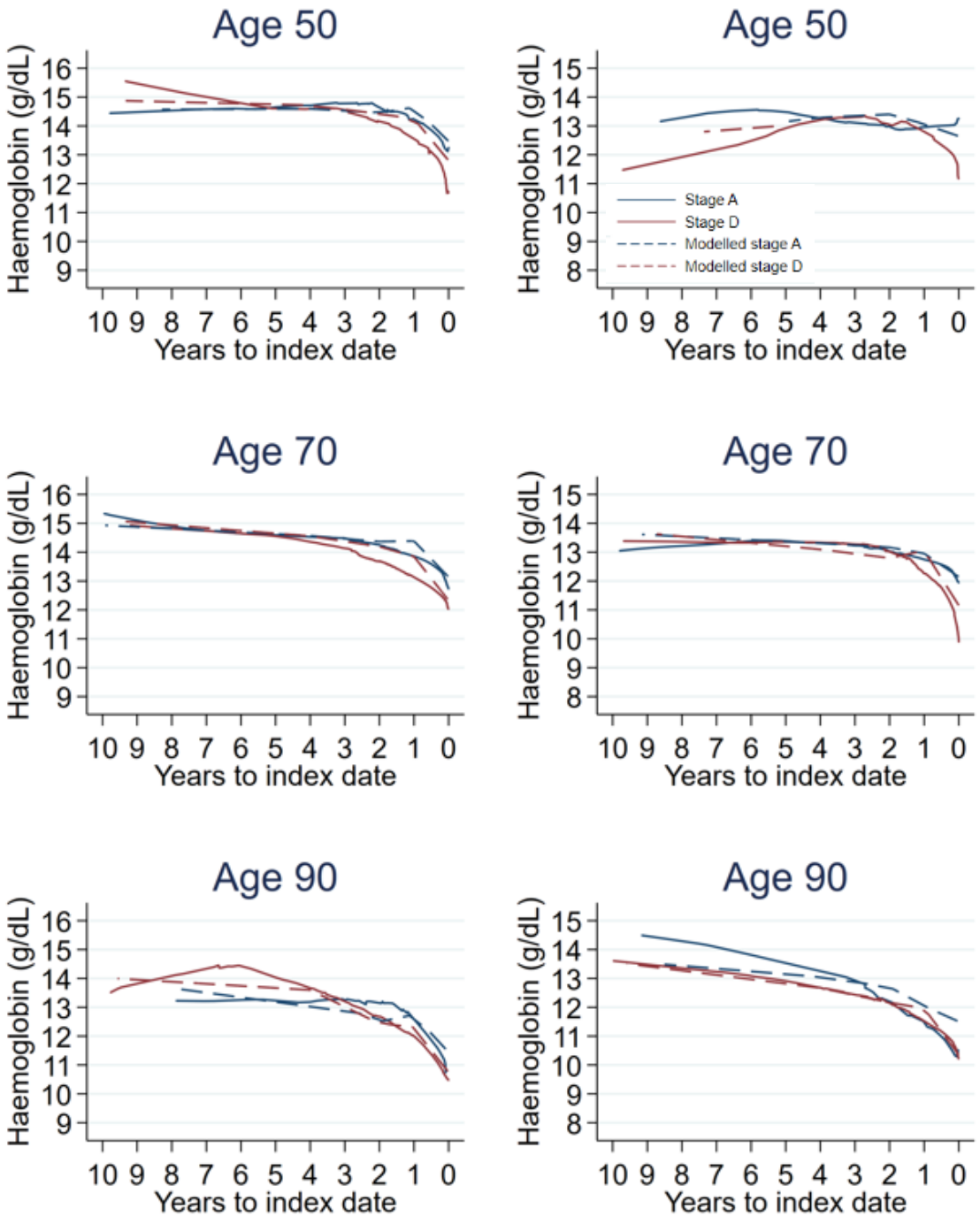
C9. Trajectories by Dukes' tumour stage

Figure C61: Trajectories (LOWESS and fixed-effects) in red blood cell count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

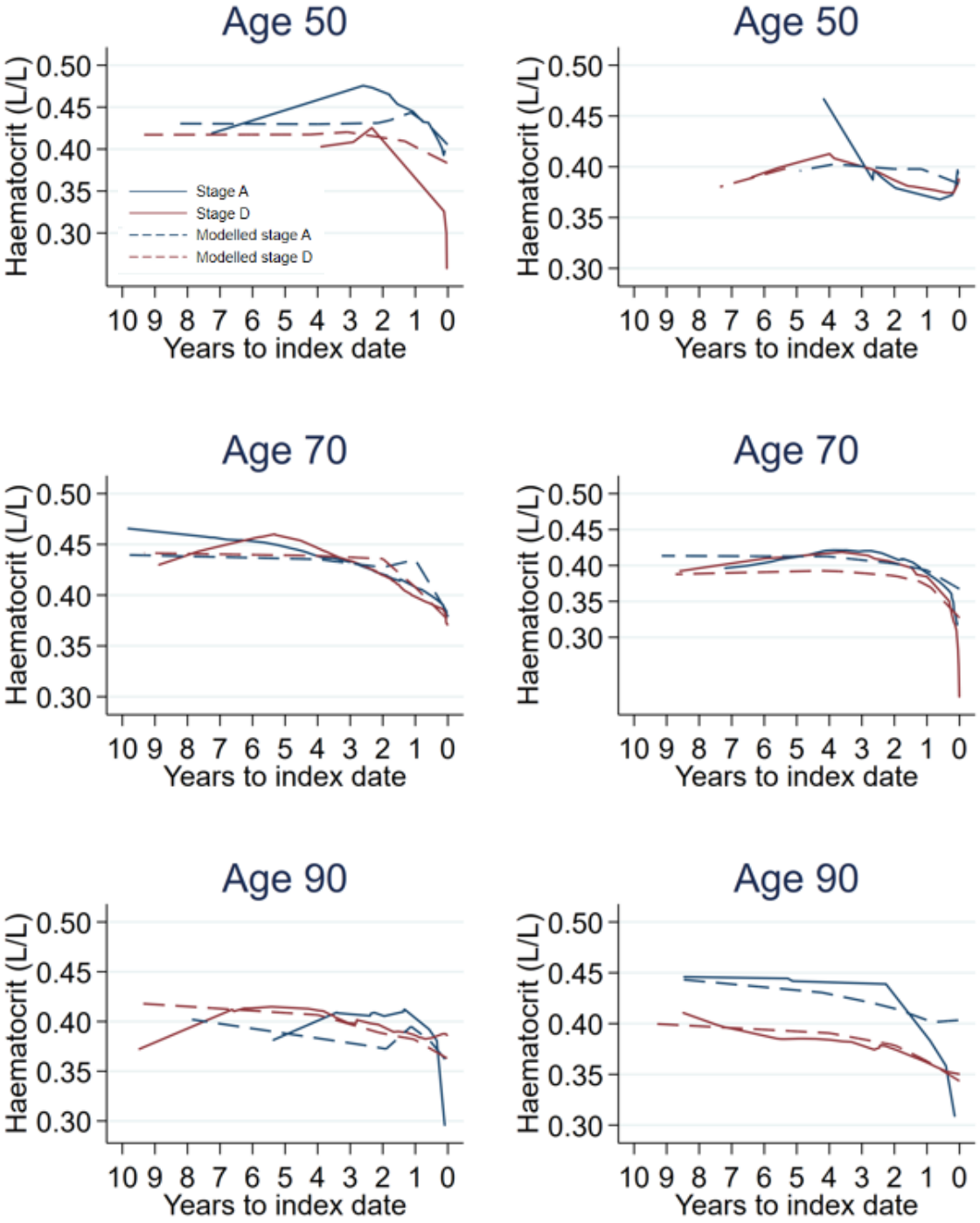
Figure C62: Trajectories (LOWESS and fixed-effects) in haemoglobin for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.

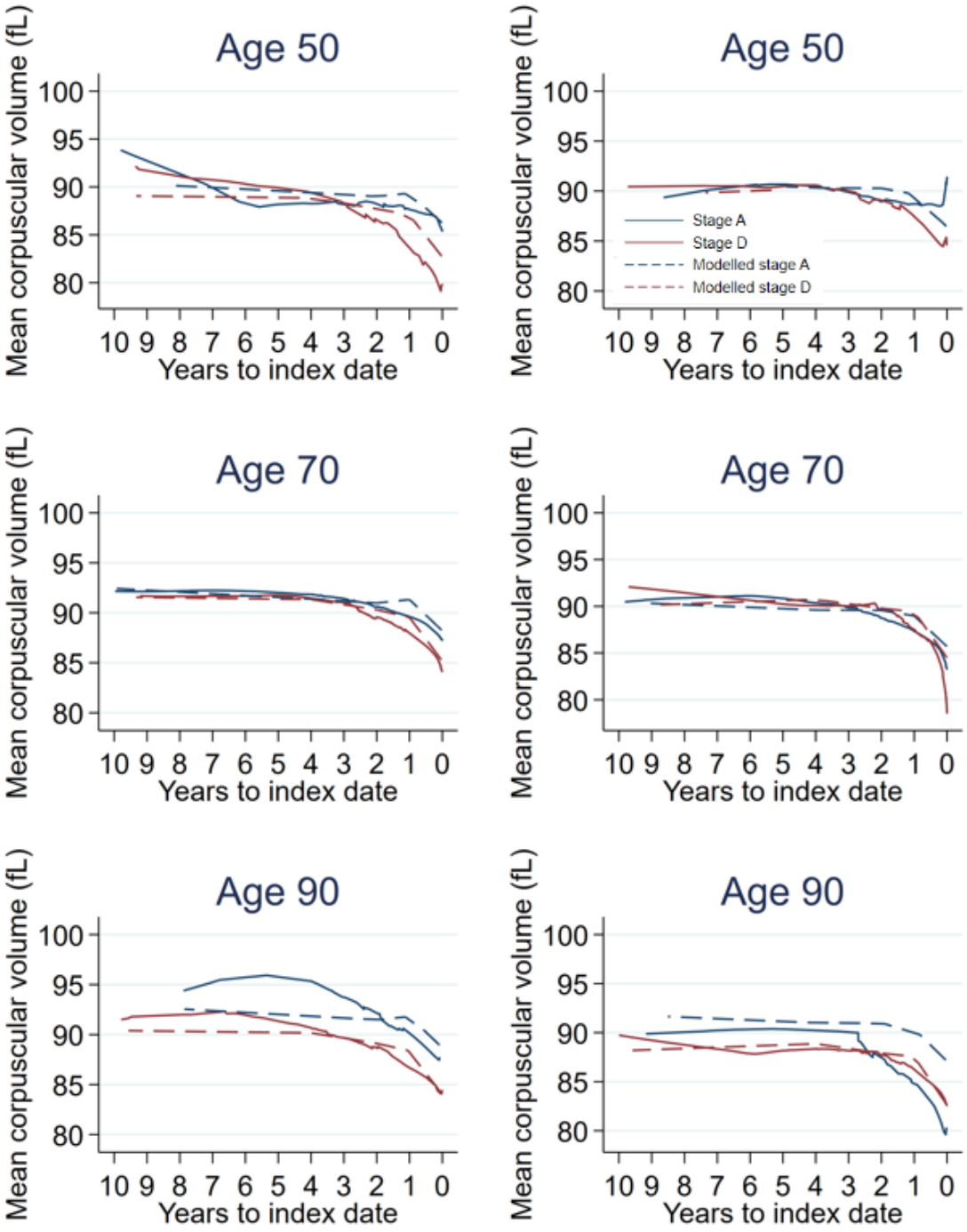
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C63: Trajectories (LOWESS and fixed-effects) in haematocrit for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



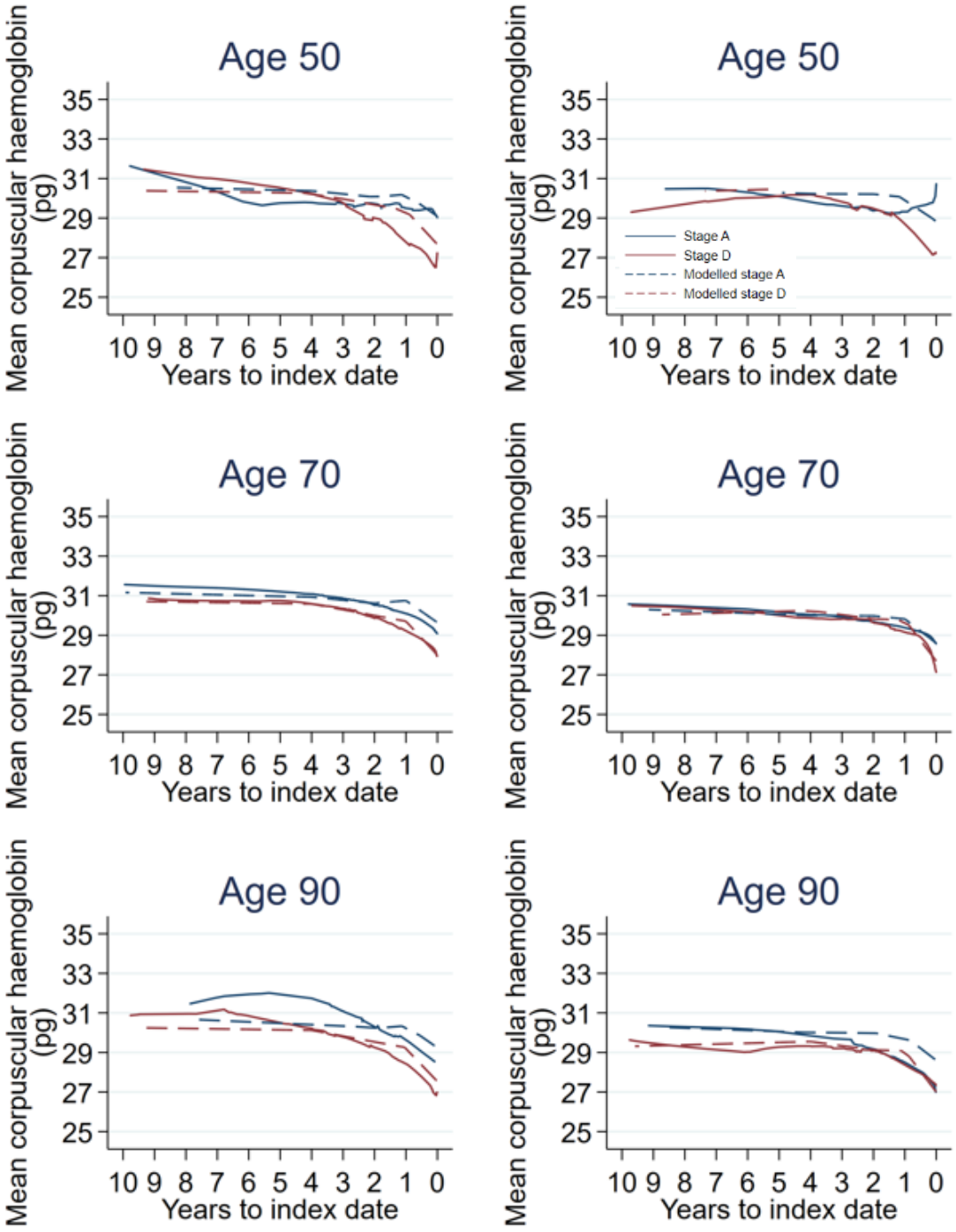
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C64: Trajectories (LOWESS and fixed-effects) in mean corpuscular volume for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



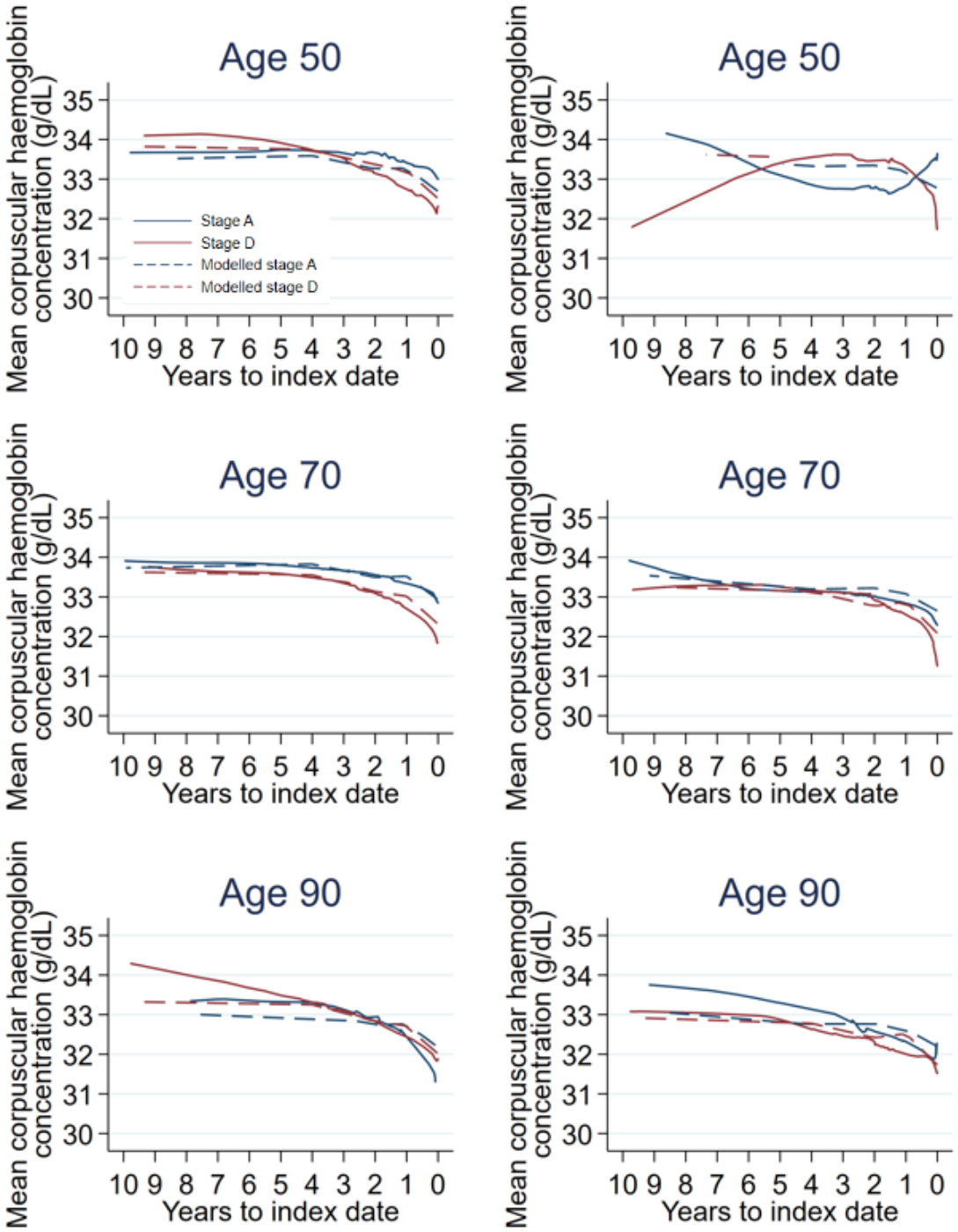
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C65: Trajectories (LOWESS and fixed-effects) in mean corpuscular haemoglobin for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



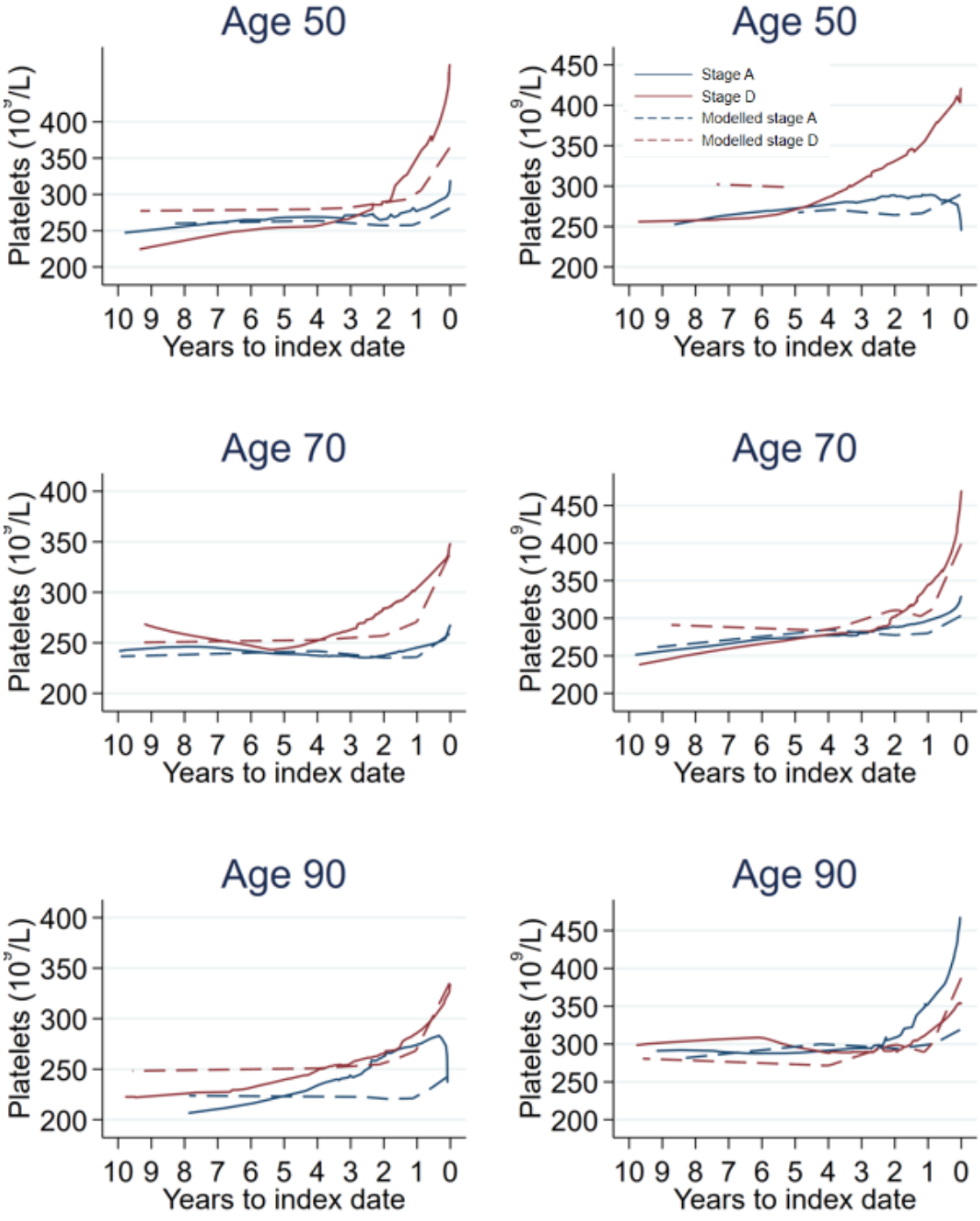
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C66: Trajectories (LOWESS and fixed-effects) in mean corpuscular haemoglobin concentration for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



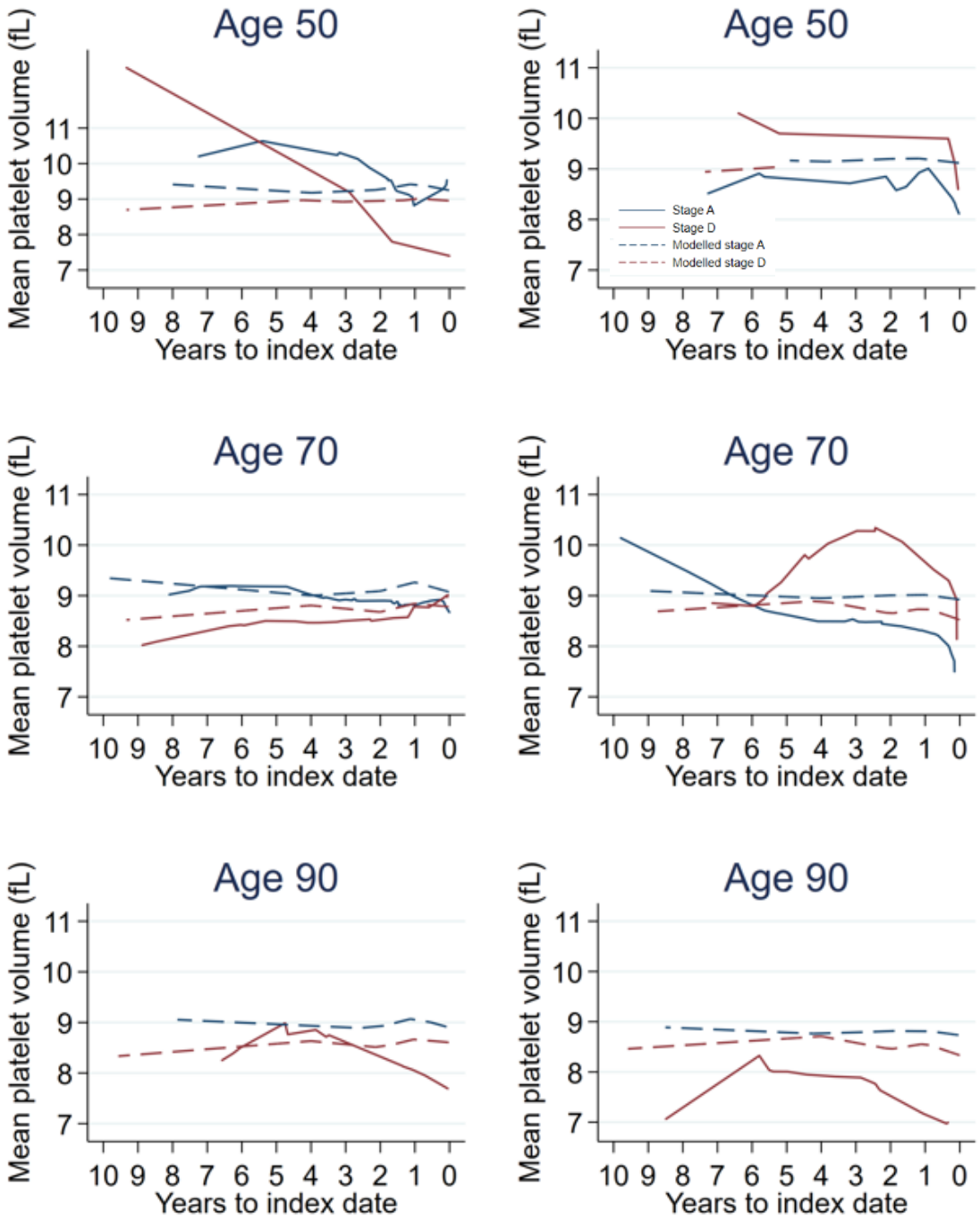
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C67: Trajectories (LOWESS and fixed-effects) in platelets for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



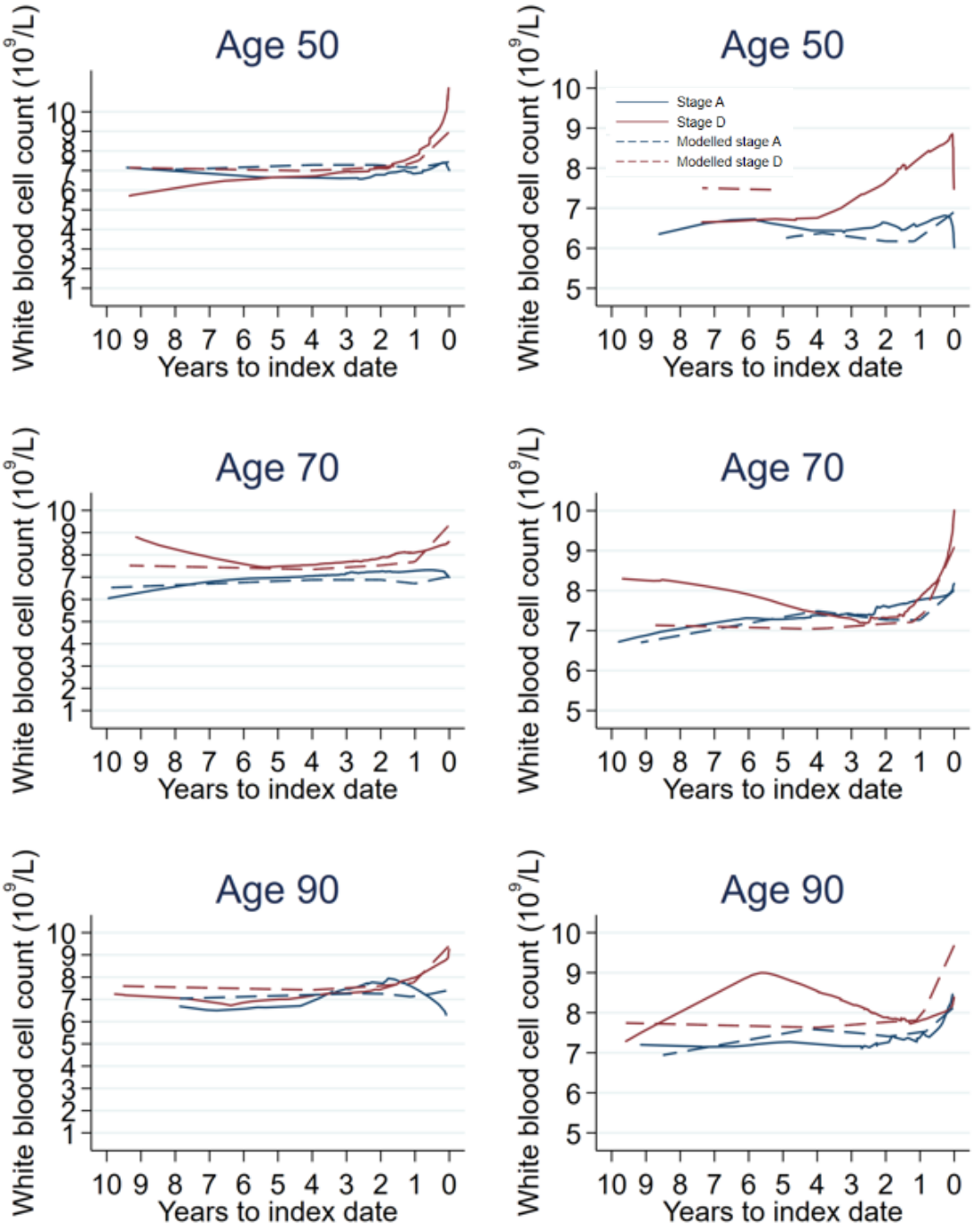
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C68: Trajectories (LOWESS and fixed-effects) in mean platelet volume for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



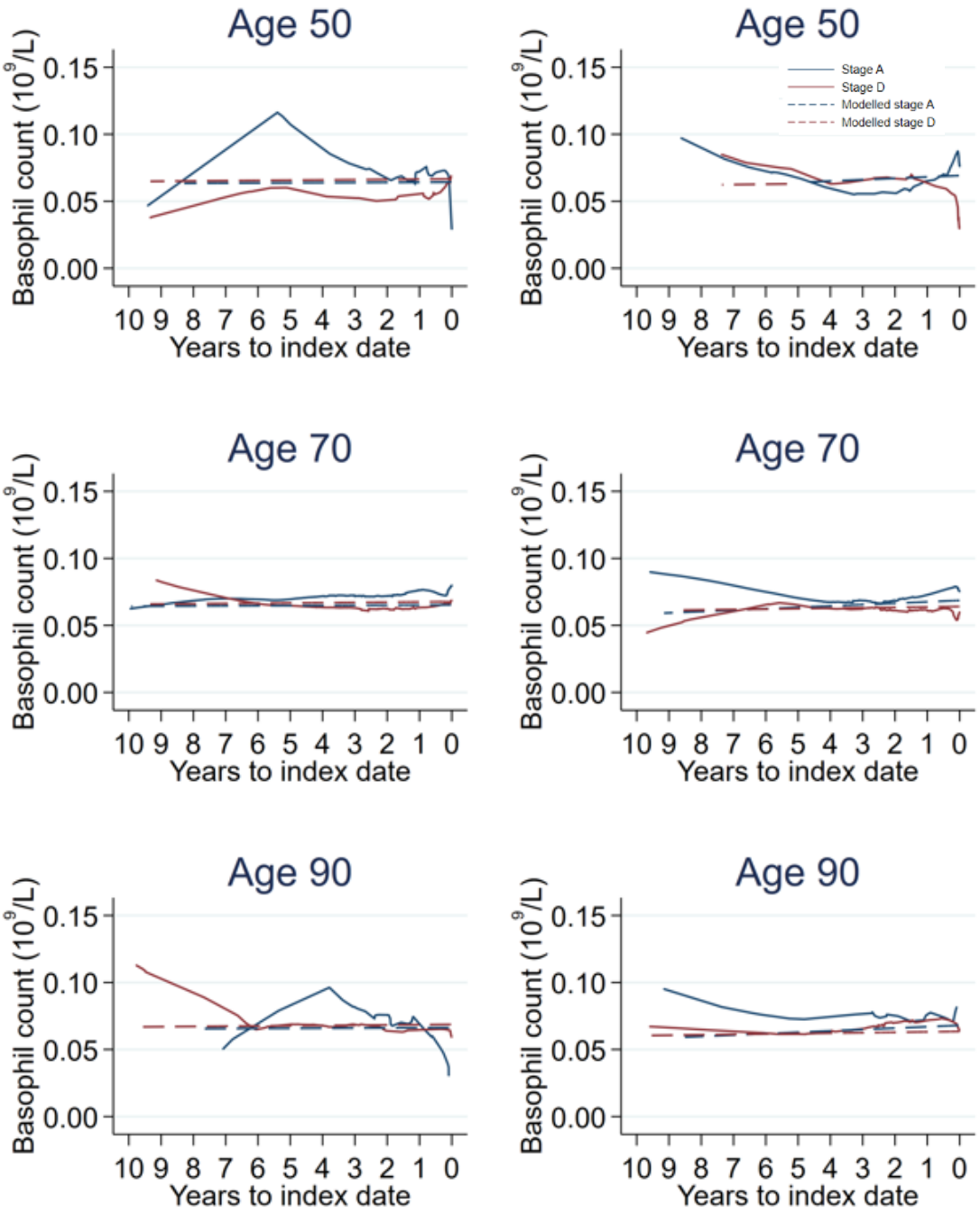
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C69: Trajectories (LOWESS and fixed-effects) in white blood cell count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



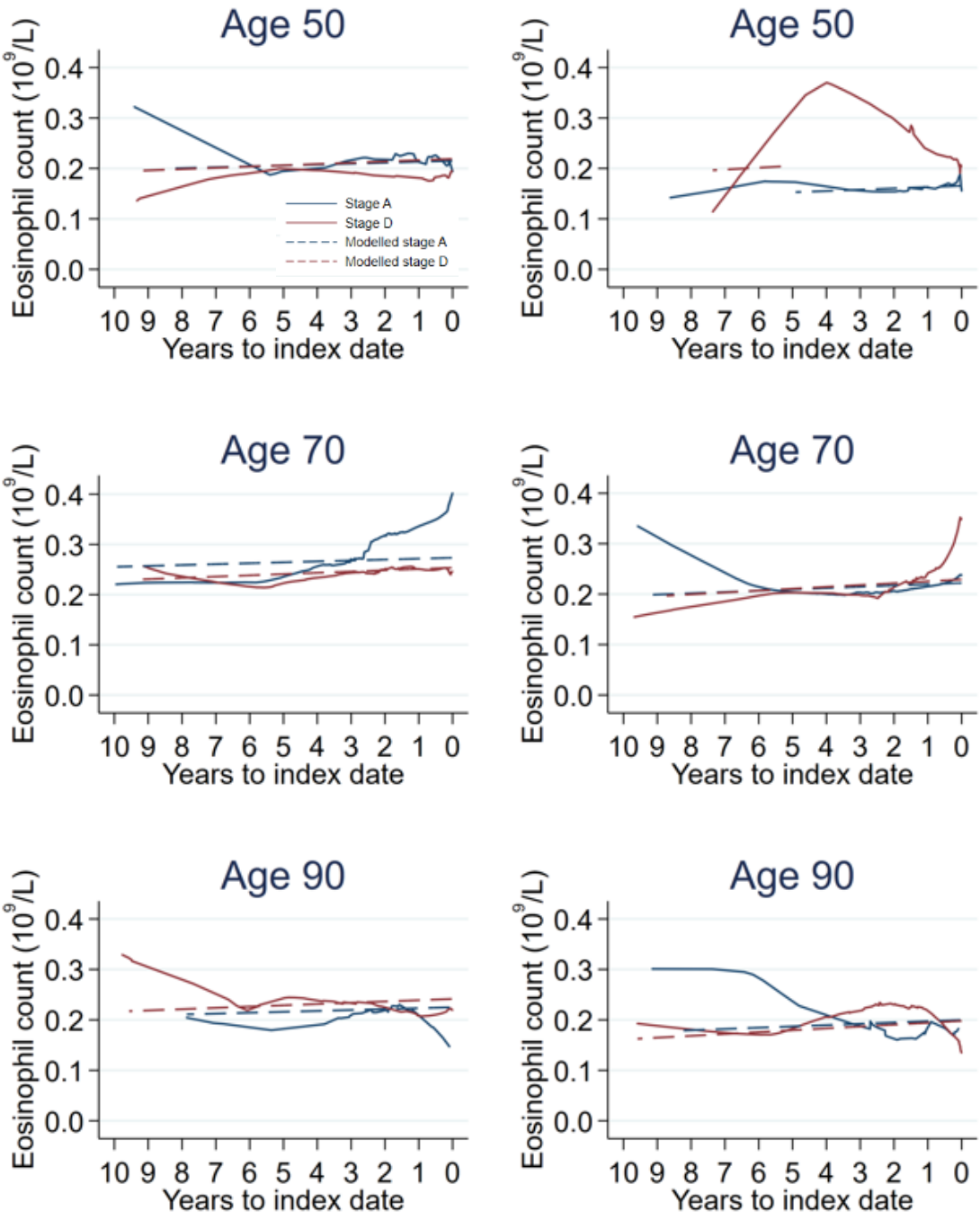
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C70: Trajectories (LOWESS and fixed-effects) in basophil count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



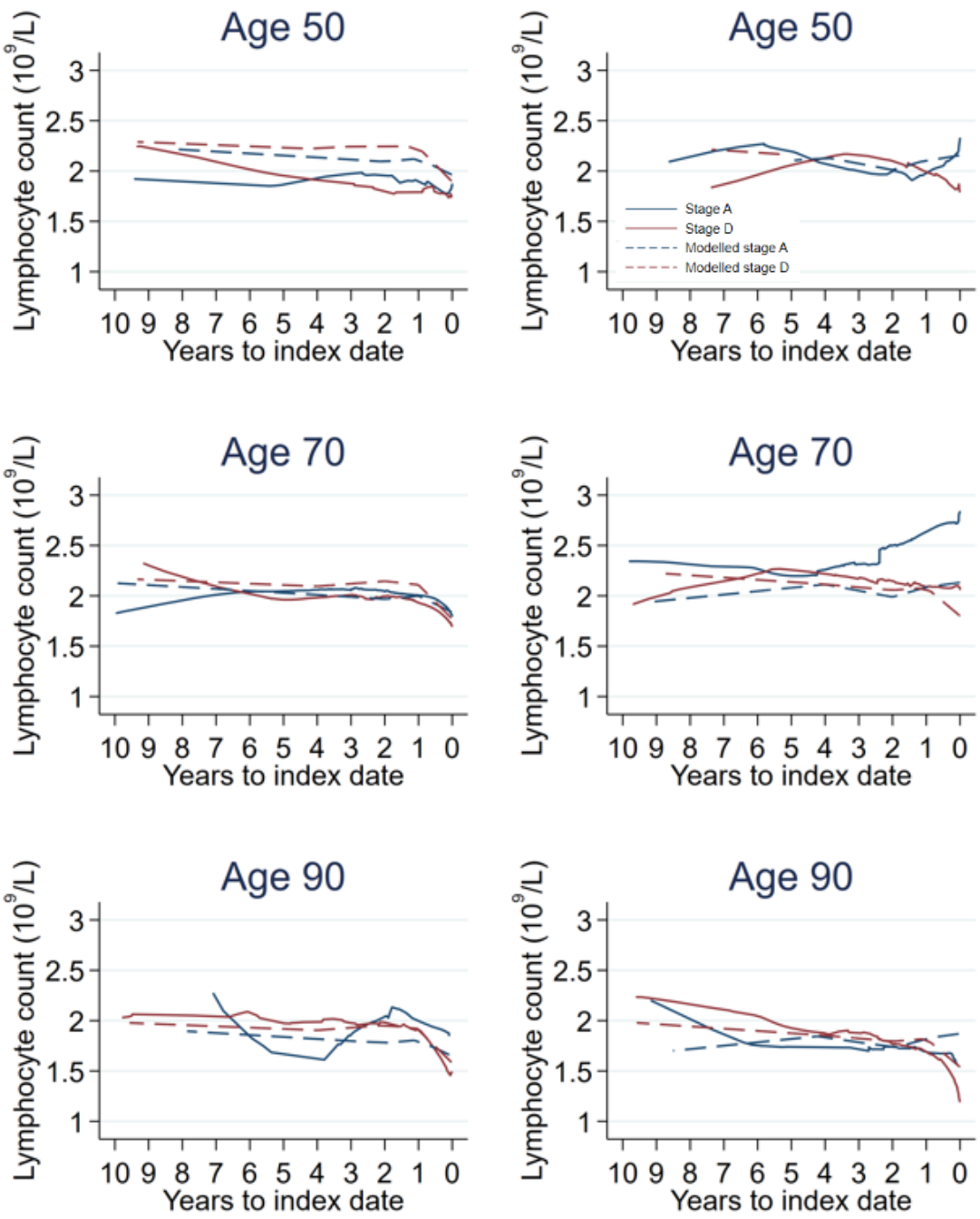
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C71: Trajectories (LOWESS and fixed-effects) in eosinophil count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



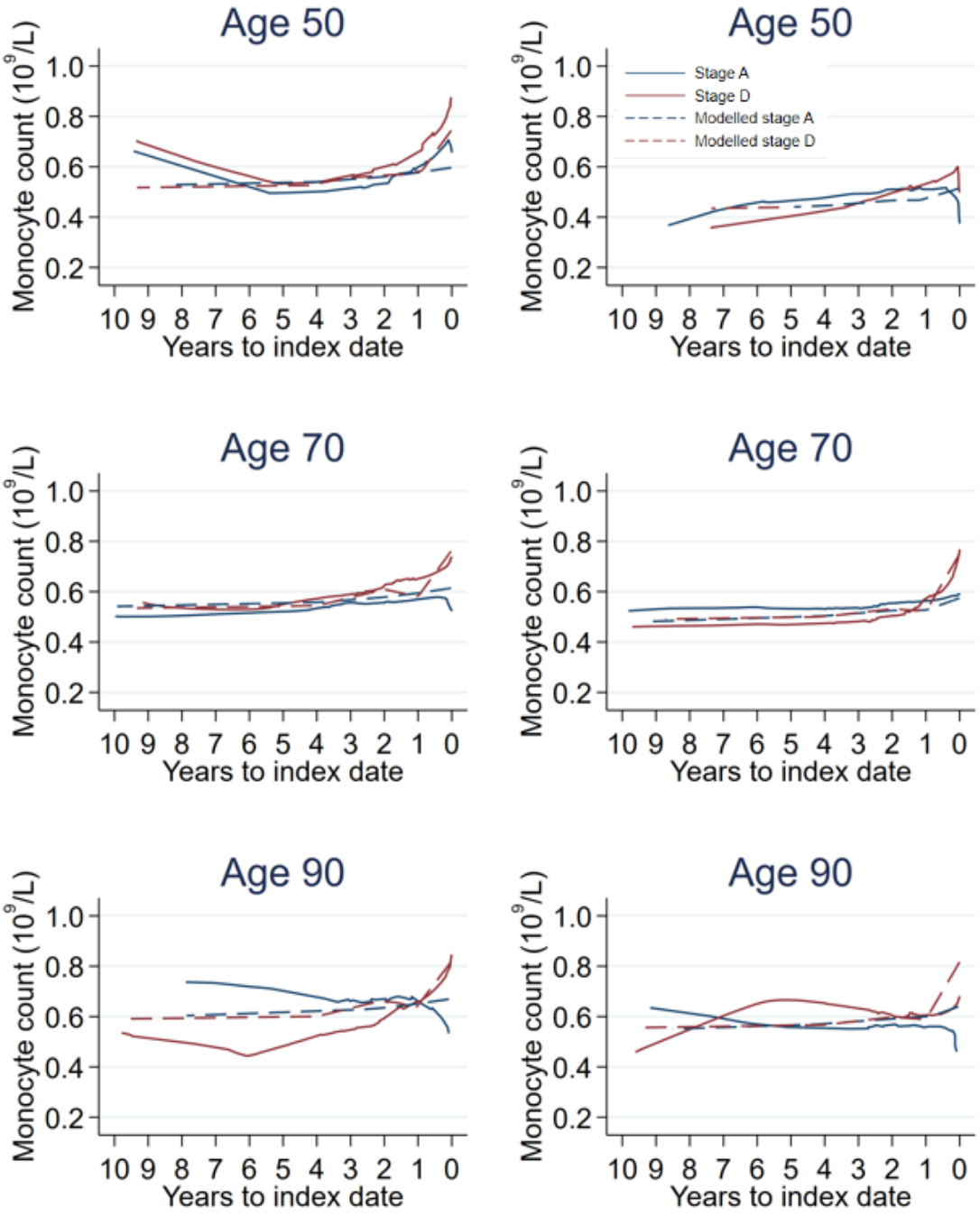
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C72: Trajectories (LOWESS and fixed-effects) in lymphocyte count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



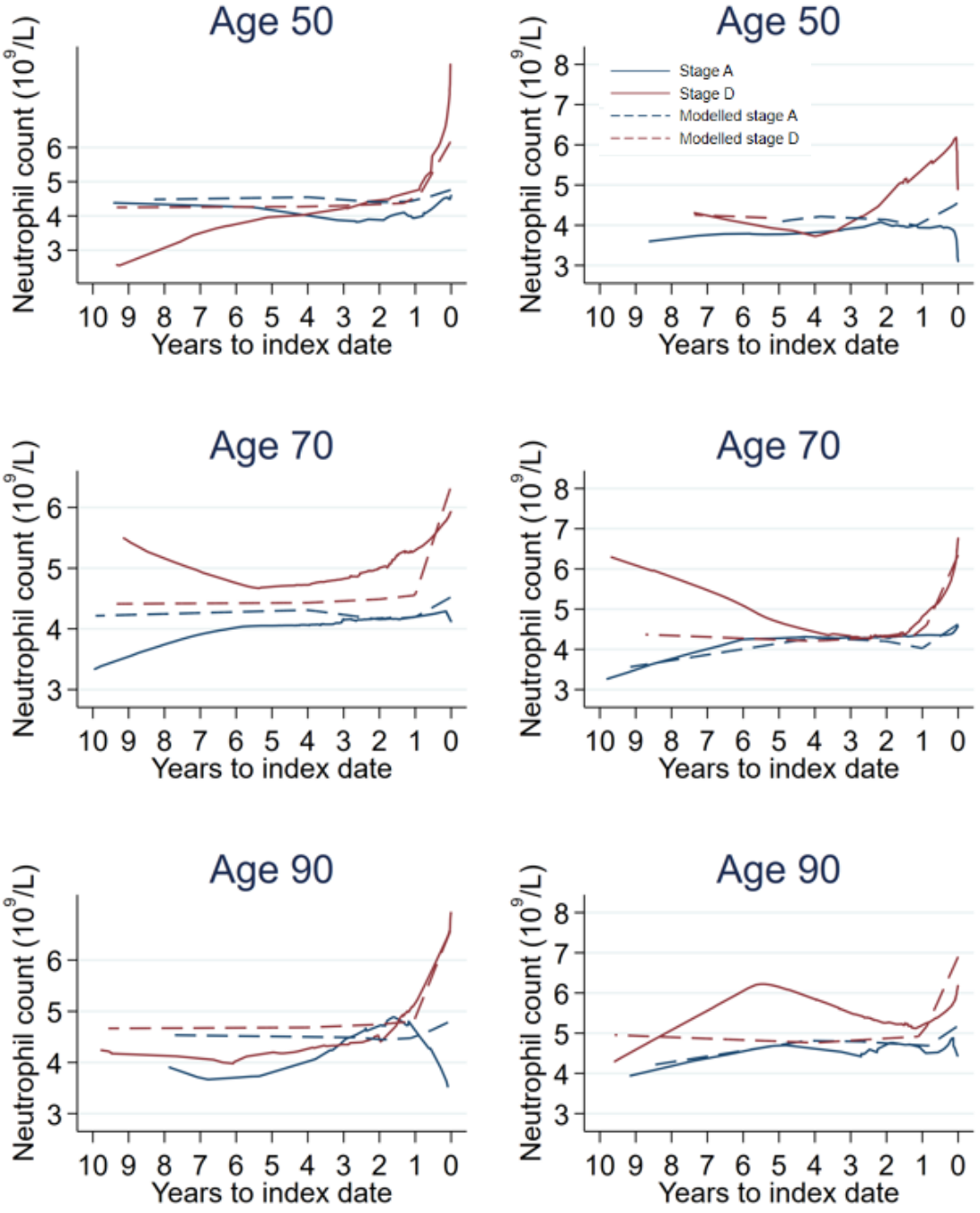
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C73: Trajectories (LOWESS and fixed-effects) in monocyte count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

Figure C74: Trajectories (LOWESS and fixed-effects) in neutrophil count for stage A vs. D tumour – males (left) and females (right)[1, 2, 3]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
[2]LOWESS trends are age (+/- 3 years) at index and modelled (fixed-effects) trends are estimated at that specific age.

C10. Mixed models for Dukes' tumour stage

Table C8: Mixed-effect model coefficients (95% CI) to assess red blood cell parameters by Dukes' Stage A vs. D

Variable	Males						Females					
	RBC	**Hb**	**Hc**	**MCV**	**MCH**	**MCHC**	**RBC**	**Hb**	**Hc**	**MCV**	**MCH**	**MCHC**
N	1515	1571	423	1551	1503	1427	1155	1213	332	1194	1134	1100
n cases	725	736	214	711	706	661	499	529	152	522	489	469
Fixed effects:												
Constant	4.4035 (3.5802, 5.2267)	9.9164 (6.8338, 12.9990)	0.4394 (0.2527, 0.6261)	76.8717 (67.8548, 85.8886)	26.1164 (22.8572, 29.3757)	30.6694 (28.9221, 32.4166)	4.6958 (3.7582, 5.6334)	12.6106 (9.0741, 16.1471)	0.4568 (0.2469, 0.6667)	77.7683 (66.4981, 89.0385)	26.8108 (22.9838, 30.6379)	33.4073 (31.2499, 35.5647)
Age at index date[1] (years)	0.0059 (-0.0088, 0.0206)	0.0674 (0.0122, 0.1225)	-0.0006 (-0.0040, 0.0027)	0.1869 (0.0271, 0.3467)	0.0597 (0.0019, 0.1174)	0.0403 (0.0094, 0.0713)	-0.0054 (-0.0221, 0.0114)	0.0046 (-0.0586, 0.0679)	-0.0012 (-0.0050, 0.0026)	0.1753 (-0.0262, 0.3769)	0.0408 (-0.0275, 0.1092)	-0.0123 (-0.0508, 0.0262)
Age at index date[1] – knot at 60 (years)	-0.0341 (-0.0595, -0.0088)	-0.1571 (-0.2520, -0.0621)	-0.0019 (-0.0077, 0.0038)	-0.1782 (-0.4420, 0.0856)	-0.0656 (-0.1608, 0.0296)	-0.0575 (-0.1081, -0.0069)	0.0065 (-0.0224, 0.0355)	-0.0601 (-0.1692, 0.0491)	-0.0013 (-0.0078, 0.0053)	-0.4503 (-0.7993, -0.1013)	-0.1080 (-0.2259, 0.0098)	0.0091 (-0.0563, 0.0745)

Age at index date[1] – knot at 70 (years)	0.0255 (0.0007, 0.0503)	0.0398 (-0.0524, 0.1321)	0.0030 (-0.0024, 0.0084)	-0.1188 (-0.3732, 0.1356)	-0.0565 (-0.1482, 0.0351)	-0.0092 (-0.0575, 0.0392)	-0.0251 (-0.0524, 0.0022)	-0.0546 (-0.1570, 0.0479)	0.0014 (-0.0049, 0.0076)	0.2975 (-0.0389, 0.6338)	0.0211 (-0.0926, 0.1348)	-0.0229 (-0.0836, 0.0378)
Age at index date[1] – knot at 80 (years)	-0.0526 (-0.0852, -0.0200)	-0.0548 (-0.1697, 0.0601)	-0.0011 (-0.0071, 0.0048)	0.2693 (-0.0910, 0.6297)	0.0847 (-0.0445, 0.2139)	-0.0229 (-0.0930, 0.0471)	0.0205 (-0.0085, 0.0495)	0.1426 (0.0365, 0.2488)	0.0070 (0.0002, 0.0138)	0.0828 (-0.2991, 0.4647)	0.0892 (-0.0411, 0.2195)	0.0060 (-0.0618, 0.0739)
Time to index date[1] (years)	0.3864 (-0.5294, 1.3021)	5.3873 (1.7396, 9.0351)	0.0942 (-0.1498, 0.3382)	3.1593 (2.5114, 3.8072)	1.1156 (0.8740, 1.3571)	0.6050 (0.4330, 0.7770)	-0.3456 (-1.4876, 0.7965)	-2.5994 (-6.8355, 1.6367)	-0.2053 (-0.5182, 0.1076)	3.3268 (2.5960, 4.0576)	1.2300 (0.9639, 1.4962)	0.4343 (0.2523, 0.6162)
Time to index date[1] date – knot at 1 (years)	-0.2071 (-1.8449, 1.4307)	-7.4027 (-13.8704, -0.9350)	-0.1852 (-0.5750, 0.2045)	-3.5320 (-4.6623, -2.4018)	-1.2569 (-1.6765, -0.8373)	-0.6380 (-0.9382, -0.3377)	0.1585 (-2.0315, 2.3485)	3.2106 (-4.7104, 11.1316)	0.2821 (-0.2574, 0.8216)	-2.7413 (-3.9939, -1.4888)	-1.0680 (-1.5217, -0.6143)	-0.2886 (-0.5989, 0.0217)
Time to index date[1] date – knot at 2 (years)	-0.2240 (-1.4605, 1.0125)	1.4891 (-3.3180, 6.2961)	0.0021 (-0.3250, 0.3292)	0.5712 (-0.2145, 1.3568)	0.2999 (0.0082, 0.5916)	0.2018 (-0.0068, 0.4104)	0.0743 (-1.6653, 1.8139)	-1.0014 (-7.1529, 5.1501)	-0.0790 (-0.4257, 0.2677)	-0.5796 (-1.4373, 0.2781)	-0.1513 (-0.4610, 0.1583)	-0.1554 (-0.3666, 0.0558)
Time to index date[1] date – knot at 4 (years)	-0.1963 (-0.8706, 0.4779)	0.2407 (-2.4176, 2.8990)	0.0804 (-0.2513, 0.4120)	-0.0318 (-0.3913, 0.3277)	-0.1184 (-0.2531, 0.0163)	-0.1839 (-0.2781, -0.0897)	0.1521 (-0.7876, 1.0918)	0.1373 (-3.2328, 3.5075)	-0.0025 (-0.3025, 0.2975)	0.1417 (-0.2589, 0.5423)	0.0508 (-0.0943, 0.1960)	0.0769 (-0.0214, 0.1751)

Dukes' D (vs. A)	1.2082 (0.1845, 2.2319)	-0.3514 (-3.8808, 3.1780)	-0.0991 (-0.3154, 0.1172)	-13.3449 (-27.1015, 0.4117)	-5.5770 (-10.5296, -0.6245)	-0.6897 (-3.3618, 1.9824)	-0.2849 (-1.3157, 0.7459)	0.6777 (-2.9215, 4.2769)	0.0177 (-0.1976, 0.2331)	3.1397 (-12.8396, 19.1190)	0.3363 (-5.1398, 5.8123)	-0.2814 (-3.3824, 2.8197)
Dukes' D (vs. A) by time to index date[1] interaction	0.0133 (-0.0550, 0.0817)	0.2794 (0.0148, 0.5441)	-0.0214 (-0.0367, -0.0060)	1.2163 (0.3164, 2.1162)	0.6000 (0.2613, 0.9388)	0.0991 (-0.1409, 0.3391)	0.2006 (0.1239, 0.2773)	1.0733 (0.7859, 1.3608)	0.0245 (0.0065, 0.0425)	1.6564 (0.5941, 2.7188)	0.8505 (0.4597, 1.2413)	0.3606 (0.0935, 0.6278)
Dukes' D (vs. A) by time to index date[1] (knot at 1) interaction	0.0410 (-0.0818, 0.1639)	0.0667 (-0.4056, 0.5391)	0.0558 (0.0280, 0.0835)	-0.1100 (-1.7108, 1.4908)	-0.1705 (-0.7691, 0.4280)	0.0550 (-0.3734, 0.4834)	-0.2790 (-0.4140, -0.1440)	-1.5478 (-2.0524, -1.0433)	-0.0226 (-0.0542, 0.0089)	-1.9403 (-3.8032, -0.0775)	-0.9683 (-1.6478, -0.2887)	-0.6136 (-1.0834, -0.1438)
Dukes' D (vs. A) by time to index date[1] (knot at 2) interaction	-0.0298 (-0.1197, 0.0601)	-0.2409 (-0.5837, 0.1019)	-0.0352 (-0.0562, -0.0141)	-0.7959 (-1.9493, 0.3576)	-0.2955 (-0.7248, 0.1337)	-0.1125 (-0.4221, 0.1970)	0.0709 (-0.0236, 0.1654)	0.5693 (0.2189, 0.9197)	-0.0038 (-0.0256, 0.0180)	0.7481 (-0.5538, 2.0499)	0.3231 (-0.1483, 0.7946)	0.4441 (0.1153, 0.7729)
Dukes' D (vs. A) by time to index date[1] (knot at 4) interaction	-0.0231 (-0.0664, 0.0201)	-0.0772 (-0.2428, 0.0883)	-0.0001 (-0.0118, 0.0117)	-0.4362 (-0.9805, 0.1081)	-0.1503 (-0.3543, 0.0537)	-0.0132 (-0.1575, 0.1312)	0.0386 (-0.0053, 0.0824)	-0.0479 (-0.2079, 0.1122)	0.0005 (-0.0109, 0.0118)	-0.7502 (-1.3526, -0.1477)	-0.3024 (-0.5215, -0.0832)	-0.2300 (-0.3835, -0.0765)
Time by age interaction	-0.0052 (-0.0215, 0.0111)	-0.0768 (-0.1417, -0.0118)	-0.0010 (-0.0053, 0.0033)				0.0082 (-0.0119, 0.0284)	0.0549 (-0.0201, 0.1299)	0.0040 (-0.0015, 0.0096)			

Time (knot at 1) by age interaction	0.0028 (-0.0264, 0.0320)	0.1098 (-0.0053, 0.2250)	0.0025 (-0.0044, 0.0094)				-0.0053 (-0.0437, 0.0330)	-0.0611 (-0.2004, 0.0783)	-0.0057 (-0.0153, 0.0039)			
Time (knot at 2) by age interaction	0.0035 (-0.0185, 0.0256)	-0.0207 (-0.1062, 0.0647)	0.0001 (-0.0056, 0.0059)				-0.0004 (-0.0307, 0.0299)	0.0152 (-0.0924, 0.1227)	0.0018 (-0.0044, 0.0080)			
Time (knot at 4) by age interaction	0.0029 (-0.0091, 0.0148)	-0.0070 (-0.0540, 0.0399)	-0.0014 (-0.0071, 0.0043)				-0.0042 (-0.0206, 0.0123)	-0.0081 (-0.0673, 0.0510)	-0.0002 (-0.0054, 0.0050)			
Time by age (knot at 60) interaction	0.0215 (-0.0059, 0.0488)	0.1336 (0.0256, 0.2415)	0.0042 (-0.0025, 0.0108)				-0.0215 (-0.0539, 0.0110)	-0.0584 (-0.1808, 0.0639)	-0.0053 (-0.0145, 0.0040)			
Time (knot at 1) by age (knot at 60) interaction	-0.0229 (-0.0721, 0.0263)	-0.1587 (-0.3506, 0.0332)	-0.0069 (-0.0181, 0.0044)				0.0241 (-0.0348, 0.0829)	0.0458 (-0.1729, 0.2646)	0.0109 (-0.0053, 0.0270)			
Time (knot at 2) by age (knot at 60) interaction	-0.0038 (-0.0403, 0.0326)	-0.0063 (-0.1461, 0.1335)	0.0009 (-0.0081, 0.0100)				-0.0070 (-0.0513, 0.0373)	-0.0004 (-0.1620, 0.1613)	-0.0058 (-0.0164, 0.0049)			

Time (knot at 4) by age (knot at 60) interaction	0.0049 (-0.0136, 0.0234)	0.0310 (-0.0411, 0.1030)	0.0010 (-0.0064, 0.0085)				0.0126 (-0.0114, 0.0365)	0.0428 (-0.0447, 0.1304)	0.0006 (-0.0065, 0.0077)			
Time by age (knot at 70) interaction	-0.0288 (-0.0534, -0.0042)	-0.0989 (-0.1946, -0.0032)	-0.0062 (-0.0120, -0.0005)				0.0266 (-0.0011, 0.0542)	0.0572 (-0.0477, 0.1621)	0.0012 (-0.0063, 0.0087)			
Time (knot at 1) by age (knot at 70) interaction	0.0374 (-0.0065, 0.0814)	0.1150 (-0.0548, 0.2849)	0.0086 (-0.0019, 0.0191)				-0.0204 (-0.0681, 0.0273)	0.0328 (-0.1478, 0.2133)	-0.0055 (-0.0185, 0.0075)			
Time (knot at 2) by age (knot at 70) interaction	0.0040 (-0.0271, 0.0351)	0.0200 (-0.0989, 0.1389)	-0.0022 (-0.0100, 0.0055)				-0.0038 (-0.0367, 0.0292)	-0.0902 (-0.2144, 0.0341)	0.0033 (-0.0057, 0.0122)			
Time (knot at 4) by age (knot at 70) interaction	-0.0196 (-0.0338, -0.0053)	-0.0401 (-0.0950, 0.0149)	0.0006 (-0.0037, 0.0049)				-0.0115 (-0.0273, 0.0042)	-0.0464 (-0.1054, 0.0125)	0.0013 (-0.0030, 0.0056)			
Time by age (knot at 80) interaction	0.0345 (0.0115, 0.0575)	0.0858 (-0.0041, 0.1756)	0.0027 (-0.0020, 0.0074)				-0.0184 (-0.0432, 0.0064)	-0.1069 (-0.2011, -0.0128)	-0.0036 (-0.0095, 0.0023)			

Time (knot at 1) by age (knot at 80) interaction	-0.0484 (-0.0889, -0.0078)	-0.1608 (-0.3179, -0.0037)	-0.0063 (-0.0148, 0.0021)				0.0009 (-0.0425, 0.0444)	0.0114 (-0.1529, 0.1756)	0.0034 (-0.0069, 0.0138)			
Time (knot at 2) by age (knot at 80) interaction	0.0031 (-0.0270, 0.0332)	0.0706 (-0.0441, 0.1853)	0.0034 (-0.0031, 0.0099)				0.0158 (-0.0142, 0.0459)	0.0992 (-0.0132, 0.2117)	0.0025 (-0.0047, 0.0097)			
Time (knot at 4) by age (knot at 80) interaction	0.0161 (0.0008, 0.0313)	0.0029 (-0.0560, 0.0619)	0.0003 (-0.0036, 0.0042)				0.0097 (-0.0035, 0.0228)	0.0357 (-0.0130, 0.0843)	-0.0027 (-0.0064, 0.0010)			
Dukes' D (vs. A) by age interaction	-0.0232 (-0.0415, -0.0050)	-0.0068 (-0.0696, 0.0559)	0.0015 (-0.0023, 0.0053)	0.1972 (-0.0471, 0.4414)	0.0831 (-0.0048, 0.1710)	0.0103 (-0.0370, 0.0577)	0.0010 (-0.0175, 0.0196)	-0.0334 (-0.0980, 0.0312)	-0.0006 (-0.0044, 0.0032)	-0.1143 (-0.4021, 0.1736)	-0.0245 (-0.1233, 0.0743)	-0.0005 (-0.0563, 0.0554)
Dukes' D (vs. A) by age (knot at 60) interaction	0.0375 (0.0065, 0.0685)	0.0010 (-0.1040, 0.1061)	-0.0011 (-0.0074, 0.0052)	-0.3402 (-0.7459, 0.0654)	-0.1877 (-0.3337, -0.0417)	-0.0645 (-0.1428, 0.0137)	0.0027 (-0.0308, 0.0363)	0.0653 (-0.0498, 0.1805)	-0.0016 (-0.0082, 0.0049)	0.3876 (-0.1364, 0.9116)	0.0511 (-0.1294, 0.2316)	-0.0243 (-0.1242, 0.0755)
Dukes' D (vs. A) by age (knot at 70) interaction	-0.0200 (-0.0502, 0.0103)	0.0313 (-0.0685, 0.1311)	-0.0010 (-0.0068, 0.0047)	0.3235 (-0.0607, 0.7077)	0.2145 (0.0766, 0.3524)	0.0881 (0.0146, 0.1615)	-0.0003 (-0.0340, 0.0334)	-0.0143 (-0.1279, 0.0993)	0.0047 (-0.0018, 0.0112)	-0.4225 (-0.9505, 0.1055)	-0.0322 (-0.2135, 0.1490)	0.0490 (-0.0481, 0.1462)

Dukes' D (vs. A) by age (knot at 80) interaction	0.0258 (-0.0112, 0.0628)	-0.0375 (-0.1562, 0.0812)	0.0024 (-0.0039, 0.0088)	-0.5334 (-0.9972, -0.0697)	-0.2267 (-0.3930, -0.0604)	-0.0222 (-0.1116, 0.0672)	-0.0032 (-0.0364, 0.0301)	-0.0517 (-0.1629, 0.0594)	-0.0071 (-0.0138, -0.0004)	-0.0146 (-0.5356, 0.5064)	-0.0571 (-0.2347, 0.1205)	-0.0383 (-0.1326, 0.0559)
Random effects:												
Intercept for patient (variance)	0.1923 (0.1756, 0.2107)	2.9902 (2.7319, 3.2730)	0.0023 (0.0019, 0.0027)	45.3333 (41.6183, 49.3800)	5.3157 (4.8602, 5.8138)	1.3501 (1.2131, 1.5027)	0.1515 (0.1361, 0.1685)	2.4850 (2.2428, 2.7534)	0.0019 (0.0015, 0.0024)	46.3906 (42.0590, 51.1684)	4.7755 (4.3019, 5.3013)	1.4257 (1.2693, 1.6013)
Slope for time to index date[1] (variance)	0.0027 (0.0021, 0.0033)	0.0770 (0.0650, 0.0913)	0.0001 (0.0000, 0.0001)	0.8162 (0.6868, 0.9701)	0.1059 (0.0895, 0.1252)	0.0260 (0.0210, 0.0321)	0.0021 (0.0017, 0.0027)	0.0564 (0.0465, 0.0684)	0.0001 (0.0000, 0.0001)	0.8589 (0.6965, 1.0592)	0.0950 (0.0773, 0.1169)	0.0259 (0.0204, 0.0330)
Covariance between Intercept for patient and slope for time to index date (covariance)	-0.0117 (-0.0143, -0.0090)	-0.3652 (-0.4175, -0.3128)	-0.0003 (-0.0004, -0.0002)	-4.7011 (-5.3479, -4.0543)	-0.5593 (-0.6397, -0.4789)	-0.1499 (-0.1752, -0.1245)	-0.0097 (-0.0121, -0.0072)	-0.2864 (-0.3327, -0.2401)	-0.0003 (-0.0004, -0.0002)	-4.1096 (-4.8676, -3.3516)	-0.4127 (-0.4939, -0.3314)	-0.1432 (-0.1718, -0.1145)
Residual (variance)	0.0600 (0.0574, 0.0627)	0.9575 (0.9180, 0.9988)	0.0007 (0.0006, 0.0008)	10.7772 (10.3277, 11.2462)	1.4429 (1.3814, 1.5071)	0.7146 (0.6834, 0.7472)	0.0582 (0.0557, 0.0608)	0.8588 (0.8231, 0.8961)	0.0006 (0.0005, 0.0007)	11.9004 (11.3826, 12.4418)	1.4914 (1.4255, 1.5603)	0.6944 (0.6633, 0.7270)

[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
Abbreviations: RBC=red blood cells, Hb=haemoglobin, Hc=haematocrit, MCV=mean corpuscular volume, MCH=mean corpuscular haemoglobin, MCHC=mean corpuscular haemoglobin concentration.

Table C9: Mixed-effect model coefficients (95% CI) to assess platelet parameters by Dukes' Stage A vs. D

Variable	Males		Females	
	Platelets	**MPV**	**Platelets**	**MPV**
N	1528	211	1181	176
n cases	94	720	68	514
Fixed effects:				
Constant	362.7745 (244.9091, 480.6400)	10.1625 (9.0258, 11.2992)	228.8615 (80.0121, 377.7110)	9.3772 (8.2775, 10.4769)
Age at index date[1] (years)	-1.6245 (-3.7115, 0.4624)	-0.0159 (-0.0322, 0.0003)	1.2336 (-1.4253, 3.8925)	-0.0065 (-0.0218, 0.0088)
Age at index date[1] – knot at 60 (years)	1.0595 (-2.3577, 4.4768)		-1.0526 (-5.6207, 3.5154)	
Age at index date[1] – knot at 70 (years)	2.4923 (-0.7897, 5.7742)		-0.7692 (-5.1260, 3.5876)	
Age at index date[1] – knot at 80 (years)	-5.2987 (-9.9446, -0.6528)		3.0129 (-1.9186, 7.9443)	
Time to index date[1] (years)	-23.8380 (-32.9754, -14.7006)	0.2090 (-0.1216, 0.5395)	-24.0242 (-33.9622, -14.0861)	0.0973 (-0.2211, 0.4158)
Time to index date[1] date – knot at 1 (years)	23.2526 (7.2207, 39.2845)	-0.3958 (-0.9658, 0.1742)	21.6704 (4.6010, 38.7399)	-0.1026 (-0.6409, 0.4356)
Time to index date[1] date – knot at 2 (years)	3.9099 (-7.2401, 15.0600)	0.1521 (-0.2582, 0.5624)	5.8590 (-5.7724, 17.4905)	-0.0305 (-0.3859, 0.3249)
Time to index date[1] date – knot at 4 (years)	-4.1553 (-9.1666, 0.8560)	0.0945 (-0.1040, 0.2929)	-7.3941 (-12.7141, -2.0742)	0.0728 (-0.0984, 0.2441)
Dukes' D (vs. A)	165.7902 (-11.5791, 343.1595)	-0.2669 (-0.7052, 0.1715)	316.5678 (104.2109, 528.9247)	-0.4065 (-0.8700, 0.0570)
Dukes' D (vs. A) by time to index date[1] interaction	-44.5130 (-57.1080, -31.9180)	-0.1359 (-0.5956, 0.3238)	-75.3165 (-89.7311, -60.9018)	0.1330 (-0.3165, 0.5825)
Dukes' D (vs. A) by time to index date[1] (knot at 1) interaction	31.4208 (8.9414, 53.9002)	0.1574 (-0.6725, 0.9873)	89.7211 (64.4038, 115.0384)	-0.2428 (-1.0078, 0.5222)
Dukes' D (vs. A) by time to index date[1] (knot at 2) interaction	7.8357 (-8.3477, 24.0191)	0.0784 (-0.5564, 0.7132)	-31.7440 (-49.3822, -14.1057)	0.2785 (-0.2305, 0.7875)
Dukes' D (vs. A) by time to index date[1] (knot at 4) interaction	5.6127 (-1.9590, 13.1844)	-0.1989 (-0.5133, 0.1155)	22.9637 (14.9723, 30.9552)	-0.2549 (-0.5125, 0.0028)
Dukes' D (vs. A) by age interaction	-1.6210 (-4.7706, 1.5287)		-3.8954 (-7.7167, -0.0740)	
Dukes' D (vs. A) by age (knot at 60) interaction	2.6723 (-2.5642, 7.9088)		5.0388 (-1.8602, 11.9379)	

Dukes' D (vs. A) by age (knot at 70) interaction	-4.3882 (-9.3424, 0.5659)		-1.5712 (-8.4503, 5.3079)	
Dukes' D (vs. A) by age (knot at 80) interaction	7.9080 (1.9337, 13.8823)		-2.1388 (-8.8827, 4.6052)	
Random effects:				
Intercept for patient (variance)	7.0e+03 (6.4e+03, 7.7e+03)	1.5550 (1.2196, 1.9824)	8.4e+03 (7.6e+03, 9.2e+03)	1.4818 (1.1520, 1.9060)
Slope for time to index date[1] (variance)	104.8360 (85.6861, 128.2657)	0.0265 (0.0159, 0.0441)	96.2314 (77.1465, 120.0376)	0.0194 (0.0121, 0.0312)
Covariance between Intercept for patient and slope for time to index date (covariance)	-6.7e+02 (-7.7e+02, -5.7e+02)	-0.1070 (-0.1661, -0.0478)	-6.6e+02 (-7.8e+02, -5.5e+02)	-0.0873 (-0.1350, -0.0396)
Residual (variance)	2.1e+03 (2.1e+03, 2.2e+03)	0.3302 (0.2892, 0.3770)	2.2e+03 (2.1e+03, 2.3e+03)	0.2551 (0.2274, 0.2862)

[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
Abbreviations: MPV=mean platelet volume.

Table C10: Mixed-effect model coefficients (95% CI) to assess white blood cell parameters by Dukes' Stage A vs. D

Variable	Males						Females					
	WBC	**Basophils**	**Eosinophils**	**Lymphocytes**	**Monocytes**	**Neutrophils**	**WBC**	**Basophils**	**Eosinophils**	**Lymphocytes**	**Monocytes**	**Neutrophils**
N	1466	1381	1385	1419	1414	1428	1137	1055	1057	1073	1071	1077
n cases	673	654	656	670	667	691	471	462	463	470	468	494
Fixed effects:												
Constant	7.7472 (3.6877, 11.8068)	0.0674 (0.0512, 0.0837)	0.0812 (-0.0790, 0.2414)	2.7383 (2.0743, 3.4024)	0.6958 (0.3986, 0.9931)	6.2180 (3.5621, 8.8738)	5.7967 (1.1308, 10.4626)	0.0685 (0.0518, 0.0853)	0.0165 (-0.1111, 0.1440)	2.0711 (1.6091, 2.5331)	0.5706 (0.3092, 0.8320)	5.5388 (3.2693, 7.8083)
Age at index date[1] (years)	-0.0077 (-0.0794, 0.0641)	-0.0000 (-0.0003, 0.0002)	0.0027 (0.0002, 0.0052)	-0.0131 (-0.0226, -0.0036)	-0.0017 (-0.0069, 0.0036)	-0.0321 (-0.0785, 0.0144)	0.0229 (-0.0602, 0.1060)	0.0000 (-0.0002, 0.0002)	0.0029 (0.0009, 0.0050)	0.0010 (-0.0054, 0.0074)	-0.0008 (-0.0055, 0.0039)	-0.0186 (-0.0584, 0.0212)
Age at index date[1] – knot at 60 (years)	-0.0008 (-0.1151, 0.1136)				0.0038 (-0.0046, 0.0122)	0.0565 (-0.0019, 0.1150)	0.0544 (-0.0831, 0.1920)				0.0066 (-0.0018, 0.0149)	0.0410 (-0.0101, 0.0921)
Age at index date[1] – knot at 70 (years)	0.0355 (-0.0571, 0.1282)		-0.0050 (-0.0097, -0.0003)		0.0007 (-0.0054, 0.0068)		-0.0761 (-0.1823, 0.0300)		-0.0043 (-0.0078, -0.0007)		-0.0051 (-0.0132, 0.0030)	

Age at index date[1] – knot at 80 (years)				0.0120 (-0.0303, 0.0544)		-0.0239 (-0.1081, 0.0603)				-0.0381 (-0.0657, -0.0104)	0.0046 (-0.0031, 0.0122)	0.0323 (-0.0185, 0.0832)
Age at index date[1] – knot at 85 (years)	-0.0098 (-0.3422, 0.3226)						0.0569 (-0.1531, 0.2669)					
Age at index date[1] – knot at 90 (years)										0.1158 (0.0017, 0.2298)		
Time to index date[1] (years)	-0.3386 (-0.6635, -0.0137)	-0.0001 (-0.0010, 0.0009)	-0.0015 (-0.0046, 0.0015)	0.1539 (0.0033, 0.3044)	-0.0216 (-0.0636, 0.0205)	-0.3035 (-0.6557, 0.0487)	-0.7255 (-1.2054, -0.2455)	-0.0010 (-0.0019, -0.0000)	-0.0024 (-0.0051, 0.0002)	-0.0529 (-0.1992, 0.0934)	-0.0489 (-0.0835, -0.0144)	-0.6451 (-1.0030, -0.2871)
Time to index date[1] date – knot at 1 (years)	0.4575 (-0.1122, 1.0271)			-0.1788 (-0.4439, 0.0864)	0.0066 (-0.0684, 0.0816)	0.2374 (-0.3948, 0.8697)	0.7098 (-0.1280, 1.5477)			-0.0228 (-0.2689, 0.2233)	0.0491 (-0.0109, 0.1091)	0.8206 (0.1925, 1.4486)
Time to index date[1] date – knot at 2 (years)	-0.1188 (-0.5173, 0.2798)			0.0463 (-0.1407, 0.2333)	0.0040 (-0.0494, 0.0573)	0.1494 (-0.3053, 0.6040)	0.1504 (-0.4295, 0.7304)			0.1343 (-0.0333, 0.3019)	-0.0121 (-0.0537, 0.0295)	-0.1369 (-0.5769, 0.3031)

Time to index date[1] date – knot at 4 (years)	-0.0243 (-0.2011, 0.1526)			-0.0019 (-0.0843, 0.0804)	0.0095 (-0.0139, 0.0328)	-0.1059 (-0.3126, 0.1009)	-0.2785 (-0.5369, -0.0200)			-0.0957 (-0.1750, -0.0164)	0.0089 (-0.0099, 0.0278)	-0.1555 (-0.3541, 0.0430)
Dukes' D (vs. A)	-1.9183 (-8.0906, 4.2541)	0.0025 (-0.0034, 0.0084)	0.0425 (-0.1947, 0.2797)	-0.0489 (-0.2636, 0.1657)	0.1502 (0.1103, 0.1902)	-0.4425 (-4.4857, 3.6007)	7.2915 (0.7634, 13.8195)	-0.0047 (-0.0112, 0.0019)	0.1932 (0.0069, 0.3796)	-0.3276 (-0.5498, -0.1055)	0.1740 (0.1376, 0.2104)	1.6704 (1.3286, 2.0122)
Dukes' D (vs. A) by time to index date[1] interaction	-1.3254 (-1.7719, -0.8789)	-0.0001 (-0.0015, 0.0013)	-0.0005 (-0.0050, 0.0040)	0.1569 (-0.0509, 0.3647)	-0.1632 (-0.2215, -0.1050)	-1.4907 (-1.9796, -1.0017)	-1.1431 (-1.8428, -0.4434)	0.0007 (-0.0008, 0.0021)	-0.0010 (-0.0049, 0.0029)	0.3316 (0.1189, 0.5443)	-0.1791 (-0.2293, -0.1288)	-1.3494 (-1.8709, -0.8280)
Dukes' D (vs. A) by time to index date[1] (knot at 1) interaction	1.0373 (0.2407, 1.8338)			-0.1110 (-0.4831, 0.2611)	0.2058 (0.1002, 0.3115)	1.4853 (0.5924, 2.3782)	1.1213 (-0.1308, 2.3734)			-0.2867 (-0.6540, 0.0806)	0.1888 (0.0991, 0.2785)	1.1001 (0.1604, 2.0399)
CRC present by time to index date[1] (knot at 2) interaction	0.2554 (-0.3221, 0.8328)			-0.0800 (-0.3519, 0.1918)	-0.0628 (-0.1404, 0.0149)	-0.1088 (-0.7715, 0.5540)	-0.1675 (-1.0561, 0.7210)			-0.0714 (-0.3286, 0.1857)	-0.0134 (-0.0772, 0.0504)	0.1630 (-0.5133, 0.8393)
Dukes' D (vs. A) by time to index date[1] (knot at 4) interaction	0.0695 (-0.1985, 0.3374)			0.0279 (-0.0981, 0.1538)	0.0201 (-0.0155, 0.0557)	0.1268 (-0.1886, 0.4423)	0.3730 (-0.0208, 0.7668)			0.0909 (-0.0321, 0.2138)	0.0060 (-0.0230, 0.0351)	0.2539 (-0.0527, 0.5605)

Dukes' D (vs. A) by age interaction	0.0721 (-0.0373, 0.1814)		-0.0009 (-0.0046, 0.0028)			0.0412 (-0.0294, 0.1117)	-0.0959 (-0.2133, 0.0215)		-0.0027 (-0.0056, 0.0002)			
Dukes' D (vs. A) by age (knot at 60) interaction	-0.1128 (-0.2888, 0.0632)					-0.0626 (-0.1508, 0.0256)	0.0633 (-0.1418, 0.2685)					
Dukes' D (vs. A) by age (knot at 70) interaction	0.0054 (-0.1323, 0.1431)		0.0025 (-0.0040, 0.0090)				0.0310 (-0.1353, 0.1973)		0.0025 (-0.0025, 0.0075)			
Dukes' D (vs. A) by age (knot at 80) interaction						0.0471 (-0.0571, 0.1513)						
Dukes' D (vs. A) by age (knot at 85) interaction	0.0766 (-0.2908, 0.4441)						0.0651 (-0.1961, 0.3264)					
Random effects:												
Intercept for patient (variance)	8.5090 (7.7695, 9.3188)	0.0017 (0.0014, 0.0019)	0.0299 (0.0267, 0.0335)	2.9781 (2.7073, 3.2759)	0.0508 (0.0452, 0.0570)	1.7479 (1.4838, 2.0591)	4.6264 (4.0525, 5.2815)	0.0017 (0.0015, 0.0020)	0.0188 (0.0167, 0.0212)	2.2505 (2.0433, 2.4788)	0.0356 (0.0314, 0.0403)	2.2038 (1.9020, 2.5535)

Slope for time to index date[1] (variance)	0.1028 (0.0806, 0.1311)	0.0000 (0.0000, 0.0000)	0.0002 (0.0001, 0.0003)	0.0111 (0.0073, 0.0170)	0.0003 (0.0002, 0.0005)	0.0550 (0.0419, 0.0722)	0.0449 (0.0283, 0.0713)	0.0000 (0.0000, 0.0000)	0.0002 (0.0002, 0.0003)	0.1144 (0.1011, 0.1295)	0.0004 (0.0003, 0.0005)	0.0190 (0.0114, 0.0319)
Covariance between Intercept for patient and slope for time to index date (covariance)	-0.7927 (-0.9275, -0.6580)	-0.0001 (-0.0002, -0.0001)	-0.0016 (-0.0021, -0.0010)	-0.1506 (-0.1994, -0.1018)	-0.0034 (-0.0043, -0.0024)	-0.1551 (-0.2124, -0.0979)	-0.3046 (-0.4090, -0.2001)	-0.0001 (-0.0002, -0.0001)	-0.0014 (-0.0018, -0.0011)	-0.4654 (-0.5169, -0.4139)	-0.0028 (-0.0035, -0.0021)	-0.1071 (-0.1568, -0.0575)
Residual (variance)	2.3451 (2.2429, 2.4520)	0.0016 (0.0015, 0.0017)	0.0213 (0.0203, 0.0223)	0.4792 (0.4578, 0.5016)	0.0421 (0.0403, 0.0440)	3.1919 (3.0522, 3.3380)	5.1725 (4.9509, 5.4041)	0.0014 (0.0013, 0.0014)	0.0080 (0.0076, 0.0084)	0.3481 (0.3309, 0.3661)	0.0233 (0.0222, 0.0245)	2.7399 (2.6150, 2.8707)

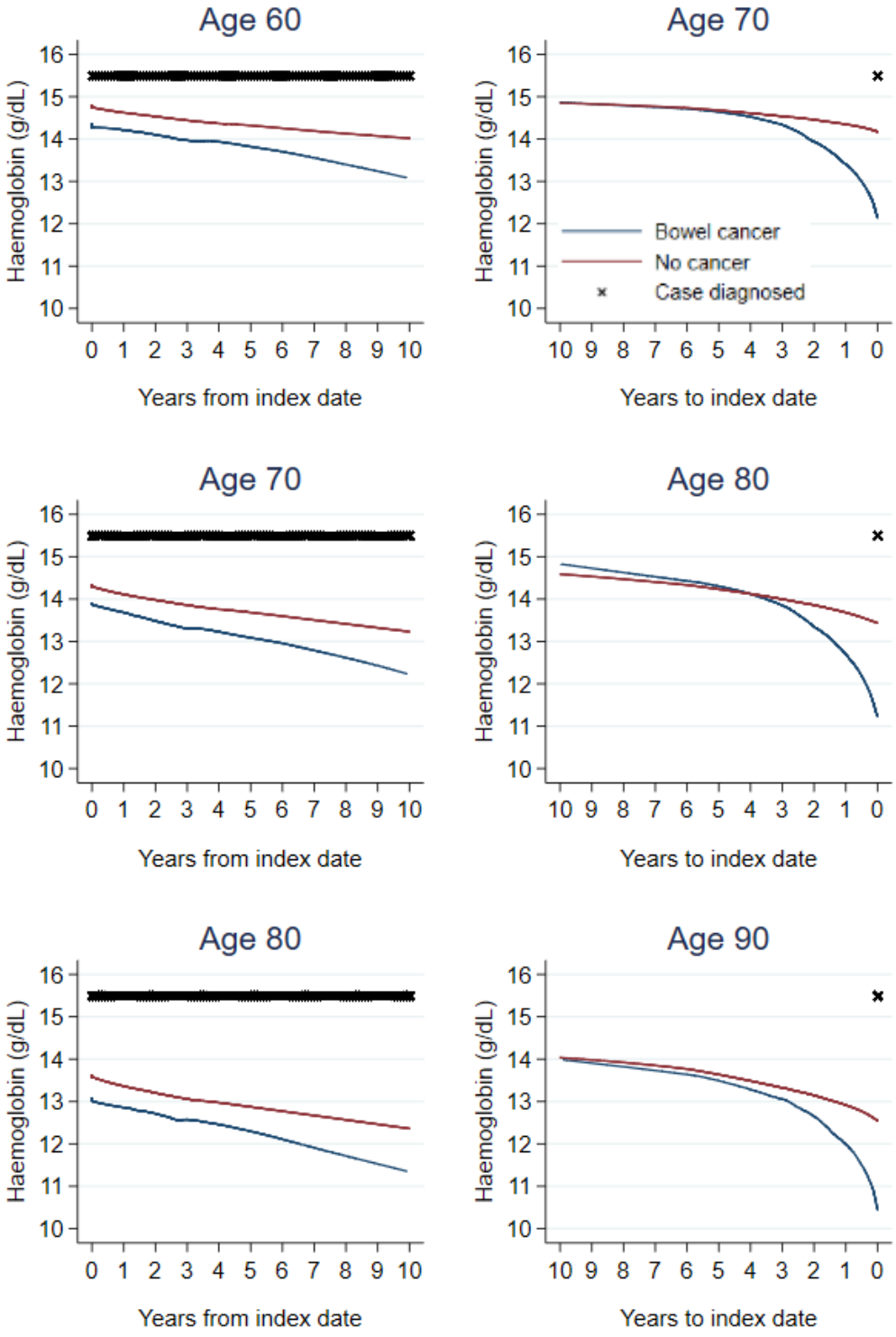
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls.
Abbreviations: WBC=white blood cells.

C11. Sensitivity analysis 1 – cohort study design

In the cohort design, 9,407 (1.5%) of 648,973 males and 8,352 (1.0%) of 820,441 females within 10 years of index (first FBC). In the cohort study, there were more cases and patients overall than in the case-control study because patients without FBCs before the randomly selected index date (controls only), without any FBC within 10 years prior to index, or diagnosed after study exit were excluded from the case-control study. The median (IQR) number of years between index and event/censor date was 3.0 (0.8, 5.7) for male cases, 5.7 (2.8, 8.9) for male non-cases, 3.2 (1.0, 6.0) for female cases, and 6.2 (3.1, 9.5) for female non-cases.

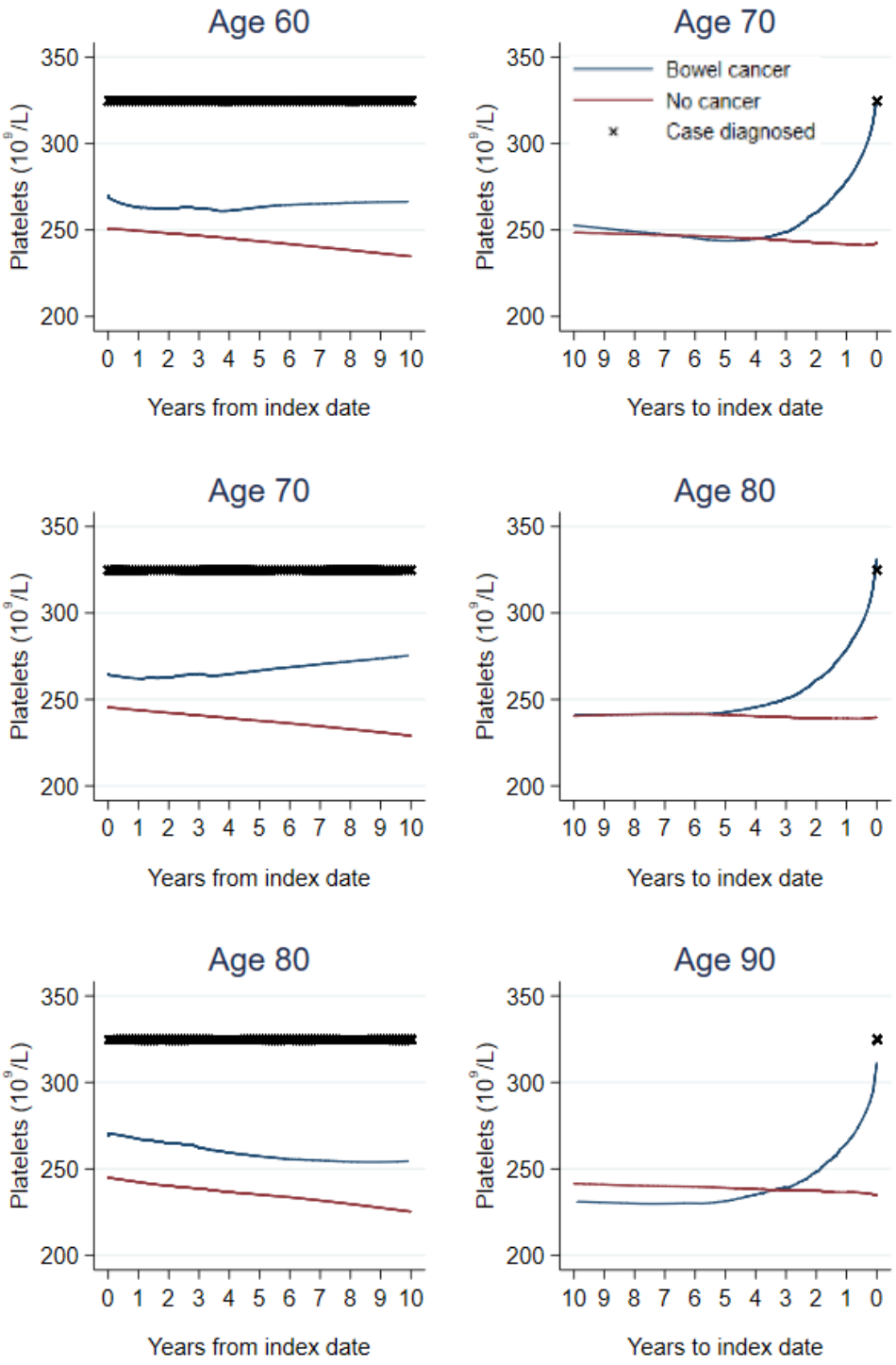
Figure C75 and Figure C76 show how haemoglobin and platelet trajectories are influenced by the use of a cohort study design. Haemoglobin trajectories in patients without bowel cancer are similar between case-control and cohort study designs for each age group but differ slightly for platelets. Trajectories in patients diagnosed are substantially different between study designs. For example, lower haemoglobin and higher platelet levels, which are more likely to represent abnormalities, are observed at index date (first FBC) for cases in the cohort design, whereas they are comparable between groups in the case-control design and diverge around four years before diagnosis. Additionally, blood levels change at a greater rate over the entire 10-year period in cases than non-cases in the cohort design, which is only observed within four years of diagnosis in the case-control design. These differences can be explained by the lack of a standardised time point (diagnosis) to identify trajectories in cases in the cohort study design. For example, diagnoses are made sporadically throughout the 10-year window in a cohort design but only at index date in the case-control design (represented by an ‘X’ in Figure C75 and Figure C76).

Figure C75: Trajectories (LOWESS) in haemoglobin for males with and without bowel cancer – cohort (left) vs. case-control (right) design[1, 2, 3]



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls in the case-control design and date of the first FBC in the cohort study design.
[2]LOWESS trends are age (+/- 3 years) at index.
[3]Rows correspond to the same age group of patients. For example, patients aged 70 at index in the cohort design and aged 80 at index in the case-control design correspond to the same age group.

Figure C76: Trajectories (LOWESS) in platelets for males with and without bowel cancer – cohort (left) vs. case-control (right) design[1, 2, 3]



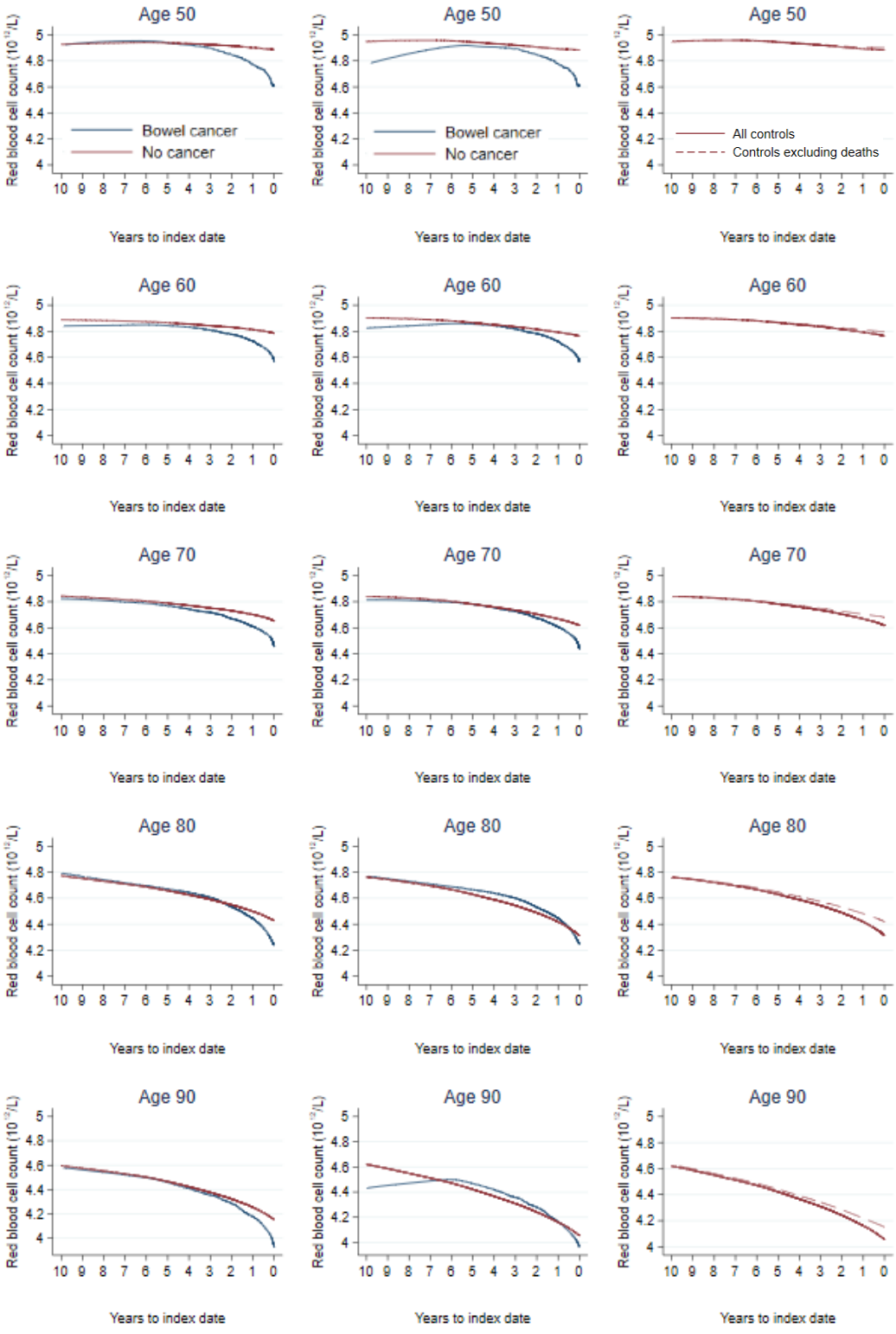
[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period for controls in the case-control design and date of the first FBC in the cohort study design.
[2]LOWESS trends are age (+/- 3 years) at index.
[3]Rows correspond to the same age group of patients. For example, patients aged 70 at index in the cohort design and aged 80 at index in the case-control design correspond to the same age group.

C12. Sensitivity analysis 2 – use study exit as index date in controls

Analyses are reported for red blood cell count and platelets for males only, as examples. Despite the eligibility criteria being the same, the number of male controls included in this analysis (n=629,635) was higher than in the case-control analysis (n=390,150). This is because the latter included only FBCs up to a randomly selected index date so patients without FBCs before this random date were excluded from the analysis. However, in this sensitivity analysis, these patients were included because they had FBCs prior to study exit (and after the random index date). Among the 629,635 males, the reason for study exit was death (7.7%, n=48,622), leaving the practice (17.5%, n=110,071), or end of study/data-cut date (74.8%, n=470,942).

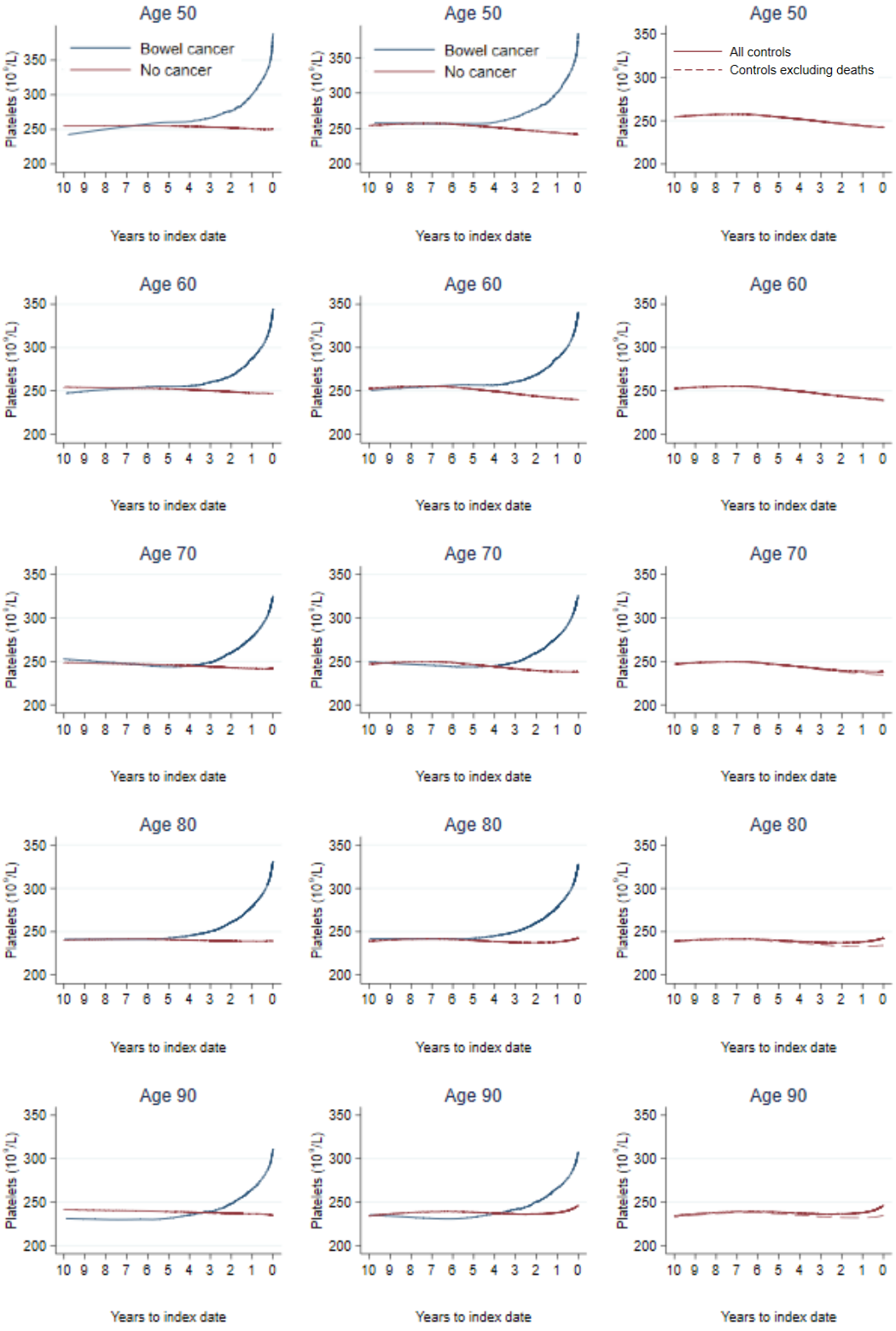
Trajectories in red blood cell count and platelet count are provided in Figure C77 and Figure C78, respectively. For cases and younger controls, the trajectories were similar when using study exit as index date (middle panel) compared to a random date in the patient's study period (left panel). However, in older controls, trajectories differed close to index date – a greater decline in red blood cell count and increase in platelets was observed when using the study exit date as index. These differences are likely caused by patients who died (right panel), which is a reason for study exit, with the patients having lowered blood levels in the run up to death than patients who did not die. This is consistent with existing literature that has shown the FBC can identify risk of death[260, 261]. When excluding controls who died from this analysis, the trajectories become comparable to those identified using a randomly selected index date. This suggests a using a randomly selected index date is ideal, as it includes all patients in the study and eliminates the confounding effect of death on FBC levels.

Figure C77: Trajectories (LOWESS) in red blood cell count for males by index date definition – random (left) vs. study exit (middle) with and without deaths (right)[1, 2]



[1]Index date was the date of diagnosis for cases and 1) a randomly selected date in the patient's study period for controls (left) or 2) study exit for controls (middle and right).
[2]LOWESS trends are age (+/- 3 years) at index.

Figure C78: Trajectories (LOWESS) in platelets for males with and without bowel cancer – random (left) vs. study exit (right) index date in controls[1, 2]



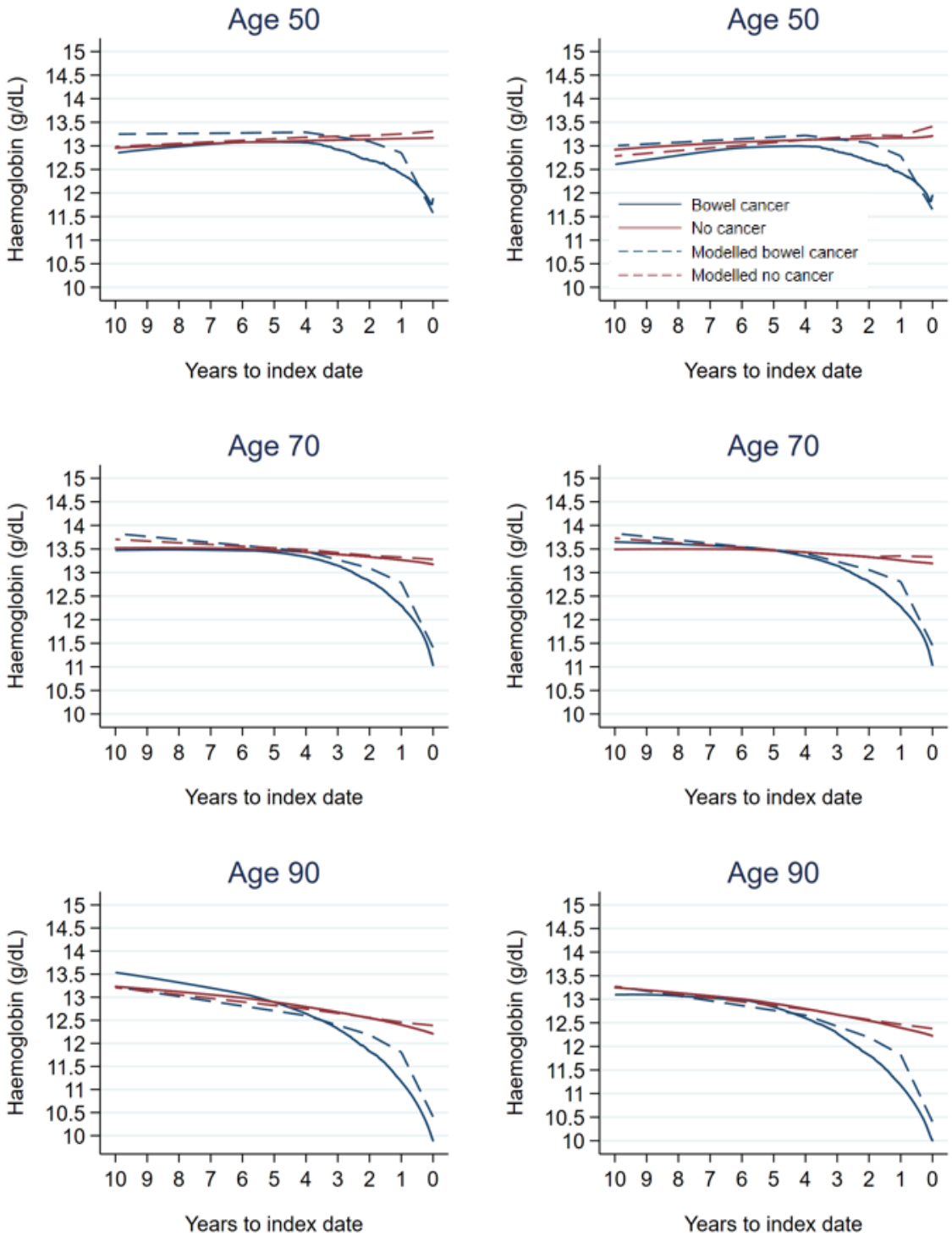
[1]Index date was the date of diagnosis for cases and 1) a randomly selected date in the patient's study period for controls (left) or 2) study exit for controls (middle and right).
[2]LOWESS trends are age (+/- 3 years) at index.

C13. Sensitivity analysis 3 – matched design

A matched design gave in 54,198 males (9,033 cases, 45,165 controls) and 48,918 females (8,153 cases, 40,765 controls). For cases and controls separately, mean (SD) age at index was 70.8 (10.6) for males and 73.0 (11.8) for females and median (min-max) follow-up time from first FBC to index was 3.1 (0-10) years for males and 3.3 (0-10) years for females.

There was no difference in graphical trends when using a matched design (see Figure C79 for haemoglobin and Figure C80 for platelets for females, as examples). Additionally, coefficients from mixed effects models (not reported here) changed only slightly, with 86.4% and 86.3% of coefficients changing by only <0.1 for males and females respectively.

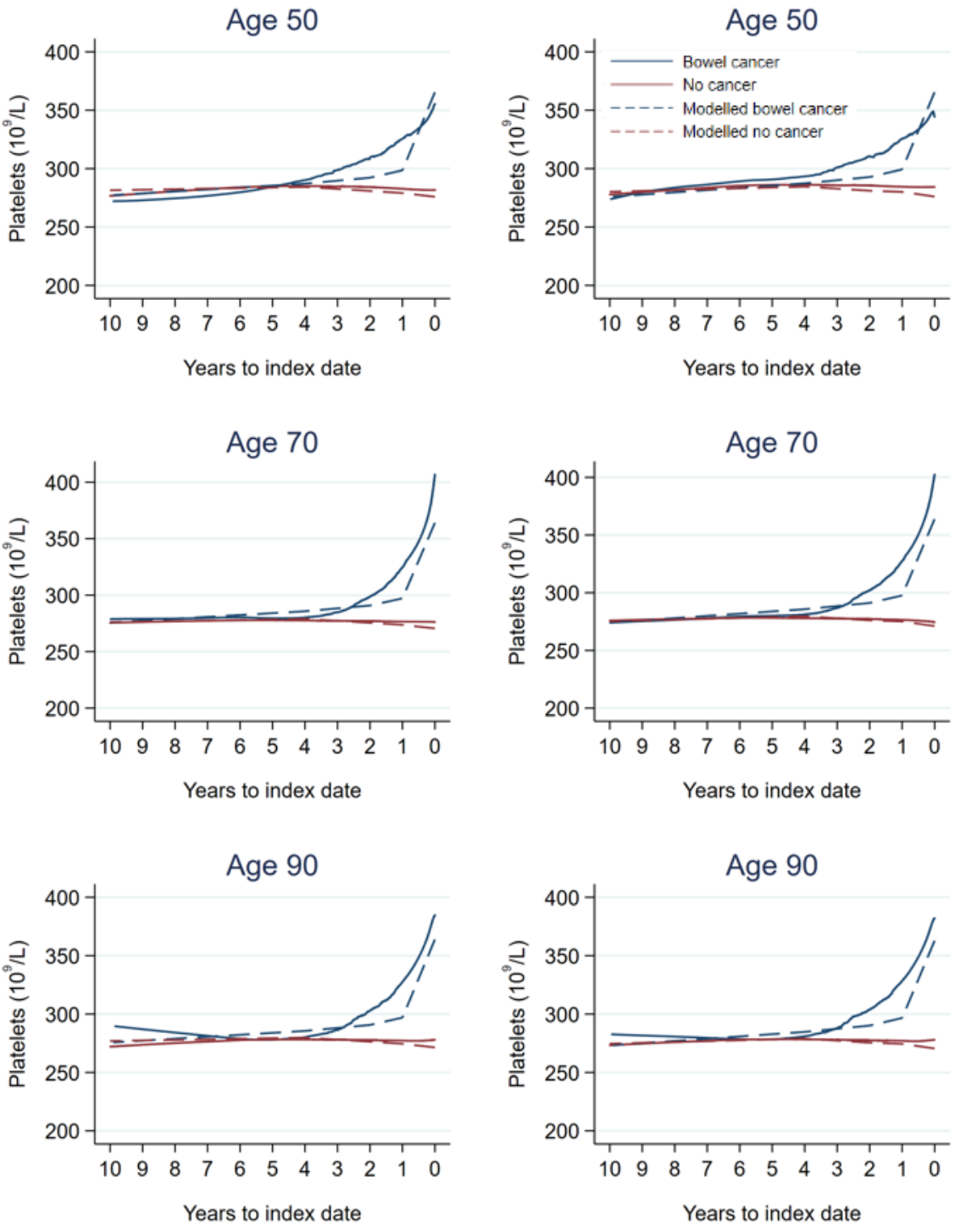
Figure C79: Haemoglobin trends by age at index[1, 2]: unmatched (left) and matched (right)



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period form controls.

[2]LOWESS trends are age (+/- 3 years) at index.

Figure C80: Platelet trends by age at index[1, 2]: unmatched (left) and matched (right)



[1]Index date was the date of diagnosis for cases and a randomly selected date in the patient's study period form controls.
[2]LOWESS trends are age (+/- 3 years) at index.

Appendix D

Appendices for Chapter 5 (joint model development)

D1. Justification for choice of candidate predictors

1. A recent systematic review identified many predictors included in existing prediction models for bowel cancer[136]. Age and sex were the most common predictors and others included body mass index, ethnicity, family history, and various factors related to lifestyle, diet, and medication use. It is generally known that many of these predictors are not accurately recorded in EHRs, particularly ethnicity, family history, and alcohol consumption levels (confirmed by Alice Fuller, the CPRD representative involved in this research). Although relevant to the disease, using inaccurately recorded predictors could compromise the reliability of a model's predictions.

2. The ColonFlag model, which is actively being used in Israel, relies on only the age, sex, and FBC results to score patients[64]. The model appeared to work well at 18-24 months prior to diagnosis (c-statistic = 0.78) in UK patient data. This suggests that age, sex, and FBC results alone may be sufficient to identify risk of diagnosis accurately.

3. Almost anything can influence FBC results and be considered confounders[262-264]. These include medication use, such as aspirin[265], and co-morbidities, such as cardiovascular disease[266]. Some of these confounding variables are not available in EHR datasets. For example, diet and vitamin supplementation can effect haemoglobin-related parameters[267] and exercise levels can effect haemoglobin-, platelet-, and white blood cell-related parameters[268]. Level of sleep is known to effect white blood cell-related parameters[269, 270]. Additionally, altitude is well known to effect haemoglobin-related parameters, with high altitude hypoxia increasing blood levels[49, 50, 271, 272].

4. Joint models are computationally intensive to develop. It would likely not have been possible to include many predictors in the model without compromising other data, such as reducing the sample size, to make the development practical and achievable.

D2. Justification for choice of final longitudinal predictors

1. **Clinical relevance**: iron-deficiency anaemia, a known risk factor for bowel cancer and used in practice to aid referral for further testing[53, 273], is reflected in haemoglobin and mean corpuscular volume. Using trends in these two blood levels could capture patients before developing iron-deficiency anaemia. Platelet count was included because the trends in Chapter 4 are suspected to be specific to bowel cancer.

2. **Systematic review findings**: the review in Chapter 2 identified six relevant parameters, of which haemoglobin, mean corpuscular volume, and platelets were either statistically significant in meta-analysis or were consistently statistically significant in each study that analysed the blood level.

3. **Amount of missing data**: the data quality assessment in Chapter 3 identified that these parameters had the least amount of missing data among all 20, missing in less than 7% of FBCs. Using these commonly available parameters would maximise the number of FBCs used to identify trajectories.

4. **Correlation**: many parameters are mathematically related so would be highly correlated. including correlated predictors in a single model can lead to bias in the estimation of regression coefficients[125, 185]. These three parameters are not mathematically related and have little-to-no correlation in levels and trends (Appendix D3). These three parameters each reflects a different aspect of the blood.

5. **Computational intensiveness**: including many more parameters would increase computer capacity and time taken to build the model. Limiting to these three parameters makes the modelling practical and achievable.

D3. Correlation between FBC parameters in the joint model

Objective:

To assess correlation between haemoglobin, mean corpuscular volume, and platelets over time.

Methods:

The data used in the case-control analysis to identify the trajectories reported in Chapter 4 was used here. This data was used instead of the one used in Chapter 5 to build the model because it gives a larger patient sample. Using scatterplots, I assessed the relationship between the three blood levels for cases and controls separately – due to non-linearity, Spearman correlation was used.

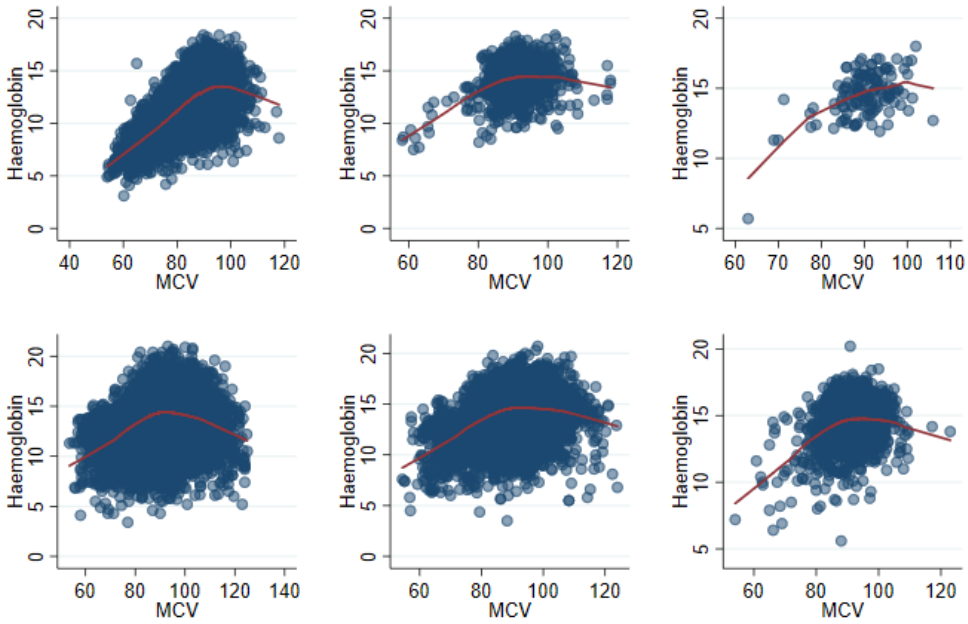
At each time point (to two decimal places) before diagnosis, the Spearman correlation coefficient (95% CI) between pairwise parameters was calculated. LOWESS curves were used to plot the trends in correlation coefficients and 95% CI over time.

Analyses were performed in Stata SE V15.1.

Results:

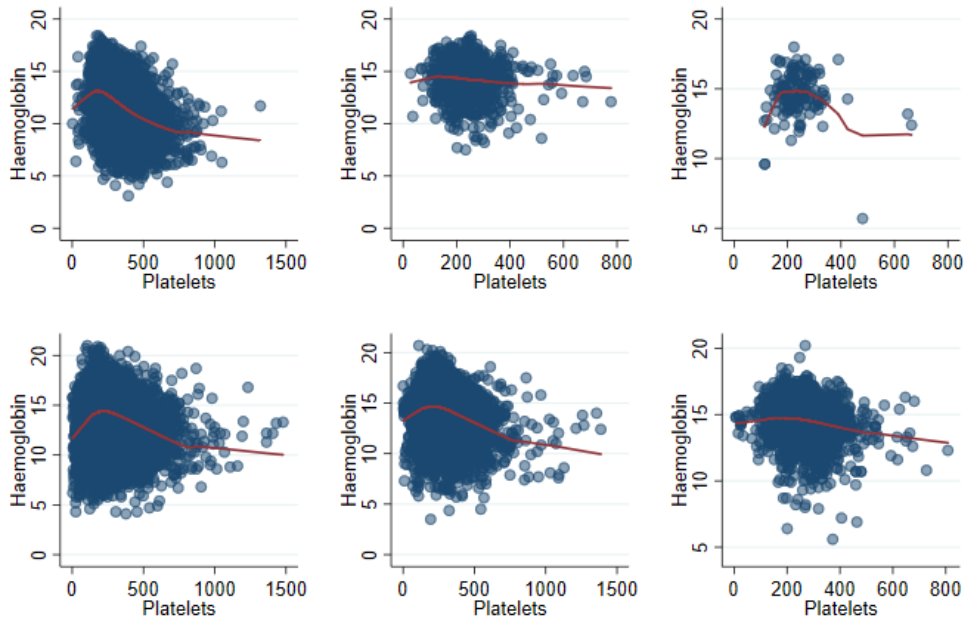
Figure D1, Figure D2, and Figure D3 show a non-linear relationship between each correlation pair at different time points. Consequently, Spearman correlation analysis was performed.

Figure D1: Relationships between haemoglobin and MCV at diagnosis/censor date (left), 5 (middle), and 10 (right) years prior for cases (top) and controls (bottom)[1]



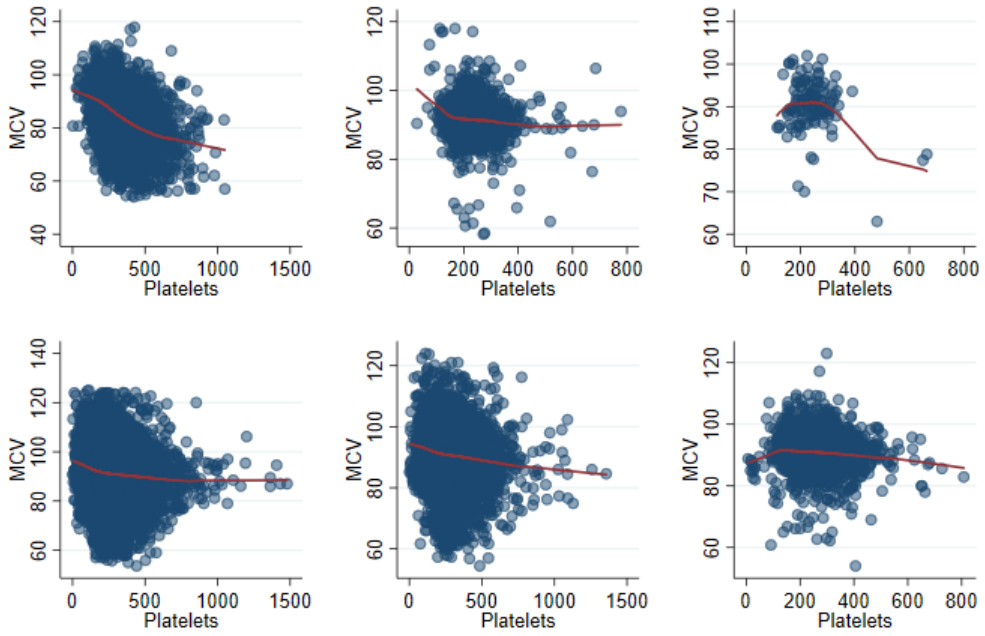
Abbreviations: MCV=mean corpuscular volume.
[1]Blue dots are scatter points, red line is the LOWESS smoother

Figure D2: Relationships between haemoglobin and platelets at diagnosis/censor date (left), 5 (middle), and 10 (right) years prior for cases (top) and controls (bottom)[1]



[1]Blue dots are scatter points, red line is the LOWESS smoother

Figure D3: Relationships between MCV and platelets at diagnosis/censor date (left), 5 (middle), and 10 (right) years prior for cases (top) and controls (bottom)[1]



Abbreviations: MCV=mean corpuscular volume.
[1]Blue dots are scatter points, red line is the LOWESS smoother

As multiple correlation analyses were performed per correlation pair, one for each time point before diagnosis/censor date, Table D1 shows a summary of the number of patients analysed among the correlation analyses.

Table D1: Summary of number of patients per correlation pair

Correlation pair	Mean	SD	Median	Min	Max
Cases:					
Haemoglobin, MCV	36.1	43.9	25	2	419
Haemoglobin, platelets	35.8	43.5	25	2	424
MCV, platelets	35.2	43.4	24	2	421
Controls:					
Haemoglobin, MCV	1084.7	846.0	835	65	3407
Haemoglobin, platelets	1073.0	838.2	821	63	3363
MCV, platelets	1056.6	829.9	804	60	3360

Table D2 shows a summary of the Spearman correlation coefficients estimated for each correlation pair. On average, there was weak correlation between each pair of FBC parameters (coefficients are close to zero).

Table D2: Summary of Spearman correlation coefficients per correlation pair

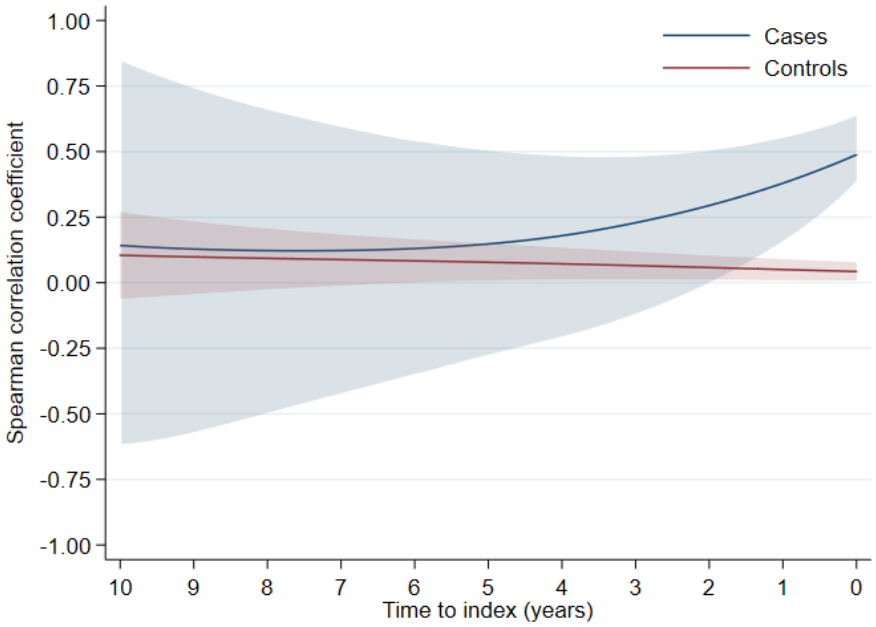
Correlation pair	Mean	SD	Median	Min	Max
Cases:					
Haemoglobin, MCV	0.18	0.32	0.20	-1	1
Haemoglobin, platelets	-0.14	0.30	-0.15	-1	1
MCV, platelets	-0.13	0.31	-0.17	-1	1
Controls:					
Haemoglobin, MCV	0.08	0.05	0.07	-0.14	0.31
Haemoglobin, platelets	-0.10	0.05	-0.10	-0.32	0.06
MCV, platelets	-0.08	0.05	-0.08	-0.23	0.16

Abbreviations: MCV=mean corpuscular volume.

Trends in the correlation coefficient over time for each correlation pair are in Figure D4, Figure D5, and Figure D6. In controls, correlation was close to zero at all times, suggesting the three parameters have weak correlation. This was similar in cases, but correlation grew from four years towards diagnosis, which corresponds to the time where trends in cases diverged from controls (Chapter 4). However, correlation remained low overall (confined within -0.5 and 0.5).

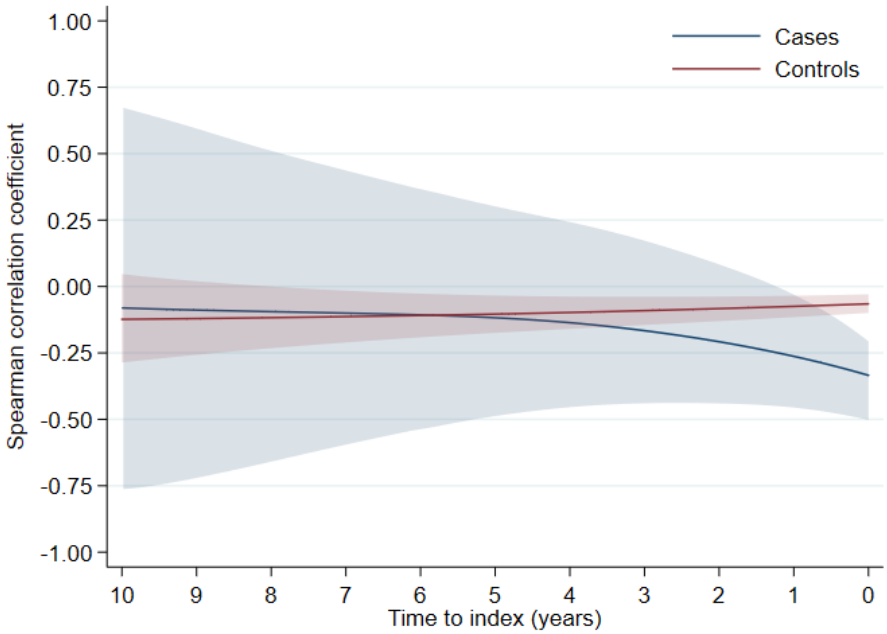
The 95% CIs between cases and controls overlapped up until close to diagnosis. At diagnosis, the correlation coefficient between haemoglobin and mean corpuscular volume reached around 0.5 (95% CI=0.4-0.6), suggesting moderate correlation.

Figure D4: Trends in the Spearman correlation coefficient and 95% CI (shaded area) between haemoglobin and mean corpuscular volume towards index date[1]



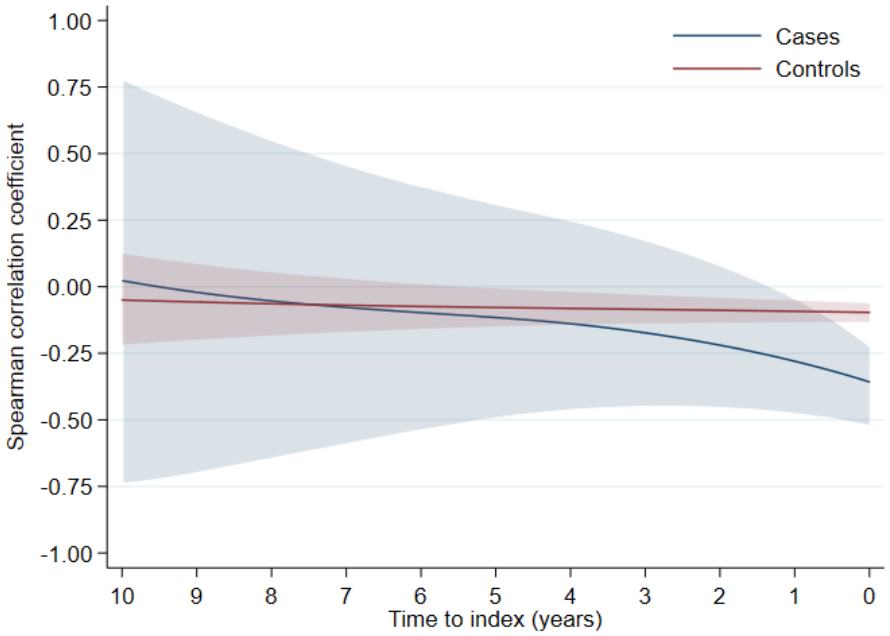
[1]Index date was the date of diagnosis (cases) or a randomly selected date in the study period (controls)

Figure D5: Trends in the Spearman correlation coefficient and 95% CI (shaded area) between haemoglobin and platelets towards index date[1]



[1]Index date was the date of diagnosis (cases) or a randomly selected date in the study period (controls)

Figure D6: Trends in the Spearman correlation coefficient and 95% CI (shaded area) between mean corpuscular volume and platelets towards index date[1]



[1]Index date was the date of diagnosis (cases) or a randomly selected date in the study period (controls)

Conclusion:

Correlation between the three parameters was low overall. In cases, it became moderate between haemoglobin and mean corpuscular volume close to diagnosis, but this may be because both are red blood cell-related parameters and are more closely related than with platelets. Additionally, for each pair, it is likely that correlation is increasing due to the presence of bowel cancer, which effects the FBC levels, and therefore may not reflect a genuine correlation, particularly as correlation was low throughout for controls.

These findings suggest these three FBC parameters are suitable to include in the multivariate joint model (Chapter 5), as the model uses trajectories up to two years prior to the event date, where correlation remains low and 95% Cis overlap.

D4. Justification for using joint modelling

1. My area of expertise is statistics and I therefore focus on statistical methods, as opposed to methods like machine-learning.

2. Conventional approaches to develop prediction models use logistic and Cox regression. These rely on risk factors measured at a single, baseline time-point to estimate risk of a future event. However, joint models use risk factors measured at multiple time-points before baseline to estimate this risk. They could therefore use relevant trajectories in risk factors to identify risk of a future event for each patient, such as those identified in Chapter 4.

3. A joint model consists of a mixed-effect sub-model and Cox sub-model. These conventional model types are well understood in the scientific community, so are easier to describe, interpret, and clinically contextualise. They are also easier to embed into general practice compared to methods like machine-learning, which are often considered “black box” algorithms because their methodology is not transparent and do not instil confidence in clinicians to use such algorithms to make health-related decisions. Therefore, my supervisors and I anticipate that the joint model will be relatively straightforward to implement and use in practice.

4. Joint models can account for three types of known correlation: 1) between repeated measures within subjects, 2) between multivariate longitudinal outcomes, and 3) between the mixed-effects and Cox sub-models. Other approaches, such as two-stage approaches and extended Cox models (which contain time-varying covariates) do not account for such correlation. Sweeting and Thompson compared joint model types to two-stage approaches and reported drawbacks of the latter include that it does not

correct for event-dependent dropout[201]. Additionally, uncertainty in the estimates and BLUPs from the mixed-effects model are not carried forward to the Cox model, resulting in too precise estimates.

5. Sweeting and Thompson performed simulation studies to compare joint models with two-stage approaches and extended Cox models. Their simulations identified that two-stage approaches can over-estimate the baseline survival function by 13% (for repeated measures every two years). Additionally, modelling longitudinal data as time-dependent covariates in Cox models can result in severe underestimation of the true association between trajectories and risk, due to regression dilution (estimates biased towards the null). Both two-stage approaches and extended Cox models were shown to have more bias than joint models.

6. Joint modelling requires less parameter estimation than other approaches, particularly machine-learning algorithms, such as neural networks and random forests. Simulation studies comparing logistic regression to various machine-learning algorithms showed the latter were instable and had more overfitting than logistic regression[274]. Stability and overfitting can improve with increased sample sizes, however simulation studies have shown machine-learning algorithms had reduced stability and a high optimism even with >200 events per parameter. Joint models can therefore be more appropriate for this research, given the available data used to build the joint model in Chapter 5.

7. A recent systematic review reviewed methodological frameworks for incorporating longitudinal data into risk prediction models[216]. Seven frameworks were identified, including joint models, machine-learning, two-stage approaches, extended Cox models,

and landmarking[217]. They report advantages of joint models compared to the other approaches, primarily those described in previous points.

D5. Comparing functional forms for age in the Cox sub-model

Figure D7 shows a non-linear functional form of age at baseline FBC for two-year (observed) bowel cancer status.

Figure D7: LOWESS curve of bowel cancer presence by age at index FBC to assess functional form for males (left) and females (right)

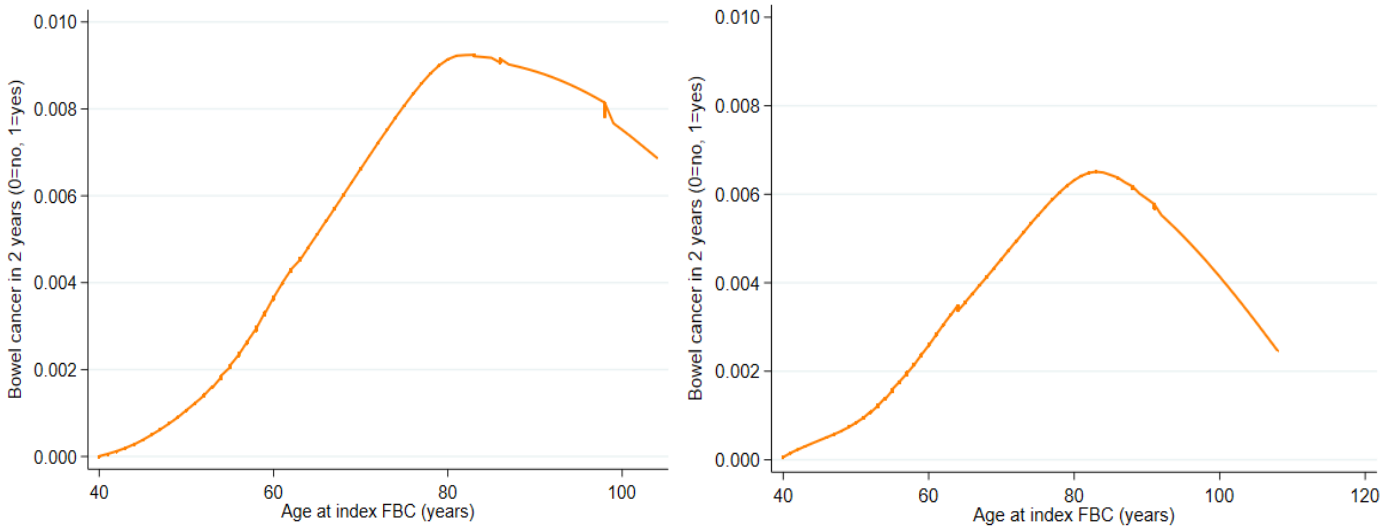


Table D3 shows the AIC and BIC for joint models fitted with various functional forms for age in the Cox sub-model. A lower AIC and BIC indicate a better fitting model. For males, fractional polynomials gave the better fitting model, with lowest AIC and BIC. For females, restricted cubic splines gave the better

fitting model, but AIC and BIC were very similar to the model with fractional polynomials, which also had a higher AUC, suggesting the latter fits similarly well.

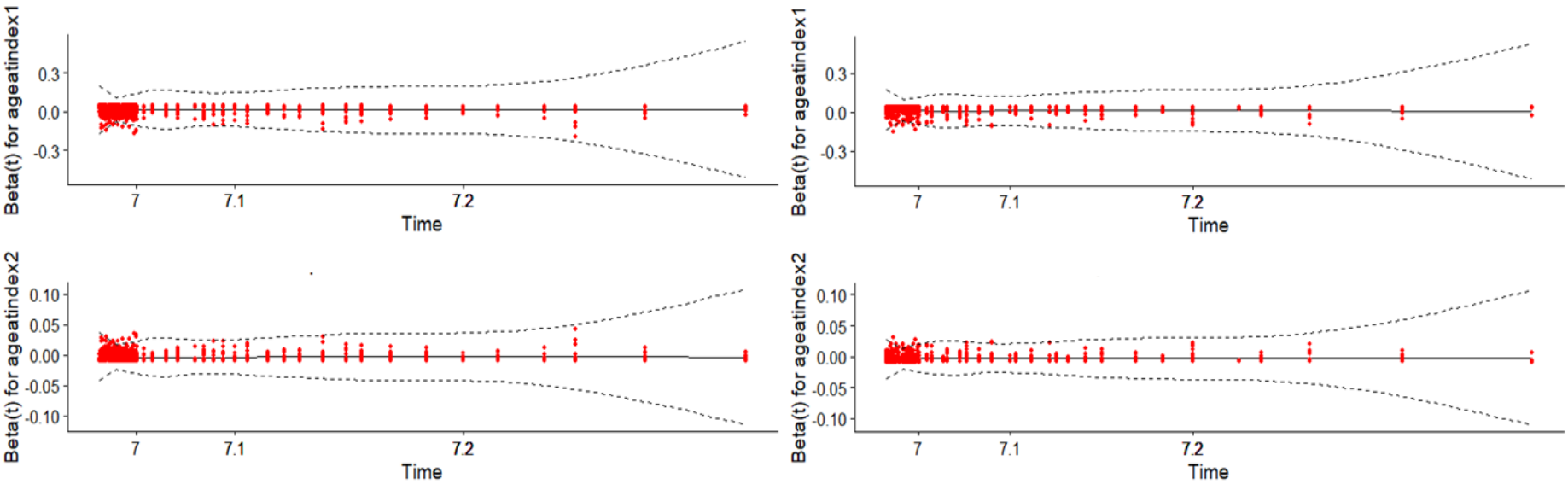
Table D3: AIC and BIC of models using various functional forms for age at baseline FBC in the Cox sub-model of a joint model[1]

Model	Males				Females			
	N	AIC	BIC	AUC	N	AIC	BIC	AUC
No age	226130	13517681	13518218	0.500	226914	15808346	15808883	0.500
Categorised[2]	226130	13517108	13517655	0.741	226914	15807925	15808473	0.740
Linear	226130	13517186	13517734	0.750	226914	15808003	15808550	0.741
Fractional polynomials[3]	226130	13517044	13517602	0.749	226914	15807878	15808436	0.756
Linear splines[4]	226130	13517236	13517804	0.731	226914	15807967	15808536	0.747
Restricted cubic splines[4]	226130	13517054	13517612	0.748	226914	15807876	15808434	0.755

Abbreviations: AIC=akaike information criteria, BIC=bayesian information criteria, FBC=full blood count, AUC=area under the curve.
[1]Each model was adjusted for trajectories in haemoglobin, mean corpuscular volume, and platelet count from mixed-effects sub-models.
[2]Categories: age≤47, 47<age≤55, 55<age≤63, 63<age≤72, and age>72 years.
[3]Fractional polynomial powers: 2, 2 (determined by the software).
[4]Knot locations: 60, 80, and 90 years (determined based on visual inspection).

Fractional polynomials did not violate the proportional hazards assumption, as the Schoenfeld residuals hover around y=0 over time (Figure D8).

Figure D8: Schoenfeld residuals to assess the proportional hazards assumption for age (fractional polynomial powers: 2, 2) for males (left) and females (right)



D6. Diagnostic test accuracy in the split sample internal validation cohort

Table D4: Diagnostic test accuracy measures of the joint models for males (split sample internal validation)

Risk centile	Risk (%) cut-off	True positives	False positives	True negatives	False negatives	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
1%	0.0099%	1039	280661	2845	1	99.90	1.00	0.37	99.96
2%	0.0108%	1038	277817	5689	2	99.81	2.01	0.37	99.96
3%	0.0116%	1038	274971	8535	2	99.81	3.01	0.38	99.98
4%	0.0125%	1038	272126	11380	2	99.81	4.01	0.38	99.98
5%	0.0135%	1037	269281	14225	3	99.71	5.02	0.38	99.98
6%	0.0144%	1037	266436	17070	3	99.71	6.02	0.39	99.98
7%	0.0155%	1034	263593	19913	6	99.42	7.02	0.39	99.97
8%	0.0166%	1034	260748	22758	6	99.42	8.03	0.39	99.97
9%	0.0177%	1033	257903	25603	7	99.33	9.03	0.40	99.97
10%	0.0190%	1032	255059	28447	8	99.23	10.03	0.40	99.97
11%	0.0202%	1032	252214	31292	8	99.23	11.04	0.41	99.97
12%	0.0216%	1032	249368	34138	8	99.23	12.04	0.41	99.98
13%	0.0230%	1032	246523	36983	8	99.23	13.04	0.42	99.98
14%	0.0246%	1030	243679	39827	10	99.04	14.05	0.42	99.97
15%	0.0261%	1028	240836	42670	12	98.85	15.05	0.43	99.97
16%	0.0278%	1028	237990	45516	12	98.85	16.05	0.43	99.97
17%	0.0296%	1026	235147	48359	14	98.65	17.06	0.43	99.97
18%	0.0315%	1026	232301	51205	14	98.65	18.06	0.44	99.97
19%	0.0335%	1025	229457	54049	15	98.56	19.06	0.44	99.97
20%	0.0356%	1024	226613	56893	16	98.46	20.07	0.45	99.97
21%	0.0377%	1024	223767	59739	16	98.46	21.07	0.46	99.97
22%	0.0401%	1023	220923	62583	17	98.37	22.07	0.46	99.97
23%	0.0424%	1023	218077	65429	17	98.37	23.08	0.47	99.97
24%	0.0449%	1022	215233	68273	18	98.27	24.08	0.47	99.97

25%	0.0477%	1020	212389	71117	20	98.08	25.08	0.48	99.97
26%	0.0503%	1019	209545	73961	21	97.98	26.09	0.48	99.97
27%	0.0533%	1018	206700	76806	22	97.88	27.09	0.49	99.97
28%	0.0564%	1016	203857	79649	24	97.69	28.09	0.50	99.97
29%	0.0596%	1015	201012	82494	25	97.60	29.10	0.50	99.97
30%	0.0628%	1014	198168	85338	26	97.50	30.10	0.51	99.97
31%	0.0665%	1012	195325	88181	28	97.31	31.10	0.52	99.97
32%	0.0701%	1008	192483	91023	32	96.92	32.11	0.52	99.96
33%	0.0741%	1003	189643	93863	37	96.44	33.11	0.53	99.96
34%	0.0783%	1001	186799	96707	39	96.25	34.11	0.53	99.96
35%	0.0828%	998	183957	99549	42	95.96	35.11	0.54	99.96
36%	0.0872%	993	181116	102390	47	95.48	36.12	0.55	99.95
37%	0.0919%	988	178276	105230	52	95.00	37.12	0.55	99.95
38%	0.0969%	983	175435	108071	57	94.52	38.12	0.56	99.95
39%	0.1019%	981	172592	110914	59	94.33	39.12	0.57	99.95
40%	0.1072%	978	169750	113756	62	94.04	40.12	0.57	99.95
41%	0.1129%	975	166907	116599	65	93.75	41.13	0.58	99.94
42%	0.1188%	974	164063	119443	66	93.65	42.13	0.59	99.94
43%	0.1249%	969	161222	122284	71	93.17	43.13	0.60	99.94
44%	0.1311%	960	158386	125120	80	92.31	44.13	0.60	99.94
45%	0.1378%	957	155543	127963	83	92.02	45.14	0.61	99.94
46%	0.1447%	950	152705	130801	90	91.35	46.14	0.62	99.93
47%	0.1517%	943	149866	133640	97	90.67	47.14	0.63	99.93
48%	0.1590%	931	147033	136473	109	89.52	48.14	0.63	99.92
49%	0.1664%	929	144189	139317	111	89.33	49.14	0.64	99.92
50%	0.1740%	922	141351	142155	118	88.65	50.14	0.65	99.92
51%	0.1817%	911	138517	144989	129	87.60	51.14	0.65	99.91
52%	0.1897%	905	135677	147829	135	87.02	52.14	0.66	99.91
53%	0.1976%	901	132836	150670	139	86.63	53.15	0.67	99.91
54%	0.2060%	892	129999	153507	148	85.77	54.15	0.68	99.90
55%	0.2144%	889	127157	156349	151	85.48	55.15	0.69	99.90

56%	0.2230%	878	124322	159184	162	84.42	56.15	0.70	99.90
57%	0.2315%	869	121486	162020	171	83.56	57.15	0.71	99.89
58%	0.2405%	857	118652	164854	183	82.40	58.15	0.72	99.89
59%	0.2491%	848	115816	167690	192	81.54	59.15	0.73	99.89
60%	0.2583%	843	112976	170530	197	81.06	60.15	0.74	99.88
61%	0.2674%	830	110143	173363	210	79.81	61.15	0.75	99.88
62%	0.2771%	820	107308	176198	220	78.85	62.15	0.76	99.88
63%	0.2867%	812	104470	179036	228	78.08	63.15	0.77	99.87
64%	0.2965%	800	101637	181869	240	76.92	64.15	0.78	99.87
65%	0.3066%	792	98799	184707	248	76.15	65.15	0.80	99.87
66%	0.3168%	782	95964	187542	258	75.19	66.15	0.81	99.86
67%	0.3275%	774	93126	190380	266	74.42	67.15	0.82	99.86
68%	0.3383%	765	90290	193216	275	73.56	68.15	0.84	99.86
69%	0.3497%	751	87458	196048	289	72.21	69.15	0.85	99.85
70%	0.3606%	744	84620	198886	296	71.54	70.15	0.87	99.85
71%	0.3720%	723	81796	201710	317	69.52	71.15	0.88	99.84
72%	0.3840%	704	78969	204537	336	67.69	72.15	0.88	99.84
73%	0.3955%	693	76135	207371	347	66.63	73.15	0.90	99.83
74%	0.4064%	670	73312	210194	370	64.42	74.14	0.91	99.82
75%	0.4175%	644	70493	213013	396	61.92	75.14	0.91	99.81
76%	0.4286%	622	67669	215837	418	59.81	76.13	0.91	99.81
77%	0.4393%	608	64838	218668	432	58.46	77.13	0.93	99.80
78%	0.4498%	589	62011	221495	451	56.63	78.13	0.94	99.80
79%	0.4600%	562	59193	224313	478	54.04	79.12	0.94	99.79
80%	0.4701%	542	56368	227138	498	52.12	80.12	0.95	99.78
81%	0.4803%	527	53537	229969	513	50.67	81.12	0.97	99.78
82%	0.4904%	508	50711	232795	532	48.85	82.11	0.99	99.77
83%	0.5002%	484	47889	235617	556	46.54	83.11	1.00	99.76
84%	0.5102%	468	45060	238446	572	45.00	84.11	1.03	99.76
85%	0.5203%	438	42244	241262	602	42.12	85.10	1.03	99.75
86%	0.5310%	418	39419	244087	622	40.19	86.10	1.05	99.75

87%	0.5416%	385	36606	246900	655	37.02	87.09	1.04	99.74
88%	0.5522%	373	33773	249733	667	35.87	88.09	1.09	99.73
89%	0.5633%	341	30959	252547	699	32.79	89.08	1.09	99.72
90%	0.5756%	312	28143	255363	728	30.00	90.07	1.10	99.72
91%	0.5887%	285	25325	258181	755	27.40	91.07	1.11	99.71
92%	0.6024%	263	22501	261005	777	25.29	92.06	1.16	99.70
93%	0.6183%	240	19679	263827	800	23.08	93.06	1.20	99.70
94%	0.6372%	214	16859	266647	826	20.58	94.05	1.25	99.69
95%	0.6583%	186	14042	269464	854	17.88	95.05	1.31	99.68
96%	0.6841%	154	11228	272278	886	14.81	96.04	1.35	99.68
97%	0.7183%	117	8420	275086	923	11.25	97.03	1.37	99.67
98%	0.7688%	80	5611	277895	960	7.69	98.02	1.41	99.66
99%	0.8590%	50	2796	280710	990	4.81	99.01	1.76	99.65

Abbreviations: PPV=positive predictive value; NPV=negative predictive value

Table D5: Diagnostic accuracy measures of the joint models for females (split sample internal validation)

Risk centile	Risk (%) cut-off	True positives	False positives	True negatives	False negatives	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
1%	0.0082%	1198	423460	4288	2	99.83	1.00	0.28	99.95
2%	0.0091%	1197	419172	8576	3	99.75	2.00	0.28	99.97
3%	0.0099%	1196	414883	12865	4	99.67	3.01	0.29	99.97
4%	0.0107%	1196	410594	17154	4	99.67	4.01	0.29	99.98
5%	0.0114%	1196	406304	21444	4	99.67	5.01	0.29	99.98
6%	0.0122%	1196	402015	25733	4	99.67	6.02	0.30	99.98
7%	0.0130%	1196	397725	30023	4	99.67	7.02	0.30	99.99
8%	0.0139%	1194	393438	34310	6	99.50	8.02	0.30	99.98
9%	0.0147%	1191	389151	38597	9	99.25	9.02	0.31	99.98
10%	0.0157%	1190	384863	42885	10	99.17	10.03	0.31	99.98
11%	0.0166%	1187	380576	47172	13	98.92	11.03	0.31	99.97

12%	0.0176%	1186	376288	51460	14	98.83	12.03	0.31	99.97
13%	0.0187%	1180	372004	55744	20	98.33	13.03	0.32	99.96
14%	0.0198%	1179	367716	60032	21	98.25	14.03	0.32	99.97
15%	0.0210%	1178	363427	64321	22	98.17	15.04	0.32	99.97
16%	0.0222%	1177	359139	68609	23	98.08	16.04	0.33	99.97
17%	0.0235%	1176	354851	72897	24	98.00	17.04	0.33	99.97
18%	0.0249%	1174	350563	77185	26	97.83	18.04	0.33	99.97
19%	0.0263%	1172	346276	81472	28	97.67	19.05	0.34	99.97
20%	0.0278%	1170	341988	85760	30	97.50	20.05	0.34	99.97
21%	0.0294%	1170	337699	90049	30	97.50	21.05	0.35	99.97
22%	0.0310%	1168	333411	94337	32	97.33	22.05	0.35	99.97
23%	0.0328%	1167	329123	98625	33	97.25	23.06	0.35	99.97
24%	0.0347%	1166	324834	102914	34	97.17	24.06	0.36	99.97
25%	0.0367%	1166	320545	107203	34	97.17	25.06	0.36	99.97
26%	0.0387%	1166	316255	111493	34	97.17	26.07	0.37	99.97
27%	0.0409%	1164	311968	115780	36	97.00	27.07	0.37	99.97
28%	0.0432%	1162	307680	120068	38	96.83	28.07	0.38	99.97
29%	0.0455%	1159	303394	124354	41	96.58	29.07	0.38	99.97
30%	0.0480%	1156	299107	128641	44	96.33	30.07	0.38	99.97
31%	0.0507%	1154	294820	132928	46	96.17	31.08	0.39	99.97
32%	0.0534%	1149	290535	137213	51	95.75	32.08	0.39	99.96
33%	0.0564%	1146	286249	141499	54	95.50	33.08	0.40	99.96
34%	0.0595%	1141	281965	145783	59	95.08	34.08	0.40	99.96
35%	0.0627%	1140	277676	150072	60	95.00	35.08	0.41	99.96
36%	0.0661%	1135	273392	154356	65	94.58	36.09	0.41	99.96
37%	0.0696%	1132	269105	158643	68	94.33	37.09	0.42	99.96
38%	0.0733%	1130	264818	162930	70	94.17	38.09	0.42	99.96
39%	0.0773%	1125	260533	167215	75	93.75	39.09	0.43	99.96
40%	0.0815%	1120	256249	171499	80	93.33	40.09	0.44	99.95
41%	0.0859%	1114	251965	175783	86	92.83	41.09	0.44	99.95
42%	0.0905%	1108	247682	180066	92	92.33	42.10	0.45	99.95

43%	0.0953%	1100	243400	184348	100	91.67	43.10	0.45	99.95
44%	0.1003%	1096	239115	188633	104	91.33	44.10	0.46	99.94
45%	0.1056%	1088	234833	192915	112	90.67	45.10	0.46	99.94
46%	0.1111%	1082	230550	197198	118	90.17	46.10	0.47	99.94
47%	0.1169%	1079	226263	201485	121	89.92	47.10	0.47	99.94
48%	0.1227%	1071	221982	205766	129	89.25	48.10	0.48	99.94
49%	0.1287%	1063	217700	210048	137	88.58	49.11	0.49	99.93
50%	0.1348%	1048	213426	214322	152	87.33	50.10	0.49	99.93
51%	0.1412%	1038	209147	218601	162	86.50	51.11	0.49	99.93
52%	0.1475%	1032	204863	222885	168	86.00	52.11	0.50	99.92
53%	0.1538%	1022	200584	227164	178	85.17	53.11	0.51	99.92
54%	0.1605%	1016	196300	231448	184	84.67	54.11	0.51	99.92
55%	0.1671%	1006	192021	235727	194	83.83	55.11	0.52	99.92
56%	0.1742%	994	187743	240005	206	82.83	56.11	0.53	99.91
57%	0.1810%	989	183459	244289	211	82.42	57.11	0.54	99.91
58%	0.1879%	977	179181	248567	223	81.42	58.11	0.54	99.91
59%	0.1949%	969	174900	252848	231	80.75	59.11	0.55	99.91
60%	0.2021%	962	170617	257131	238	80.17	60.11	0.56	99.91
61%	0.2093%	949	166341	261407	251	79.08	61.11	0.57	99.90
62%	0.2165%	944	162056	265692	256	78.67	62.11	0.58	99.90
63%	0.2237%	929	157782	269966	271	77.42	63.11	0.59	99.90
64%	0.2309%	919	153502	274246	281	76.58	64.11	0.60	99.90
65%	0.2381%	901	149231	278517	299	75.08	65.11	0.60	99.89
66%	0.2454%	887	144955	282793	313	73.92	66.11	0.61	99.89
67%	0.2526%	873	140680	287068	327	72.75	67.11	0.62	99.89
68%	0.2597%	854	136410	291338	346	71.17	68.11	0.62	99.88
69%	0.2665%	828	132146	295602	372	69.00	69.11	0.62	99.87
70%	0.2733%	806	127879	299869	394	67.17	70.10	0.63	99.87
71%	0.2802%	792	123603	304145	408	66.00	71.10	0.64	99.87
72%	0.2869%	774	119332	308416	426	64.50	72.10	0.64	99.86
73%	0.2934%	763	115053	312695	437	63.58	73.10	0.66	99.86

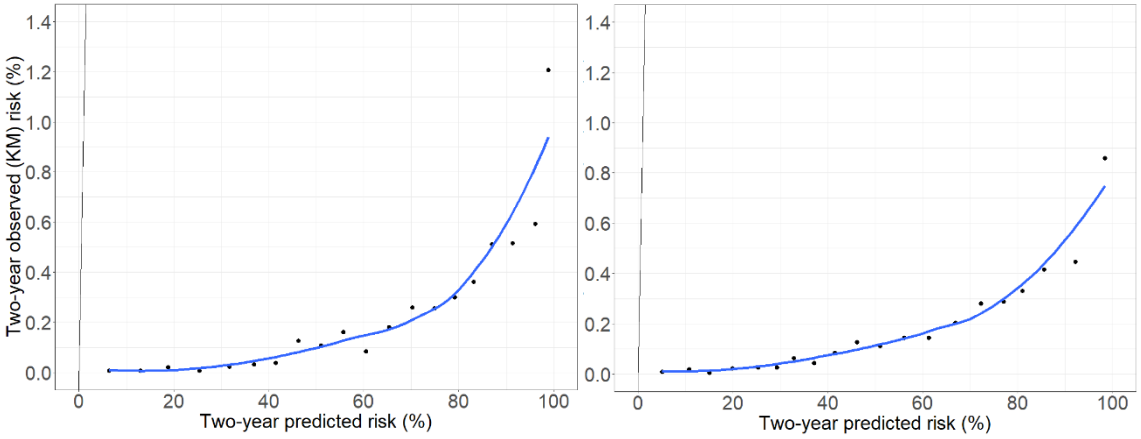
74%	0.2999%	743	110784	316964	457	61.92	74.10	0.67	99.86
75%	0.3064%	732	106505	321243	468	61.00	75.10	0.68	99.85
76%	0.3126%	710	102238	325510	490	59.17	76.10	0.69	99.85
77%	0.3191%	700	97958	329790	500	58.33	77.10	0.71	99.85
78%	0.3256%	684	93685	334063	516	57.00	78.10	0.72	99.85
79%	0.3322%	663	89416	338332	537	55.25	79.10	0.74	99.84
80%	0.3387%	637	85153	342595	563	53.08	80.09	0.74	99.84
81%	0.3453%	605	80895	346853	595	50.42	81.09	0.74	99.83
82%	0.3521%	595	76616	351132	605	49.58	82.09	0.77	99.83
83%	0.3593%	571	72350	355398	629	47.58	83.09	0.78	99.82
84%	0.3667%	538	68094	359654	662	44.83	84.08	0.78	99.82
85%	0.3741%	521	63822	363926	679	43.42	85.08	0.81	99.81
86%	0.3820%	510	59543	368205	690	42.50	86.08	0.85	99.81
87%	0.3905%	480	55284	372464	720	40.00	87.08	0.86	99.81
88%	0.3994%	455	51019	376729	745	37.92	88.07	0.88	99.80
89%	0.4092%	428	46757	380991	772	35.67	89.07	0.91	99.80
90%	0.4194%	400	42495	385253	800	33.33	90.07	0.93	99.79
91%	0.4305%	382	38224	389524	818	31.83	91.06	0.99	99.79
92%	0.4429%	353	33963	393785	847	29.42	92.06	1.03	99.79
93%	0.4572%	314	29713	398035	886	26.17	93.05	1.05	99.78
94%	0.4741%	282	25455	402293	918	23.50	94.05	1.10	99.77
95%	0.4948%	242	21206	406542	958	20.17	95.04	1.13	99.76
96%	0.5211%	201	16957	410791	999	16.75	96.04	1.17	99.76
97%	0.5572%	167	12702	415046	1033	13.92	97.03	1.30	99.75
98%	0.6108%	127	8452	419296	1073	10.58	98.02	1.48	99.74
99%	0.7187%	76	4214	423534	1124	6.33	99.01	1.77	99.74

Abbreviations: PPV=positive predictive value; NPV=negative predictive value

D7. Calibration of the ColonFlag machine-learning algorithm

Figure D9 shows that calibration for a monotone score cannot be assessed. This is because the score does not reflect absolute risk so cannot be compared to observed risks to assess model accuracy. For example, Figure D9 shows the mean of the predicted ColonFlag score for each 20th of predicted risk, but this is a monotone score of 0-100, hence the predictions do not coincide with the observed risks. Calibration cannot therefore be assessed.

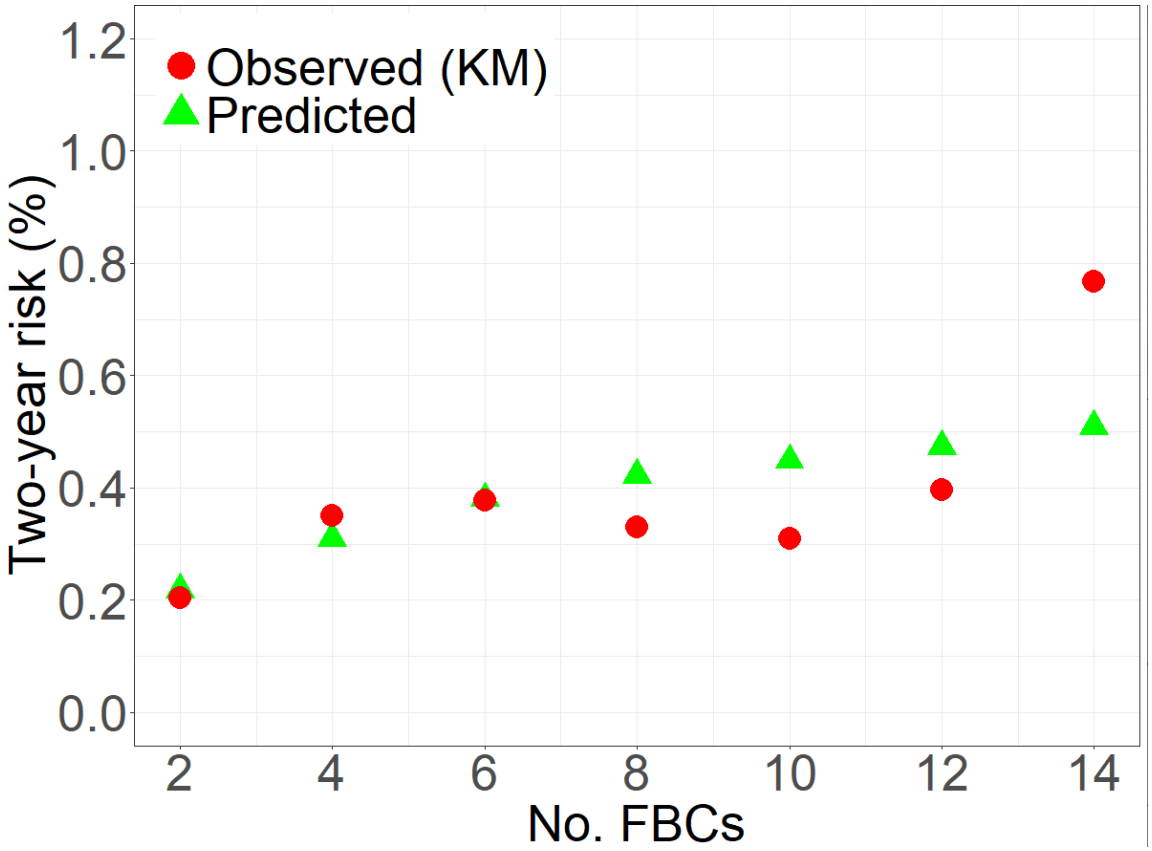
Figure D9: Calibration plot for the ColonFlag algorithm in males (left) and females (right) in the split sample internal validation cohort



Abbreviations: KM=Kaplan-Meier

D8. Calibration by number of FBCs for males in the development cohort

Figure D10: Calibration plot for the joint models by number of FBCs for males in the development cohort



Appendix E

Appendices for Chapter 6 (discussion)

E1. How this research help me: a personal note

- **My first grant application**: my DPhil is NIHR funded. Although I was involved in some sections of other grant applications from other researchers, this was the first grant application I had ever fully developed and led.

- **Led a systematic review**: I have been involved in many systematic reviews but this DPhil gave me the opportunity to perform one myself. I led all aspects of the review.

- **Led my own team of researchers**: the systematic review also gave me the opportunity to build and lead my own small team of researchers to assist in the work. I managed my own timelines but also my team's timelines and workload.

- **PPI advisors**: this research introduced me to PPI, which I had never experienced beforehand. As statisticians, opportunities to build and lead PPI representatives is rare but I was fortunate to have had this opportunity from this DPhil. I learnt how to approach PPI representatives, the importance of patient involvement, and how to present complex study designs and advanced statistical methods using lay terminology.

- **EHR/routine data**: I had some experience of working with EHR data prior to this research. However, my experience and skillset developed significantly in this DPhil. I was involved in the data acquisition, preparation, and analysis of EHR data, which I led and managed predominantly independently (with appropriate oversight by my supervisors and collaborators).

- **Prediction modelling and joint models**: I had no practical experience of developing prediction models at the highest technical level. This research gave me the opportunity to develop and test my own models. I also learnt about joint models, such as how they compare to other modelling approaches.

- **Advanced computers and technology**: the use of big data and advanced statistical modelling were computationally burdensome. I explored and learnt of many different ways to analyse big data and perform computationally intensive analyses and learnt about advanced computer systems.

- **Programming skills**: I am proficient in the use of Stata software and used it before and throughout this DPhil. I had some knowledge of R software before this DPhil commenced but by actively using this software to develop and validate the model, I have developed skills in R programming and become a better statistical programmer.

- **Handling missing data**: I had used multiple imputation in prior research studies to account for missing data, which was often advised by senior statisticians. This DPhil gave me the opportunities to develop a fundamental understanding of the data available, consider different approaches to handle missing data in practice, and consider when it is appropriate to use them.

- **Build collaborations**: I was able to work with clinicians, haematologists, members of the public, and data programmers outside of my direct research team. I therefore built important collaborations and have plans to work with these collaborators in future research.

- **Gaining international recognition**: I have met with a company in the US a few times over the last two years, who contacted me. This company is interested in this research and would like to work together in the future.

- **Enhanced publication record**: I had publications in peer-reviewed journals prior to this DPhil but the DPhil gave me opportunities to perform and publish my own research. I currently have three publications from this research, with a further two planned.

- **Conference proceedings**: I have been given many opportunities to present my own research at national conferences. I presented eight research abstracts from this research at six different conferences.

- **Ideas for future research**: this DPhil has highlighted many areas for future work. I will aim to pursue these areas in practice and perform my own research. These opportunities have evolved from this research.

- **Training courses**: historically, I would only go to training courses that were relevant to my projects. This DPhil ensured I had a training programme that would not only relate to this research, but to my own career development too.

- **General skill development**: all of the individual items listed above (and more) have developed my skillset and confidence in many ways. These skills are essential for independent researchers. Through this DPhil, I am now better equipped to perform my own research in the future.