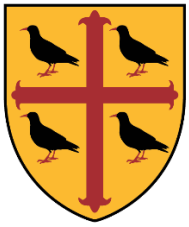
Big drug data for big drug trials: development, validation, and implementation of routinely-collected data on medications within large trials in cardiovascular disease and COVID-19 in England and Scotland

Oxford Population Health

Dr Guilherme Pessoa-Amorim, MD

St Edmund Hall, University of Oxford





Thesis submitted for the degree of

Doctor of Philosophy

2024

Nihil nisi amor

How much has happened, in the world and in my life, since I was admitted to this DPhil program back in 2020. A global pandemic, a wedding, a home. A new life born. A dearly loved one lost. Laughs and tears, together occasionally. And a broken leg (advice often given, but seldom followed). This work is dedicated to my family, in particular my wife and best friend Ana, who was there throughout it all, kept me afloat when I needed most, and is a permanent source of love, comfort, and purpose. To our baby son Eduardo, the most precious gift one could ask for, and a daily reminder of what really matters. To my brother and my parents, whose unwavering love and support has been my bedrock for so many years. Nothing in life is worth achieving without you in it.

Last but not least, this thesis is submitted in loving memory of José Neto Pessoa de Amorim Rosa. He might have been satisfied.

Acknowledgements

I wish to start by thanking the contribution of the many thousands of participants in the RECOVERY, LENS, and ORION-4 trials. Neither those trials, nor the work presented here, would have been possible without them. Similarly, I wish to acknowledge the contributions of the NHS and research staff who have worked to deliver these trials, and those whose daily work caring for NHS patients ultimately led to the generation of the data sources assessed within this thesis. I would also like to thank the support and advice received from staff at NHS England (NHSE) and Public Data Scotland on the use of these datasets.

Secondly, I am deeply grateful for the support and advice given by my supervisors: Professor Louise Bowman, Professor Sir Martin Landray, and Dr Will Stevens. The path leading from the start of my DPhil to submission of this thesis was long, hard, and seldom clear. None of it would have been possible without your guidance and your commitment throughout, as well as the confidence you have always shown in my capacity to deliver this project. I am particularly grateful to Professor Bowman for the kind and generous support that she has always granted me, ever since I first moved to Oxford.

Third, the work presented here was also made possible or significantly enhanced by the contribution of multiple investigators and trial staff at Oxford Population Health. In particular, I wish to thank Professor Richard Haynes, Associate Professor Marion Mafham, Associate Professor David Preiss, and Associate Professor Natalie Staplin for their comments on this thesis, as well as Rijo Kurien, Michelle Nunn, and Charles Crichton for their support with trial data management. A particular word also for my dear colleagues Dr Charlie Harper and Dr Emily Sammons for their encouragement, guidance, and support.

Finally, I want to thank the Medical Research Council through its Population Health Research Unit at Oxford Population Health, whose studentship has supported my studies.

To all of you, I am deeply grateful.

Abstract

Background

Randomised controlled trials (RCTs) are the established gold-standard for the reliable assessment of medical interventions. However, trials have become increasingly complex and costly, in part due to inefficient trial designs or data collection and verification procedures. Routinely-collected healthcare data (RCD) are frequently used in observational studies, and may contribute to streamlining RCTs, but widespread use of these data within trials requires appropriate validation and methodological development. This thesis aimed to: 1) develop the methodology required to implement novel or little used sources of RCD on medications in the United Kingdom (UK) within 3 distinct trials in Coronavirus Disease (COVID-19) and cardiovascular disease; 2) ascertain whether those RCD are appropriate to identify drug exposure for trial participants; and 3) test the impact of using these data within potential use-case scenarios in each trial.

Methods

First, a scoping literature review was performed in PubMed to gather previous studies comparing RCD on medications with other sources of the same information, across either RCTs or observational studies. The aim of this review was to describe the settings, populations, and medication groups considered in previous validation studies, as well as the methodologies used to extract drug exposure from RCD, the statistical methods employed to compare different sources of data, and the results of direct comparisons between sources.

The original analyses presented here involved the use of four distinct datasets within three large randomised controlled trials in England and Scotland. Specifically, these were:

1) the General Practice Extraction Service Data for Pandemic Planning and Research (GDPPR, or "GP") dataset (which includes nationwide primary care records in England) and the Medicines Dispensed in the Community ("Dispensing") dataset (a nationwide collection of data on medications dispensed in the community in England and submitted for reimbursement), both used to identify medication exposure in the community among 36,202 participants recruited in England within the Randomised Evaluation of COVID-19 treatment (RECOVERY) trial of treatments for patients hospitalised with COVID-19;

2) the Electronic Prescribing and Medicines Administration (EPMA) dataset (which includes data on medication use from a subset of secondary care hospitals in England), used to identify in-hospital medication exposure among 4,313 RECOVERY trial participants recruited in England;

3) the Prescribing Information Services (PIS) dataset (which captures medication dispensing in the community setting in Scotland), used to identify community drug exposure for 1,125 participants in the Lowering Events in Non-proliferative retinopathy in Scotland (LENS) trial of fenofibrate to prevent diabetic retinopathy disease progression; and

4) the Dispensing dataset (as described above), used to identify community drug exposure among 10,020 participants recruited in England within the ORION-4 trial of inclisiran for secondary cardiovascular disease prevention.

In each trial, data quality assessments were performed by investigating data features including completeness (the degree of capture of the intended trial population and medications of interest), and accuracy/agreement with medication data as collected in each trial's case report form (CRF) (the latter generated from medical records review in RECOVERY, and participant self-report in LENS and ORION-4). To guide use of these

datasets, methodological development exercises were performed to investigate which possible drug exposure derivation algorithms yielded the best agreement between RCD and the trial CRF (used as reference). These included 1) choosing an observation window to identify drug exposure at randomisation from RCD, 2) handling RCD records in the month of randomisation (for datasets that did not provide exact event dates), and 3) testing the accuracy of drug doses recorded in RCD. Finally, the impact of using RCD in each trial was tested by comparing the estimates obtained using from RCD versus those from the CRF on specific use-case scenarios. The main measure of agreement used was Cohen's Kappa coefficient (κ), complemented with sensitivity, specificity, and positive and negative predictive values (PPV/NPV).

Results

The scoping review identified a total of 192 references comparing RCD on medications with other sources, published up until March 2023, 18 of which within RCTs and 174 within observational studies. These studies spanned a wide variety of countries, study populations, and medication groups. Claims/pharmacy data was the most common RCD source used, and self-reported drug exposure the most common comparator source (against RCD). RCD were the most common reference chosen for comparisons between data sources. The median (interquartile range [IQR]) aggregate percent agreement across all studies was 0.87 (0.7-0.95), κ 0.68 (0.46-0.82), sensitivity 0.79 (0.63-0.9), specificity 0.98 (0.92-0.99), PPV 0.82 (0.68-0.92), and NPV 0.96 (0.88-0.99). Agreement was generally similar across the different RCD sources used, but varied greatly with geographical location and population or medication groups. Comparisons performed within RCTs exhibited generally higher agreement results when compared with those within observational studies, and most studies had overall positive agreement results (as interpreted by their authors).

Within the RECOVERY trial, 91% (32,768/36,202) of participants recruited in England were captured in both the GP and Dispensing datasets, and 98% (35,344/36,202) in at least one. Both datasets captured most of a wide range of 17 medication groups of interest. A 3-month lookback period to identify drug exposure at randomisation maximised agreement between CRF and RCD for medications commonly used in the community setting (κ=0.76 for antiplatelets; κ=0.67 for oral anticoagulants). Compared with Dispensing data only, inclusion of GP data improved agreement with CRF marginally (κ: 0.75 to 0.76 for antiplatelets; κ: 0.68 to 0.69 for oral anticoagulants). RCD yielded lower estimates for drug exposure at randomisation versus CRF (antiplatelets: 14.6% RCD vs 16.3% CRF; oral anticoagulants: 8.0% RCD vs 9.9% CRF).

Secondary-care medications data included in the EPMA dataset captured a representative sample of 10.6% (4,313/40,860) of RECOVERY participants recruited in England. Overall agreement between EPMA and the trial CRF was high for in-hospital use of a group of medications of interest to the trial, at both randomisation and during follow-up (κ=0.72, PPV=75%, NPV=95% for baseline exposure at randomisation across all medications combined; κ=0.77, PPV=85%, NPV=95% for follow-up exposure). Both sources yielded similar overall drug exposure prevalence estimates. Low sensitivity was the major contributor for discrepancies in most drugs considered (baseline: 74%-92%; follow-up: 19%-92%), except for corticosteroids (low specificity; baseline: 55%, follow-up: 68%), with important variations in agreement over time (between March 2020 and May 2022) and by respiratory status at baseline.

In the LENS trial, PIS data captured >99% (1,147/1,151) of randomised participants. Including records in the month of randomisation provided very high agreement with CRF data for 18 drugs of interest across all possible lookback windows (κ=0.90-0.92), with a marginal increase if also including records in the month after randomisation (κ=0.92-0.93).

Statin doses taken from PIS data and those recorded in the randomisation CRF agreed in 98-100% of cases (depending on the specific statin). Using an observation window including 3-months pre-randomisation and the month of randomisation, agreement between PIS and CRF for baseline exposure was high (κ>0.80) across most drugs studied (except ezetimibe: κ=0.71). Sensitivity, specificity, and predictive values were high (approximately >90% for most drugs). PIS and CRF data generated similar overall estimates of drug exposure at baseline, and for initiation of prohibited drugs during follow-up. Integration of PIS data complemented the randomisation CRF by capturing initiation of non-prohibited medications post-randomisation (which are not captured in the follow-up CRF).

Finally, the Dispensing dataset captured >99% (10,020/10,038) of ORION-4 participants recruited in England. Using Dispensing records in the previous 3 calendar months to define baseline drug exposure at randomisation provided very high agreement with CRF records (κ=0.93 across all drug groups combined), with slight improvement if also including records in the month of randomisation (κ=0.94). Daily statin dose recorded in the CRF agreed with the most recent previous Dispensing statin strength record in 85-90% of cases (across different statins and visits). Overall agreement between CRF and Dispensing data across all drugs groups combined was high across all study visits (κ=0.94-0.96, with sensitivity, specificity, and predictive values generally >95%). However, Dispensing data failed to adequately capture exposure to a small number of participants using two concomitant medications prohibited by the trial protocol (non-study inclisiran and proprotein convertase subtilisin/kexin type 9 inhibitors), but these were not expected to be captured in Dispensing data as they are not usually managed in the community setting. Dispensing data yielded generally similar or slightly higher estimates for medication use prevalence at randomisation and during follow-up when compared with the CRF, while

agreeing with the CRF for statin treatment intensity at randomisation for 85% of participants.

Conclusions

In summary, results from this thesis show that RCD on medication use in the community setting in England and Scotland provide excellent coverage of participants in large trials of COVID-19 and cardiovascular disease, and yield high agreement for drug exposure estimates when compared with existing data collection methods. Data on in-hospital medication use had limited population coverage, and issues with both sensitivity and specificity. Expanded use of the community data sources may help streamline trial data collection, so that more trials can be conducted, at faster speed, and lower cost, as long as the limitations of using RCD are acknowledged and mitigated. Future work will be required to validate the use of RCD in distinct trial settings and for other medication groups, but the results presented here strongly support their integration within future cardiovascular UK trials.

Statement of contribution to work

The work presented in this thesis is my own. It has been supervised by Professor Louise Bowman, Professor Sir Martin Landray, and Dr Will Stevens. I have independently planned and conducted all data management, analysis, visualisation, interpretation, and writing. All medication codelists have been produced by myself, and validated by either Dr Mark Campbell, Dr Waseem Karsan, or Dr Nathan Samuel.

I was not involved in the design of the RECOVERY, LENS, or ORION-4 trials, but was involved in conducting RECOVERY and ORION-4 and collecting data in both trials. I also led the application for data linkage within the ORION-4 study, which allowed access to the NHSE dataset assessed within that trial. Besides my supervisors, the work presented in this thesis has been informed by comments and advice received from Professor Richard Haynes (for analyses concerning the RECOVERY trial), Associate Professor Marion Mafham (for analyses concerning the RECOVERY and ORION-4 trials), Associate Professor David Preiss (for analyses concerning the LENS and ORION-4 trials), and Associate Professor Natalie Staplin (across all chapters). I have also received advice from Professor Will Herrington as part of my Transfer and Confirmation assessments.

I am aware of and understand the University's policy on plagiarism. I certify that this thesis is my own work, except where indicated by references. The work presented in it has not been submitted in support of another degree or qualification from this or any other university or institute of learning.

Funding

This DPhil was funded by a DPhil studentship awarded by the Medical Research Council Population Health Research Unit, at the Oxford Population Health Department.

The RECOVERY trial (ISRCTN number: 50189673, ClinicalTrials.gov number: NCT04381936) was supported by grants to the University of Oxford from UK Research and Innovation and the National Institute for Health and Care Research (NIHR), the Wellcome Trust through the COVID-19 Therapeutics Accelerator, the Oxford BHF Centre of Research Excellence, and Health Data Research UK, and by core funding provided by the NIHR Oxford Biomedical Research Centre, the Wellcome Trust, the Bill and Melinda Gates Foundation, the Foreign, Commonwealth and Development Office, the Medical Research Council Population Health Research Unit, the NIHR Health Protection Unit in Emerging and Zoonotic Infections and NIHR Clinical Trials Unit Support Funding.

The LENS trial (ISRCTN number: 15073006, ClinicalTrials.gov number: NCT03439345, European Union Clinical Trials registry number: 2016-002656-24) is funded by the NIHR.

The ORION-4 trial (ISRCTN number: 88876914, ClinicalTrials.gov number: NCT03705234) is funded by Novartis (previously The Medicines Company).

1 Table of contents

Acknowledgements iii

Abstract iv

Background iv

Methods iv

Results vi

Conclusions ix

Statement of contribution to work x

Funding xi

Table of contents xii

List of abbreviations and acronyms xxi

List of tables xxiv

List of figures xxviii

CHAPTER 1: Introduction 1

1.1 Chapter summary 1

1.2 Randomised trials in Medicine – general concepts and contemporary challenges 3

1.3 Routinely-collected healthcare data in clinical research 4

1.4 Medication data collection and reporting in trials 7

1.5 Routinely collected medications data – previous uses in clinical research, and a novel landscape within the UK 12
1.6 Overall thesis aims 16
1.7 Thesis outline 17

CHAPTER 2: Validation of routinely-collected medications data – a literature review . 19

2.1 Chapter summary 19
2.2 Introduction 22
2.3 Methods 25
2.3.1 Review scope 25
2.3.2 Search strategy 26
2.3.3 Data collection 31
2.3.4 Statistical analysis, data, and code 32
2.4 Results 33
2.4.1 Included studies – geographical settings, populations, and medications 33
2.4.2 Included studies – RCD sources, comparator sources, and reference used . 38
2.4.3 Included studies – drug exposure methodology 40
2.4.4 Agreement metrics and results 42
2.5 Discussion 57
2.5.1 Limitations 62
2.6 Conclusion 63

CHAPTER 3: Routinely-collected community prescribing and dispensing data in England: validation and implementation in the RECOVERY trial of COVID-19 treatments among hospitalised patients 65

3.1 Chapter summary 65

3.2 Introduction 68

3.3 Methods 69

3.3.1 Study population, design, and setting 69

3.3.2 Data collection 69

3.3.2.1 Case report form data 69

3.3.2.2 Routinely-collected data sources 70

3.3.3 Data analysis 76

3.3.3.1 Cohort characteristics 76

3.3.3.2 Data quality assessment 77

3.3.3.3 Data-driven methodology development 79

3.3.3.4 Use-case scenarios 82

3.3.3.5 Analysis periods 83

3.3.3.6 Statistical analyses 83

3.3.3.7 Data and code 84

3.3.4 Ethics 84

3.4 Results 85

3.4.1 Study population 85

3.4.2 Data quality assessment 85

3.4.2.1 Linkage coverage....85
3.4.2.2 Drug coverage....90
3.4.2.3 Timeliness and stability....92
3.4.3 Data-driven methodology development....94
3.4.3.1 Lookback windows....94
3.4.3.2 Handling Dispensing record dates....100
3.4.3.3 Combining different RCD sources....102
3.4.4 Use-case scenarios....104
3.5 Discussion....112
3.5.1 Limitations....117
3.6 Conclusion....119

CHAPTER 4: Routinely-collected nationwide inpatient medications data in England – validation and implementation within the RECOVERY trial of COVID-19 treatments....120

4.1 Chapter summary....120
4.2 Introduction....122
4.3 Methods....123
4.3.1 Study population, design, and setting....123
4.3.2 Data collection....123
4.3.2.1 RECOVERY case report form data....123
4.3.2.2 EPMA dataset....127
4.3.3 Data analysis....132
4.3.3.1 Data quality – completeness/coverage....132

4.3.3.2 Data quality – stability and timeliness 133
4.3.3.3 Data methodology development 133
4.3.3.4 Agreement between EPMA and CRF data 134
4.3.3.5 Use-case scenarios 135
4.3.3.6 Statistical analyses 135
4.3.3.7 Data and code 137
4.3.4 Ethics 138
4.4 Results 139
4.4.1 Data quality assessments 139
4.4.1.1 Study cohort 139
4.4.1.2 Drug coverage 142
4.4.1.3 Timeliness and stability 145
4.4.2 Data-driven methodology development 148
4.4.3 Agreement between EPMA and CRF data 155
4.4.3.1 Binary baseline and follow-up drug exposure 155
4.4.3.2 Quantitative follow-up exposure 163
Use-case scenarios 165
4.4.4 165
4.5 Discussion 169
4.5.1 Limitations 175
4.6 Conclusion 177

CHAPTER 5: Routinely-collected nationwide Scottish medications data - development, validation, and application in the LENS trial of fenofibrate in diabetic retinopathy 178

5.1 Chapter summary 178

5.2 Introduction 180

5.3 Methods 182

5.3.1 Study population and time period 182

5.3.2 Medication data sources 182

5.3.2.1 Case report forms 182

5.3.2.2 PIS data 182

5.3.3 Data analysis 185

5.3.3.1 Data quality - coverage and timeliness 185

5.3.3.2 Drug exposure methodology development 186

5.3.3.3 Data quality - Agreement between PIS and CRF data 189

5.3.3.4 Use-case scenarios 189

5.3.3.5 Statistical analyses 190

5.3.3.6 Data and code 190

5.3.4 Ethics 191

5.4 Results 192

5.4.1 Study population 192

5.4.1.1 Data coverage and timeliness 192

5.4.2 Methodology development 194

5.4.2.1 Lookback windows and handling peri-randomisation records 194
5.4.2.2 Imputing dispensing dates 200
5.4.2.3 Validating PIS drug doses 201
5.4.3 Agreement between PIS and CRF data 202
5.4.4 Use-case scenarios 206
5.5 Discussion 210
5.5.1 Limitations 214
5.6 Conclusion 215

CHAPTER 6: Validation, development, and implementation of routinely-collected data on medications in England in a phase-III, regulatory submission trial of inclisiran for secondary cardiovascular disease prevention (ORION-4) 216

6.1 Chapter summary 216
6.2 Introduction 219
6.3 Methods 221
6.3.1 Study population and time period 221
6.3.2 Medication data sources 221
6.3.2.1 Case report forms 221
6.3.2.2 Medicines Dispensed in the Community dataset 223
6.3.3 Data analysis 224
6.3.3.1 Data quality – coverage and timeliness 225
6.3.3.2 Drug exposure methodology development 225
6.3.3.3 Data quality – agreement between RCD and CRF 228

6.3.3.4 Use-case scenarios 228
6.3.3.5 Statistical analyses 228
6.3.3.6 Data and code availability 229
6.3.4 Ethics 229
Results 230
6.3.5 Study population 230
6.3.6 Drug coverage and timeliness 233
6.3.7 Methodology development 235
6.3.7.1 Lookback windows and handling of peri-randomisation records 235
6.3.7.2 Handling exact dates for drug initiation during follow-up 242
6.3.7.3 Validating statin drug doses from RCD records 244
6.3.8 Agreement for drug exposure between Dispensing data and case report form (CRF) 246
6.3.8.1 Binary drug exposure at each visit 246
6.3.8.2 Time to drug initiation after randomisation 251
6.3.9 Use-case scenarios 254
6.4 Discussion 259
6.4.1 Limitations 264
6.5 Conclusion 265
CHAPTER 7: Overall discussion 267
7.1 Chapter summary 267

7.2 Introduction 267

7.3 Main findings 268

7.3.1 Previous studies 268

7.3.2 Data on medication use in the community within the RECOVERY trial 269

7.3.3 Inpatient medication data in the RECOVERY trial 271

7.3.4 Community dispensing data in Scotland in the LENS trial 272

7.3.5 Community dispensing data in England in the ORION-4 trial 272

7.4 Strengths and limitations 273

7.5 Implications for future clinical trials 277

7.6 Implications for future research 280

7.7 Implications for data controllers and other stakeholders 281

7.8 Final summary 285

References 287

1

List of abbreviations and acronyms

ACE - Angiotensin-converting enzyme

ARB - Angiotensin receptor blockers

BNF - British National Formulary

CCB - calcium channel blockers

CHI - Community Health Index

CI – confidence interval

CM – concomitant medication

CONSORT - Consolidated Standards of Reporting Trials

COVID-19 – Coronavirus Disease

CPRD - Clinical Practice Research Datalink

CRF - case report form

dm+d - Dictionary of Medicine and Devices

DPP-4 - Dipeptydil peptidase-4

ECMO - extracorporeal membrane oxygenation

EHR - electronic health record

EMA – European Medicines Agency

EPMA - Electronic Prescribing and Medicines Administration

EPS - Electronic Prescription Service

EU - European Union

FDA – Food and Drug Agency

GDPPR - General Practice Extraction Service (GPES) Data for Pandemic Planning and Research dataset

GLP-1 - Glucagon-like peptide 1

GPES - General Practice Extraction Service

GP – general practitioner

HES – Hospital Episode Statistics

ICC - intraclass correlation coefficient

IMV – invasive mechanical ventilation

IQR – interquartile range

ISD - Information Services Division

κ - Cohen's Kappa coefficient

LENS – Lowering Events in Non-proliferative retinopathy in Scotland

LMIC - low-and-middle-income countries

LMWH - low-molecular weight heparins

LOA – limits of agreement

MeSH - Medical Subject Header

MHRA – Medicines and Healthcare products Regulatory Agency

NHS – National Health Service

NHSBSA - National Health Service Business Services Authority

NHSE – National Health Service England

NIV - non-invasive ventilation

NOACs - Non-vitamin K oral anticoagulants

NPV – negative predictive value

PABAK - prevalence-and-bias-adjusted-kappa

PCSK9 - Proprotein convertase subtilisin/kexin 9

PIS - Prescribing Information Service

PPV – positive predictive value

Q1 – quartile 1

Q3 – quartile 3

RAAS - Renin-angiotensin-aldosterone system

RCT – randomised controlled trial

RCD – routinely-collected data

RECOVERY – Randomised Evaluation of COVID-19 Therapy

RWE – real-world evidence

SAIL - Secure Anonymised Information Linkage

SD – standard deviation

SEARCH - Study of the Effectiveness of Additional Reductions in Cholesterol and Homocysteine

SGLT2 - Sodium glucose transporter 2

Sn - sensitivity

SNOMED - Systematized Nomenclature of Medicine Clinical Terms

Sp - specificity

UK – United Kingdom

US – United States

WOSCOPS - West of Scotland Coronary Prevention Study

List of tables

Table 2.1 - Illustrative search strategy and results for first review round 26

Table 2.2 - Tentative systematic search strategy 29

Table 2.3 - Characteristics of included studies: size and geographical location 36

Table 2.4 - Characteristics of included studies: population and medication group 37

Table 2.5 - Characteristics of included studies: RCD source, comparator source, and reference used 39

Table 2.6 - Characteristics of included studies: drug exposure and analytical method 42

Table 2.7 - Characteristics of included studies: agreement metrics and interpretation 43

Table 3.1 - Drug groups of interest to RECOVERY selected for analysis 71

Table 3.2 - GP dataset structure 73

Table 3.3 - Dispensing dataset structure 75

Table 3.4 - Baseline cohort characteristics, by linkage status 86

Table 3.5 - Agreement for baseline drug exposure between case report form (CRF) and routinely-collected data (RCD) using different lookback periods before randomisation 99

Table 3.6 - Agreement for baseline drug exposure between case report form (CRF) and routinely-collected data (RCD) using different lookback periods before randomisation 101

Table 3.7 - Impact of using different data sources on the agreement for baseline antiplatelet and oral anticoagulant exposure between case report form (CRF) and routinely-collected data (RCD) 103

Table 3.8- Aggregate estimates for different drug exposures at randomisation by RCD source 109

Table 3.9 - Agreement metrics for different drug initiations after randomisation between the GP and Dispensing datasets 111

Table 4.1 - RECOVERY case report form drug exposure collection (for drugs used in this analysis) 125

Table 4.2 - Drugs and exposure metrics selected for analysis 126

Table 4.3 - Number of participants and records included in each EPMA data table 128

Table 4.4 - Overview of EPMA data tables 128

Table 4.5 - EPMA data tables and fields selected for analysis 130

Table 4.6 - Baseline cohort characteristics for participants with vs without EPMA data 140

Table 4.7 - Sensitivity analysis: administration table date fields 149

Table 4.8 - Sensitivity analysis: prescription vs administration tables (combined and by drug) 150

Table 4.9 - Sensitivity analysis: prescription type (combined and by drug) 151

Table 4.10 - Sensitivity analysis: prescription status (combined and by drug) 152

Table 4.11 - Sensitivity analysis: prescription formulation (combined and by drug) 153

Table 4.12 - Sensitivity analysis: handling of data recorded in the day of randomisation (combined and by drug) 154

Table 4.13 - Agreement between EPMA and CRF data for baseline exposure 155

Table 4.14 – Agreement metrics between EPMA and case report form (CRF) data for follow-up exposure 156

Table 4.15 - Agreement metrics between EPMA and case report form (CRF) data for baseline exposure, stratified by respiratory status at baseline 158

Table 4.16 - Agreement metrics between EPMA and case report form (CRF) for follow-up exposure, stratified by respiratory status at baseline 159

Table 5.1 - Medications included in the LENS PIS data extraction..... 183

Table 5.2 - PIS data structure..... 184

Table 5.3 - Drugs select for analysis..... 184

Table 5.4 - Baseline characteristics of LENS participants, split by PIS data linkage status 192

Table 5.5 - Performance of different rules to derive baseline drug exposure from PIS data vs case report form (CRF) - lookback periods and handling peri-randomisation records 197

Table 5.6 - Performance of different rules to derive baseline drug exposure from PIS data vs case report form (CRF) – imputing dispensing dates 200

Table 5.7 - Agreement between PIS data and case report form (CRF) for baseline exposure 203

Table 6.1 - Drug groups of interest to ORION-4 included for analysis 223

Table 6.2 - Baseline characteristics of ORION-4 participants recruited in England and submitted in data linkage request, split by Dispensing dataset linkage status 231

Table 6.3 - Performance of different rules to derive baseline drug exposure from Dispensing data when compared with self-reported exposure at the randomisation visit 239

Table 6.4 - Agreement between Dispensing data and case report form (CRF) for drug exposure at each visit type, across all drug groups 246

Table 6.5 - Agreement for drug exposure at screening between Dispensing data and case report form (CRF), by drug group 247

Table 6.6 - Agreement for drug exposure at randomisation between Dispensing data and case report form (CRF), by drug group 248

Table 6.7 - Agreement for drug exposure at any point during follow-up between Dispensing data and case report form (CRF), by drug group 249

Table 6.8 - Agreement for drug exposure at each follow-up visit between Dispensing data and case report form (CRF), by drug group 250

Table 6.9 - Agreement estimates for time to drug initiation after randomisation, by drug 253

Table 6.10 - Differences in aggregate drug exposure between routinely-collected data (RCD) and case report form (CRF), by study visit and drug group 256

List of figures

Figure 2.1 - Literature review flowchart 34

Figure 2.2 - Schematic illustration of the fixed window method for deriving drug exposure using RCD 41

Figure 2.3 - Schematic illustration of the legend-time method for deriving drug exposure using RCD 41

Figure 2.4 - Aggregated statistical metrics of agreement for comparisons on drug exposure assessments between RCD and comparator sources by study type in the literature review 45

Figure 2.5 – Study-specific interpretation of agreement metrics for comparisons on drug exposure assessments between RCD and comparator sources 46

Figure 2.6 – Linear correlation between statistical agreement metrics and study size (i.e. number of participants) for comparisons on drug exposure assessments between RCD and comparator sources 47

Figure 2.7 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by geographical location 49

Figure 2.8 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by population group 50

Figure 2.9 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by medication group 51

Figure 2.10 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by RCD source type 53

Figure 2.11 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by comparator data source type 55

Figure 2.12 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by drug exposure outcome and analysis method.....56

Figure 3.1 - Number of distinct drug codes per participant, split by drug group and dataset 81

Figure 3.2 – Study population linkage coverage and overlap between GP and Dispensing datasets 87

Figure 3.3 - Geographical distribution of RECOVERY participants, by linkage status.....88

Figure 3.4 - Monthly randomisation counts of RECOVERY participants, split by dataset linkage status.....89

Figure 3.5 - Capture of broad (BNF chapters) and specific drug groups in the linkage datasets 91

Figure 3.6 - Monthly counts of distinct participants, entries, and drug codes per dataset across consecutive data extracts 93

Figure 3.7 - Average time interval between consecutive drugs records in each dataset .95

Figure 3.8 – Cumulative proportions of participants with different average time intervals between consecutive drug records, by dataset 97

Figure 3.9 – Aggregate estimates for drug exposure at randomisation by ascertainment source (antiplatelets and oral anticoagulants).....106

Figure 3.10 - Overlap between case report form (CRF) and routinely-collected data (RCD) sources for antiplatelet and oral anticoagulant exposure at randomisation at an individual participant level.....107

Figure 3.11 - Aggregate estimates for different drug exposures at randomisation by routinely-collected data source.....108

Figure 3.12 - Aggregate estimates for drug initiation after randomisation, by drug group and RCD source 110

Figure 4.1 - Recruitment timeline for monthly counts of RECOVERY participants with vs without available EPMA data 141

Figure 4.2 - Proportion of participants with available EPMA data within each RECOVERY site 141

Figure 4.3 - Drug capture in EPMA data across BNF chapters 143

Figure 4.4 - Drug capture in EPMA data for specific drugs 144

Figure 4.5 - Time series of weekly record counts for each date field (record date and reported date), split by Administration and Prescription tables 146

Figure 4.6 - EPMA data stability across consecutive extracts (split by Prescription vs Administration tables) 147

Figure 4.7 - Sensitivity and specificity over time for binary baseline and follow-up drug exposure assessments, stratified by drug group 161

Figure 4.8 - Number of participants identified as baseline or follow-up exposure according to EPMA, case report form (CRF), or either data source 162

Figure 4.9 - Number of days of follow-up drug exposure according to EPMA vs case report form (CRF) data 163

Figure 4.10 - Proportion of days (for aspirin) or number of doses (for tocilizumab/sarilumab) received during follow-up according to EPMA vs case report form (CRF) data 164

Figure 4.11 - Use-case scenario 1: Monitoring of case report form (CRF) drug recording across study sites (using EPMA as reference) 166

Figure 4.12 - Use-case scenario 2: Use of select RECOVERY trial medications during the admission of randomisation over time according to EPMA vs case report form (CRF) data, by level of respiratory support at randomisation 168

Figure 5.1 - Baseline drug exposure derivation rule logic 188

Figure 5.2 - Cumulative monthly counts of dispensing records and randomisations over time 193

Figure 5.3 - Density plots for average time interval in days between consecutive entries of the same BNF code for each participant, per drug group 194

Figure 5.4 – Cumulative proportions of LENS participants with distinct average time intervals in days between consecutive records of the same BNF code in PIS data, by drug group 195

Figure 5.5 - Impact of different derivation rules on identification of drug exposure as baseline vs post-randomisation initiation, using PIS data vs case report form (CRF) 199

Figure 5.6 - Statin dose distributions (at randomisation) 201

Figure 5.7 - Comparison of statin dose recorded in the randomisation case report form (CRF) vs statin strength in PIS data 202

Figure 5.8 - Classification of discrepancies between case report form (CRF) and PIS data for baseline drug exposure 205

Figure 5.9 - Baseline exposure to specific drug groups in PIS data vs case report form (CRF) 207

Figure 5.10 - Post-randomisation initiation of prohibited drugs in PIS data vs case report form (CRF) 207

Figure 5.11 - Changes to drug exposure after randomisation captured by PIS data if using case report form (CRF) as baseline source 209

Figure 6.1 – Number of monthly randomisations over time, split by linkage status..... 232

Figure 6.2 - Recruitment location of the study population, split by linkage status..... 232

Figure 6.3 - Capture of drug groups of interest in the Dispensing dataset 233

Figure 6.4 - Histogram of historic Dispensing data available before randomisation for each participant 234

Figure 6.5 - Histograms of average intervals between consecutive records of the same BNF code for each participant, by drug group 236

Figure 6.6 - Cumulative proportions of ORION-4 participants with different average time intervals between consecutive records of the same BNF code, by drug group 237

Figure 6.7 - Impact of different derivation rules on identification of drug exposure as baseline vs post-randomisation initiation from Dispensing data vs case report form (CRF) 241

Figure 6.8 - Comparison of estimates for time to drug initiation since randomisation from case report form (CRF) vs Dispensing data, across combinations of possible date handling rules 243

Figure 6.9 - Statin dose distributions from case report form (CRF) data, by statin and study visit 244

Figure 6.10 - Comparison of statin dose recording in case report form (CRF) vs Dispensing data, by statin and study visit 245

Figure 6.11 – Individual estimates of time to drug initiation after randomisation according to Dispensing data and case report form (CRF), by drug 252

Figure 6.12 – Absolute differences in aggregate drug exposure estimates between routinely-collected data (RCD) and case report form (CRF), by study visit and drug group 255

Figure 6.13 - Comparison of statin intolerance characterisation between routinely-collected data (RCD) and screening case report form (CRF) 257

Figure 6.14 - Comparison of statin treatment intensity exposure at randomisation between routinely-collected data (RCD) and case report form (CRF) 258

CHAPTER 1: Introduction

1.1 Chapter summary

Randomised controlled trials (RCTs) are considered the gold standard approach for assessing safety and efficacy of medical interventions. Nonetheless, due to rising costs, overly complex designs, and regulatory bureaucracy, trials are becoming increasingly difficult to conduct, limiting the deployment of new therapies and fuelling calls for expanded use of less reliable observational approaches.

Routinely-collected data (RCD) are data generated as a by-product of standard clinical care. RCD have been repurposed for research for decades, as they generally contain the information required (demographics, diagnostics, treatment, and clinical outcomes) and are cheaper to use than data collected for a specific research purpose. However, the use of RCD within trials has been hindered by limited availability, and concerns as to whether these sources would be of sufficient quality to meet regulatory standards.

In the UK, some sources of RCD have been available for research purposes for a number of years, and often used within trials. However, these RCD sources did not contain any information on medication use, which is a crucial component of data collection in many RCTs. On the other hand, a number of research-ready populations (such as the Clinical Practice Research Datalink [CPRD]) have been built using multiple RCD sources, including prescriptions, but do not currently allow linkage to an external trial cohort.

Since 2020, three new data sources on medications in England (primary care records, medications dispensed in the community, and secondary care prescribing and administration) have become available for linkage to randomised cohorts. In Scotland, a similar dataset has been available since 2009, but its use in trials has been extremely limited. This evolving landscape opens new opportunities for integration of RCD on

medications into UK trials, potentially improving their efficiency, speed, and data completeness.

Nevertheless, widespread use of these new sources requires appropriate validation (namely in comparison with established sources of data collection in trials), particularly if these data are to be used for future submissions to regulatory agencies, and the development of source-specific data derivation methods.

This thesis aims to validate, and develop the methodology required for, the use of nationwide RCD sources on medication across England and Scotland within three UK RCTs, and to test the impact of integrating these data sources in specific use-case scenarios.

1.2 Randomised trials in Medicine – general concepts and contemporary challenges

Randomised trials are the established gold-standard for the assessment of efficacy and safety of medical interventions.[1,2] RCTs study the effect of a particular intervention in comparison with a control by making use of randomisation (i.e. allocation of study interventions at random) to untether any links between patient characteristics and treatment selection, thus removing the effect of confounding. Randomisation permits unbiased estimates of the causal relationship between interventions and their effects by yielding populations that are, on average, comparable with respect to all measured and unmeasured, known and unknown confounders – the only remaining difference being the treatment each group is allocated to. This simple yet remarkable feature separates RCTs from all other kinds of non-randomised studies, placing them at the top of the classical pyramid of medical evidence (surpassed only by meta-analyses of randomised trials).[2] RCTs have been instrumental in the massive reductions observed in major causes of morbidity and mortality (namely cardiovascular disease) since the mid-20th century, and have been recognised as one of the "10 health advances that changed the world".[3] The transformative role of randomisation in causal inference has also been acknowledged with a Nobel Prize in Economics in 2019 to investigators who applied this principle to developmental economics, transforming it "from a subjective into an empirical discipline".[4] The first RCT recognisable as such is generally considered to be the 1948 UK Medical Research Council trial of streptomycin for tuberculosis, although mentions of similar controlled experiments can be found as far back as the Bible, or in James Lind's study of treatments for scurvy in the 18th century.[5] Throughout the 20th and 21st century, the number of RCTs being conducted has risen exponentially, so that over 25,000 trials were being published annually by 2019.[6] However, similar trends can been seen in trial complexity

and costs over recent decades, sparking calls to "rescue" the feasibility of running trials in the United States (US) and elsewhere.[7–10] Increasing trial complexity and costs can be attributed to secular trends of reductions in cardiovascular and cancer mortality rates (so that larger or longer trials are required),[11,12] but also the adoption of overly complex study designs, excessively restrictive eligibility criteria, inappropriate trial bureaucracy, and inefficiencies in data collection and verification.[13–15]

1.3 Routinely-collected healthcare data in clinical research

Appropriate data collection in trials is of paramount importance to ensure participant safety, generate reliable results, and avoid resource waste,[16–18] and is highlighted by international guidelines as a critical component of good clinical practice.[19–21] This is particularly relevant for trials supporting submission of data on investigational products for regulatory approval, where previous evidence on efficacy and safety may be limited (particularly in broad population groups). As a result, trial data are usually collected for purpose, during long periods, and with a high level of detail, contributing to high trial costs and complexity.[13,22] In part, this can be explained by overly strict interpretations of clinical trials regulations, which increase bureaucracy without necessarily improving trial quality or safety.[14] Moreover, much of these data end up not being used, contributing simultaneously to high drug development costs and resource waste.[23]

Routinely-collected data (RCD) (or *real-world data)* are those generated as a result of the delivery of standard healthcare (through clinical records, insurance claims, or disease-specific registries), and collected from existing sources such as clinical databases.[24] Harnessing RCD has been proposed as a potential solution to some contemporary challenges of running trials,[10,25] fuelling calls for expanded use of "*real-world evidence*"

(RWE) – that is, observational analyses performed using RCD - to replace randomisation.[26–28] Although RCD are not generated to address any specific scientific question, these data represent information about demographics, diagnoses, treatments, and clinical events, and are therefore similar to those used in research.[29] Moreover, such data have long been used for research purposes, such as in epidemiological and drug safety studies.[30,31] The potential value of integrating RCD in trials is also being increasingly recognised, particularly through the endorsement of regulatory agencies including the United States (US) Food and Drug Agency (FDA), the United Kingdom (UK) Medicines and Healthcare products Regulatory Agency (MHRA), the European Medicines Agency (EMA), the Chinese National Medical Products Administration, and the Japanese Pharmaceutical and Medical Devices Agency, as well as political authorities such as the European Union (EU) and the UK government.[32–37] As an example, the FDA has recently granted label extensions for drugs based solely on RWE, and is adopting the use of "synthetic control arms" to replace RCT-level evidence in specific areas.[38,39] Simultaneously, the FDA is promoting the conduct of pragmatic clinical trials within the nationwide pharmacovigilance Sentinel system, including the ADAPTABLE trial of different aspirin dosing regimens (published in 2021).[40,41] The EU is also developing the EHR4CR programme, which, among other goals, aims to foster the identification of patients for clinical trials using RCD.[42] On the other hand, claims-based medication adherence metrics have been adopted by the Centers for Medicare and Medicaid Services in the US as part of their Five-Star quality rating program since 2008.[43] Altogether, these initiatives show that major international regulatory agencies and government bodies consider RCD to be potentially valuable and adequate for medical research purposes, with evidence to suggest broader support for their use from trial participants, patients in general, and the wider public.[44–46]

The use of RCD for trials offers numerous potential advantages,[32,47] such as: 1) reduced data collection costs,[48–50] 2) reduced burden to patients and staff,[51,52] 3) faster, more accurate, and more efficient recruitment;[50,53] 4) improved data coverage, diversity, and integrity (including real-time collection to inform interim safety and futility assessments),[54–56] 5) expanded research agenda due to broader data availability,[32] and 6) focus on clinically relevant outcomes (i.e. those captured by standard pathways) and thus improved generalisability.[50] RCD may enhance all stages of the trial delivery process by reducing resource waste, improving efficiency, and more closely integrating RCTs within the routine care pathways they aim to inform.[23,51] Overall, integration of RCD within RCTs may help tackle many of the current drug development challenges, such as high trial costs and difficulty in conducting studies of appropriate size or duration.[13,15,32]

In spite of the many potential benefits, and recent efforts aiming to validate their use within RCTs,[42,57,58] RCD are only being used in a minority of trials globally, and mostly restricted to studies assessing technological or behavioural interventions (rather than medications or procedures).[49,59,60] In the UK, it has been estimated that only approximately 5% of trials are using RCD (with similar figures globally), despite half of publicly-funded UK trials declaring at the grant stage that they expected to harness them.[61,62] Potential causes for this attrition include some specific challenges associated with the use of RCD, such as: 1) data availability, access, or timeliness; 2) data complexity, size, and diversity; 3) data accuracy, completeness, and documentation; and 4) novelty of the data analyses pipelines.[32] There is, then, ongoing uncertainty as to whether RCD (in their many forms) are appropriate for use as a primary (or complementary) source of data for clinical trials, particularly given the experimental nature of RCTs and the associated complex regulatory framework.[63] This is especially the case for trials of new or repurposed drugs aiming to

support subsequent regulatory submissions, which entail complex and thorough data auditing and verification procedures.

1.4 Medication data collection and reporting in trials

RCD are currently only used in a minority of trials in the UK, but are increasingly being adopted for this purpose (especially since the beginning of the COVID-19 pandemic),[33,61] including in regulatory submissions.[51] A longstanding obstacle to more extensive harnessing of RCD in the UK has been the limited availability of broader types of data beyond secondary care records, disease registries, and death records.[33,61,64] A noteworthy example is medications data. Accurate reporting of both study treatment adherence and use of concomitant medications (CMs; i.e. drugs taken by a trial participant in addition to the investigational drug under study) is a vital component of appropriate clinical trial data collection, as highlighted by different international guidelines.[19,20] Information on study treatment use is critical to assess adherence to the trial protocol, monitor safety, and interpret the study results.[65] Similarly, data on CMs may be used in trials for several purposes, namely selecting and describing the trial population, safety monitoring and reporting, and interpretation of results.[66–70]

Background medical therapy is a common part of trial eligibility, given the potential need to avoid undesirable drug interactions or excessive symbiotic effects.[71] In 2007, a review focused in six leading medical journals found that 54% of published RCTs had at least one medication-related exclusion criterion.[66] CMs are also important to describe standard of care in the study population (e.g. disease-modifying drugs in heart-failure trials, or steroids in studies of hospitalised COVID-19 patients)[72] as recommended by the CONSORT guidelines,[73] and to determine baseline disease severity (e.g. defining severe asthma based on oral steroid prescriptions).[74]

Adequate reporting of safety events requires a detailed description of all CMs (either chronic or acute) being used or recently stopped at the time of the event under investigation, which helps to inform causal assessments in relation to the trial intervention. As the trial progresses, background therapy may also determine whether the study medication is still indicated or if it should be discontinued (for example, if a contra-indicated drug or one of the same class is initiated). Moreover, CM modifications or additions may be indicators of important adverse events, such as increased need for diuretics in heart failure (which may reflect reduced exercise tolerance), initiation of glucose-lowering medications (a surrogate for new-onset diabetes), or antibacterial prescriptions (a possible surrogate for iatrogenic infections).

Finally, adequate interpretation of trial results may benefit from or require accurate and complete recording of CMs at baseline and over time, as these might influence drug safety or efficacy.[75] Such modifications of treatment effects may occur through direct or indirect interactions. Direct modifications occur due to drug-to-drug interactions. Importantly, these might be unexpected, such as the safety concern found with concomitant use of simvastatin and amiodarone in the Study of the Effectiveness of Additional Reductions in Cholesterol and Homocysteine (SEARCH) trial, highlighting the importance of thorough recording of all CMs.[70] Indirect effects may occur when both CMs and the new intervention target similar biological pathways, possibly yielding marginal added benefits (e.g. background statins in trials of lipid-lowering drugs), or independently increasing the risk of a particular adverse event (e.g. bleeding in antithrombotic treatment trials, and immunosuppression in COVID-19).[76–80]

Adequate CM recording may involve particular challenges, since trial participants are often exposed to at least one other drug upon entering the trial,[71] frequently many simultaneously, and each with different dosing and regimens.[81] This is especially the case

in trials in the inpatient setting, where participants may be exposed to larger number of drugs than they would in the community setting, and with more complex regimens. Moreover, CMs may change in number, dosing, or frequency during the study (particularly in long trials of chronic conditions), and drug information collected at baseline might poorly reflect actual drug consumption during the follow-up period.[82] Changes to CMs may occur through either drop-in (i.e. starting or intensifying treatment) or drop-out (stopping treatment or lowering dose), and as a result of progressive disease, intended study treatment effects, advances in standard-of-care, or known or unknown adverse events. If occurring to different degrees in each arm (i.e. differential drop-in/drop-out), CM changes may influence the estimated differences in surrogate endpoints and clinical events between study groups.[83,84] For example, differential drop-in to glucose-lowering medications may reduce the observed effect size in trials in diabetes (where it is desired that glycaemic levels are controlled below a certain threshold), and similarly in trials of anti-hypertensive drugs.[69] Similarly, different levels of adherence to maintenance inhaler therapy can explain a substantial part of the variance seen in surrogate outcomes of treatment efficacy in asthma add-on treatment trials (namely respiratory function). In this context, measuring adherence over time and using it to assess between-group differences could lead to important improvements in trial statistical power.[85,86] The same could be argued for other drugs and conditions (e.g. use of statins, cholesterol levels, and atherosclerotic events in cardiovascular disease trials). Importantly, increased adherence to maintenance therapy as a result of trial participation could explain some of the substantial improvements in respiratory function seen in placebo groups in some RCTs.[87,88] Nonetheless, exact starting or stopping dates may be difficult to pinpoint, complicating adherence estimates, safety reviews, and the assessment of potential impact of CM drop-

in/drop-out in trial outcomes.[69] Together, these examples highlight the need to adequately record CMs, not only at the beginning, but also throughout the duration of the trial.

Whilst medication data collection may be streamlined for drugs in which the safety profile is already established and accepted, assessment of unlicensed medications requires complete and comprehensive CM recording (as both efficacy and safety may be influenced by unexpected interactions).[70] Even when studying an already licensed drug, the importance of a given CM in relation to treatment safety or efficacy might also not always be apparent upfront.[67,70] Possibly owing to these complex challenges, CMs have been noted to be poorly reported across RCTs at all stages of the drug development process, and in different fields, limiting adequate trial interpretation and subsequent translation to practice (e.g. when the level of standard-of-care received is not clear).[68,84,89–93]

Methods used for collection of study treatment adherence and CM data in trials are varied, with no established gold-standard.[65,94,95] These may include direct observation, self-report, pill counts, electronic monitors (such as pill cartridges or inhalers), blood drug levels, medical record review, clinician assessments, or harnessing of RCD sources.[96–98] It has been estimated that most RCTs use self-report as the preferred method to compile background medication lists, making it the *de facto* standard (at least for trials in the outpatient/community setting).[99] However, the use of this subjective approach (both within and outside trials) has several known limitations. Firstly, it can be resource-consuming, particularly for participants taking a large number of medications, or in trials with long follow-up periods or frequent study visits.[100] Secondly, it can either underestimate (due to recall bias or unwillingness to disclose sensitive medications, such as antidepressants or antipsychotics)[98,100–105] or overestimate adherence to study medications (due to social desirability bias – i.e. a desire to please researchers by overstating adherence).[106] Lastly, CM recall and drug adherence metrics may vary depending on the specific method and

tools used to interview patients, for example resulting from the use of open- vs close-ended questions (in interviews or questionnaires), resorting to memory prompts, and asking participants to present current medication lists, boxes, or bottles.[100,102,107,108] As an example, in a 2021 study including 315 patients upon admission to a single US hospital, only a third were able to provide accurate medication lists when compared with manual review of local pharmacy dispensing records within the last 3 months (which was considered the gold-standard source in that case). Surprisingly, the average number of errors increased if patients were asked to use pill bottles (2.3 errors) or medication lists (1.8) as an aid memoir, vs from memory alone (1.2).[109]

Accordingly, in epidemiological and drug surveillance studies, it is generally accepted that self-report is not the most reliable approach to identify drug exposure.[41,110,111] Instead, objective measures including RCD sources (such as claims databases and prescription registries), blood drug levels, pill counts, or electronic pill bottle monitors are commonly used to validate self-reported estimates. Nonetheless, subjective measures such as self-report may well capture true medication intake patterns that may be missed by objective measures (e.g. medication which was purchased and is held at home but was not taken, or use of over-the-counter or non-reimbursed medications). Moreover, while generally considered more accurate, it has been observed that objective measures may still be subject to bias, especially if the individual under study knows that they are being monitored (for example, by putting away unused pills, opening electronic cartridges repeatedly to make up for previous non-adherence, or increasing medication adherence shortly before having bloods drawn for drug level analyses).[96,112–114] Thus, different data sources may well capture distinct aspects of the same ground truth, with no single source considered perfectly complete or accurate (and therefore a true gold-standard).[110,115–117] In

summary, triangulation between different sources is likely to provide the most reliable estimates, potentially by using distinct sources for distinct purposes.[118–121]

1.5 Routinely collected medications data – previous uses in clinical research, and a novel landscape within the UK

Medicines are the most common healthcare intervention, and a prominent feature of any medical record.[122] RCD on medications can be generated at different stages of the care delivery process, namely when a clinician prescribes a drug, a patient collects it from a pharmacy, or it is administered in-hospital. These events may then be submitted through reimbursement/claims or insurance systems, and captured in claims databases or prescription registries. Alternatively, medications may be recorded in registries focused on specific diseases or treatments (such as those used for cancer or vaccinations), or when recorded as part of regular medication lists in primary care or hospital records.[114,123]

RCD on medications has been used in medical research for decades,[30] namely by repurposing pharmacy claims data (particularly in the US)[115,124,125] or those from national disease registries (most notably those in the Scandinavian countries).[31,126] However, their use has mostly been limited to observational studies. More recently, medications data have also been incorporated in RCTs to identify eligible patients, collect information on CMs, or as a trial endpoint (such as in studies aiming to improve prescription or adherence patterns).[127–131] Nevertheless, most long-term follow-up extensions of RCTs incorporating RCD have not been able to collect information on medications. [132] Such information is required to assess use of open-label study treatment and other medications, which are particularly relevant for interpretation of the observed differences after the controlled treatment period is over. Moreover, despite extensive efforts at validating the use of RCD

sources on medications within observational studies (which have yielded mixed results, and with large heterogeneity between drug classes),[133–136] few studies have undertaken similar assessments within RCTs, and have either included small numbers of individuals or focused on limited sets of medications.[115,137–140]

The unified structure of the UK National Health Service (NHS) and the work performed by institutions such as NHS England (NHSE; previously NHS Digital), Public Health Scotland, and the Secure Anonymised Information Linkage (SAIL) Databank in Wales facilitate the aggregation and standardisation of large amounts of data generated as part of clinical care from multiple individual sources, with established pathways for access, and at a relatively low cost (although timeliness may sometimes be problematic).[140] This centralised structure helps to overcome many of the challenges associated with the use of RCD, such as interoperability, linkage costs, and data harmonisation.[32] NHS data have often been used in medical research, including in trials,[141] with recent examples including feasibility assessments,[142–144] identification and recruitment of potential participants,[142–144] communication with participants,[145] and follow-up data collection and outcome adjudication.[145–147]

Until recently, nationwide RCD available from NHSE and other organisations in England has mostly concerned demographics, death records, and clinical diagnoses from secondary care records or disease registries. In parallel, some research collaborations (such as CPRD, QResearch, and the Royal College of General Practitioners Research and Surveillance Centre) have collated primary care records from large numbers of patients.[148–150] These data are extremely valuable for observational studies. Of note, a small number of pragmatic trials have been performed via recruitment and data collection within CPRD and similar networks.[131,151–154] However, such data have been of limited use for RCTs since they do not provide nationwide coverage, thus failing to deliver complete follow-up data

collection for broader trial cohorts. Similarly, specific databases such as the Hospital Treatment Insights (which links secondary care records from Hospital Episode Statistics [HES] to inpatient pharmacy records in 43 out of 153 hospital trusts in England) have been found to be unreliable, due to issues with linkage or incomplete capture of medication records (namely antibiotics).[155]

In 2020, NHSE have released new datasets for research purposes, including:

1. The General Practice Extraction Service (GPES) Data for Pandemic Planning and Research dataset (GDPPR; hereby referred to as the Prescribing or GP dataset), comprising primary care data (which includes drug prescriptions);[156]
2. The Medicines Dispensed in the Community dataset (referred to as the Dispensing dataset), which includes dispensing records used for reimbursement purposes by the NHS Business Services Authority;[157] and
3. The Electronic Prescribing and Medicines Administration (EPMA) dataset, which collates secondary care prescriptions from a limited number of NHS Trusts in England.[158]

Although some of these data are already in use for drug safety and regulatory purposes,[29] and in some instances for medical research,[159–161] this was the first time that the full scope of the data collected became available for linkage to external cohorts at the individual record level.

In Scotland, similar RCD on medications are collected and maintained by the Information Services Division (ISD) within Public Health Scotland. ISD provide a number of datasets for research purposes, including data on community dispensing records in the Prescribing Information Service (PIS) dataset.[162] PIS includes aggregated medications data since 1993, and individual level records since 2009. These data have been harnessed in multiple

observational studies, including for drug safety research.[163] In a 2018 study, self-reported use of 8 classes of cardiovascular, psychiatry, and gynaecological drugs from >10,000 participants in the Generation Scotland cohort was compared with linked PIS data, with the results showing very good agreement for most drugs.[164] Nonetheless, and in contrast with other Scottish RCD sources,[165] use of PIS data within RCTs has been restricted to two pragmatic trials of feedback interventions to improve prescribing safety. This observation suggests a need for further validation to support implementation in broader RCT settings, including within drug intervention trials.[166,167]

Similarly to other types of RCD, medications data may help address some of the challenges facing contemporary trials by making them cheaper, faster, and more efficient. Specific use-cases for RCD on medications in trials include identification of potentially-eligible participants for clinical trials (e.g. based on current medication), phenotyping and selection of patients (e.g. defining diabetes mellitus by long-term prescriptions of antidiabetic drugs, or identifying hypertensive or asthmatic patients who are truly treatment-resistant despite appropriate adherence), and collection of follow-up data on both study treatment and CM use.[69,115,131,168,169] On a wider perspective, identifying medication use from RCD also allows research into prescription and adherence patterns at the population-level, where other sources of such information (such as self-report, pill counts, or drug levels) would be too troublesome or costly to collect.

Although these novel sources bear great promise for collecting medications data in trials, several questions need to be answered before widespread adoption.[24] Specifically, these concern 1) assessing data quality and appropriateness, 2) developing the correct methodology to derive medication exposure, and 3) establishing the feasibility and impact of integrating these data in specific analyses within trials.[32,170] Importantly, as described above, RCD on medications may take many forms (such as claims data, primary care

records, or prescription registries), with the content, structure, and quality of each data source depending on the specific data collection and reporting pathways. Likewise, recent FDA and MHRA guidance documents on the use of RCD or electronic health records for clinical research state that each data source must be assessed to determine whether it is fit for its intended use in the specific setting considered.[24,32,36,117,171] As an example, primary care data is generally considered to provide the best overall picture of an individual's health, but may miss a substantial proportion of diagnoses or medications.[56,172,173] Similarly, some drugs may be underrepresented in RCD sources due to prescription or reimbursement patterns, or remain present in the data despite having been stopped or switched.[24,174–177] Thus, adequate and purpose-specific data quality assessments are warranted, with both regulatory agencies suggesting validation against a relevant reference data source (such as that collected as part of trials), in line with standard practice for assessing data quality of medical records.[24,36,117] Such validation efforts, together with showcasing of successful uses within concrete settings, can provide valuable information, increase trust, and foster subsequent and more widespread RCD use within trials.[33]

1.6 Overall thesis aims

In summary, there is a novel opportunity to harness the evolving landscape of routinely-collected data on medications in the UK to help deliver better, faster, more efficient randomised trials. This thesis will address this knowledge gap by validating the use of four different RCD sources of medications in England and Scotland within three distinct RCTs, developing the data analysis methods required for their integration, and testing their implementation in specific use-case scenarios relevant to those trials, with the overarching objective of determining the value and feasibility of using RCD on medications for UK trials.

1.7 Thesis outline

This thesis will consist of 8 chapters, as outlined below:

1. The present Chapter 1 provides an introduction to the overall background for this thesis and the research question addressed;
2. Chapter 2 depicts a scoping literature review addressing the settings, methodology, and results of previous research studies comparing RCD against collateral sources of information on drug exposure;
3. Chapter 3 investigates the use of two novel RCD sources (GP and Dispensing datasets) on outpatient medications in England within a trial of treatments for hospitalised COVID-19 patients (Randomised Evaluation of COVID-19 Therapy - RECOVERY), in which only limited data on outpatient medications was available from study case report forms (CRF);
4. Chapter 4 explores the use of a novel RCD source on inpatient medications in England (EPMA dataset) within the RECOVERY trial in comparison with CRF data (for in-hospital medication use);
5. Chapter 5 addresses the use of an existing RCD source on medication dispensing in the community in Scotland (PIS dataset) within a trial of fenofibrate in diabetic patients (Lowering Events in Non-proliferative retinopathy in Scotland - LENS) when compared with self-reported medications;
6. Chapter 6 assesses the value of using the Dispensing dataset when compared with self-reported medications in a regulatory-submission trial of inclisiran, a new cholesterol-lowering medication for secondary cardiovascular disease prevention (ORION-4);

7. Chapter 7 concludes this thesis with a discussion of its overall results, contribution to the existing literature, future directions, and recommendations for relevant stakeholders.

CHAPTER 2: Validation of routinely-collected medications data – a literature review

2.1 Chapter summary

Aims: To describe and summarise the settings, methodology, and results of previous studies comparing RCD on medications with parallel sources.

Methods: A scoping review was performed in PubMed by combining free literature searches (i.e. exploratory searches based on relevant keywords) with footnote chasing in reference lists and the "Similar Articles" and "Cited By" features. Initial free literature searches were performed in July-August 2021, updated in January 2022, and again updated and expanded with footnote chasing in January-March 2023. This review included studies comparing any RCD source, geographical location, population, medication group, analytical method, and statistical approach. Relevant studies were grouped by research setting (RCTs vs observational studies). All searches were performed in PubMed. Agreement results were aggregated using medians and interquartile ranges (IQR).

Results: In total, approximately 37,500 titles were screened over the multiple review rounds performed (approximately 21,000 identified from the free searches, and 16,423 from the footnote chain searches). This resulted in 192 relevant studies identified, including 18 within RCTs and 174 within observational studies. Most studies were conducted in the US (67% and 37% of RCTs and observational studies, respectively), Canada (5.6% and 11%), Scandinavia (11% and 21%), Australia (11% and 4.6%), and the Netherlands (0% and 7.5%). The most common study populations identified were population cohorts (28% and 33%), followed by cardiovascular disease (39% and 14%), and oncology (5.6% and 13%). For medication groups, the most common were cardiovascular (22% and 17%),

endocrinology (10% and 10%), and mental health (14% and 8.5%). Claims/pharmacy data were the most common RCD sources used (100% and 67%), and self-reported drug exposure the most common comparator source (76% and 56%). RCD were the most common reference chosen for comparisons against other sources (39% and 37%), followed by studies where both sources under study were considered equivalent (33% and 16%), and those where medical records were used as reference (0% and 18%). Various possible drug exposure metrics were identified, with agreement assessed using multiple different statistics, including percent agreement, Cohen's Kappa coefficients (κ), sensitivity/specificity (Sn/Sp), and positive/negative predictive values (PPV/NPV). Agreement results varied greatly with geographical location, population and medication group, but specificity and NPVs were generally high across all settings. Agreement was generally similar across the different RCD sources used, albeit lower for primary care data, but there was large heterogeneity based on the parallel data source used for comparison against RCD. The median (IQR) aggregate percent agreement across all studies was 0.87 (0.7-0.95), κ 0.68 (0.46-0.82), sensitivity 0.79 (0.63-0.9), specificity 0.98 (0.92-0.99), PPV 0.82 (0.68-0.92), NPV 0.96 (0.88-0.99), and correlation 0.3 (0.17-0.72). Comparisons performed within RCTs exhibited generally higher agreement results when compared with those within observational studies. Most studies had overall positive agreement results as specified by individual study authors (56% and 71%, vs negative: 28% and 14%, and mixed: 17% and 14%), with large heterogeneity in agreement results included in each category (e.g. percent agreement: 0.87 [0.75-0.95] within studies with positive assessment of agreement, vs 0.58 [0.22-0.85] within those with negative assessments; sensitivity: 0.79 [0.66-0.90] vs 0.41 [0.25-0.72] respectively).

Conclusions:

RCD have been compared against collateral sources on medications data across multiple geographies, populations, and medication groups, within both RCTs and observational studies. The methodologies used and results obtained were varied, with important heterogeneity across different settings and sources. RCD were the reference standard most commonly used for comparisons. Agreement between sources was generally considered to be positive, but interpretation of agreement results was heterogeneous across different studies.

2.2 Introduction

RCD may offer great value to help streamline clinical research, while also potentially improving data quality. Sources of RCD on medications (such as claims data or prescription registries) are increasingly common.[178–181] Such data been used extensively in clinical research, particularly in observational research focused on drug safety, health economics, or medication adherence studies.[114,123,179,182]

Nonetheless, adoption of RCD within RCTs has been limited. This may be due to specific challenges associated with the use of RCD, such as cost, restrictive access, protracted timeframes, regulatory and ethical constraints, and uncertain data quality.[63,171,183,184] Some of these limitations can potentially be addressed retrospectively after data is collected, including through changes to access policies or pathways to allow dissemination of previously stored data.[185–187] However, using RCD inherently entails repurposing of data for a different purpose than the one they were collected for. Therefore, overcoming data quality issues retrospectively to comply with the requirements of RCTs may be more troublesome, particularly given their complex regulatory environment. Moreover, as RCD was not collected for research purposes, and without focusing on a particular research question, concerns have been raised that they may not have been collected with the same care as research data collected for a particular aim such as a randomised trial.[171] Conversely, it has been argued that data recorded by clinical staff working daily on a specific disease area, or for billing purposes, might indeed be of better quality than that captured by research staff, particularly if the data recorded by research staff is based on subjective sources (such as participant self-report).[188–190]

Data quality has been defined in many different ways.[191,192] A natural, yet stringent, definition is that data quality depends on whether the data perfectly represent the truth they aim to capture. However, the need for perfect data should also be considered in light

of the associated trade-offs in terms of practicality, speed, and cost, and these may differ from one setting to the other. In a 2013 systematic review on electronic health record (EHR) data quality dimensions and assessment methods, Weiskopf and Weng found that quality data could be articulated as "*data are of sufficient quality when they serve the needs of a given user pursuing specific goals*".[192] This formulation follows a broad, commonly used definition of quality as "*fitness-for-purpose*",[193] thus entailing that data quality dimensions and definitions are not absolute but rather context-specific. Recently, the FDA and the MHRA have issued specific guidance on the use of RCD or EHR for clinical research.[24,36] Both agencies have used similar definitions to Weiskopf and Weng, respectively stating that "*each data source should be evaluated to determine whether the available information is appropriate for addressing a specific study hypothesis*", and "*data quality processes must be appropriately validated to ensure that any process is fit for its intended use*". In sum, data quality is relative, and dependent on the specific purpose for which a given source may be employed.

In order to establish data quality, the FDA recommend that "*validation ideally involves a comparison of the exposure classification in the proposed data source with a reference data source*". Accordingly, Weiskopf and Weng found that the most commonly used approach for data quality validation was comparison with a gold-standard or reference source. Thus, implementation of RCD sources on medications within trials requires adequate validation of the quality of those data in the specific context in which they may be used (i.e. geographical setting, population, medication exposure), by comparing it to a gold-standard or reference source on that same information. This will be the general aim of this thesis, with Chapters 3-6 presenting results from direct data quality assessments performed across multiple trial settings.

In order to guide subsequent data analysis and interpretation, the present chapter depicts the results of a literature review focused on 1) describing previous comparisons of RCD sources on medication against collateral data sources, and 2) summarising results from relevant studies.

2.3 Methods

2.3.1 Review scope

This literature review focuses on methodological studies comparing RCD sources vs any other source on medication use (including subjective sources such as self-, clinician-, or relative-reported medication, and objective sources such as medical record review, pill counts, electronic monitoring, or blood drug levels). RCD were considered as databases or registries that collated information from EHR or other sources of clinical information (i.e. excluding manual review/abstraction of the EHR itself, which was included as a possible comparator source).[24] This included data sources encompassing primary care records, prescription or dispensing records, claims data, disease registries, etc. Characteristics of interest included the geographical settings, populations, and medications groups considered, the methodology used for drug exposure derivation and statistical analysis, and the RCD sources and comparators considered.

There were no restrictions in terms of geographical or clinical setting, disease, medication, drug exposure derivation methodology, or statistical approach used to assess agreement between sources. Given the remit of this thesis, the search was initially restricted to RCTs. It was later expanded to include observational studies as it was considered that results from comparative studies in this setting could also be informative.

This chapter followed the framework of a scoping review - which aims to describe a large body of literature and present an overview which informs further research - as opposed to a systematic review – which aims to systematise and aggregate results from a smaller number of studies focused on a particular research question.[194,195] The scoping review framework was chosen here given the specific aim of describing a vast number of heterogeneous studies to guide subsequent research within this thesis, rather than ensuring complete retrieval of all relevant studies focused on a particular topic. At the

same time, use of a systematic search strategy would lead to identification of a prohibitively high number of potential references for review given the broad scope in terms of medications, population groups, methodologies, and research settings (as further detailed below).

2.3.2 Search strategy

A preliminary, exploratory literature review was performed in July-August 2021 with the goal of retrieving methodological studies performed within the context of RCTs.

Table 2.1 shows an illustrative list of the string searches performed and respective results. All searches were conducted in PubMed only.

Table 2.1 - Illustrative search strategy and results for first review round

String search	Date search performed	Number of studies screened (by title)
"concomitant medication clinical trials systematic reviews"	10th July 2021	262
"((trials) OR (clinical trials) OR trial OR randomised OR randomized) AND (drug OR medication OR drugs OR medications)) AND (concomitant OR non-trial OR non-study)) AND (reporting OR source OR collection)"	11th July 2021	5,522
"drug data trials reporting concomitant"	11th July 2021	1,217
"systematic review adherence comparison medication"	27th July 2021	416
"medication database claims comparison"	27th July 2021	1,175

As this search approach yielded a very limited number of relevant studies conducted within RCTs (n=2, as selected after full text review), it was then expanded in January 2022 to include methodological studies comparing RCD with other data sources in any research

context (i.e. including observational studies outside the scope of trials). Search strings were simultaneously broadened to include combinations of the keywords “adherence”, “compliance”, “self-report”, “medication”, “drug”, “comparison”, “validation”, “claims”, and “prescription”, among others. Additional studies were also collected from reference lists and PubMed “Related articles” and “Cited by” features (from relevant articles). This expanded search strategy yielded a total of 60 relevant studies - Figure 2.1.

An update to this search was initiated in January 2023, with the initial aim of extending this review by developing a systematic search strategy. First, the list of 60 articles mentioned above was updated on the 26th January 2023 and the 8th February 2023 by repeating the free searches described above, and also by retrieving additional records from reference lists. This update resulted in 43 additional studies identified, and 6 studies being removed from the original list after further review, yielding a total of 97 relevant studies identified at this stage - Figure 2.1.

In accordance with previous reports,[196] during this process it was identified that no combination of keywords could reliably identify all relevant studies already retrieved (from the free literature searches). Potentially useful Medical Subject Header (MeSH) terms were also only included recently, namely “Routinely Collected Health Data” (introduced in 2021), “Data Accuracy” (2016), “Database” (2008), “Databases, Pharmaceutical” (2013), and “Validation Studies as Topic” (2008)”. A systematic review search query was nonetheless developed based on a combination of previous similar systematic review queries and relevant MeSH terms. Candidate MeSH terms were identified through a search of all PubMed MeSH terms using the keywords “Medication”, “Drug”, “Validation”, “Data quality”, “Database”, and “Routinely collected data”, performed on the 27th February 2023.[60,96,182,197–200] Given the broad aim of this literature review, and similarly to previous reports,[194,196,201,202] this search did not yield a reasonable limitation to the number of studies

identified, resulting in 94,103 titles requiring review. Conversely, a 2022 systematic review addressing a similar topic (adherence measures), which screened over 60,000 studies, was not able to capture some of the relevant studies already identified during the course of this review.[111] Table 2.2 shows the tentative search strategy used, which was performed on Embase and PubMed from inception until the search date (27th February 2023).

1 Table 2.2 - Tentative systematic search strategy

General search criteria	Search string
Routinely collected data	("national register" or "national registry" or "national registries" or "public register" or "public registry" or "public registries" or "national patient register" or "national patient registry" or "national patient registries" or "administrative claims" or "routinely collected data" or "routinely collected electronic patient record#" or "pharmacy claims" or "pharmacy data" or "dispensing data" or "computerized data" or "computerized database" or "administrative claims" or "prescription claims" or "prescription database" or "prescription data" or "health database" or "health care database" or "health care claims" or "insurance plan" or "administrative database" or "claims database" or "pharmacy benefits manager" or "administrative databases" or "claims databases" or "pharmacy benefits manager" or "pharmacy refill database" or "pharmacy database" or "pharmacy databases" or "Databases, factual" or "automated database" or "automated databases" or "pharmacy claims database" or "administrative data" or "administrative record#" or "claims data" or "claims record#" or "electronic health care record#" or "electronic healthcare record#" or "electronic health data" or "electronic health record#" or "electronic medical record#" or "electronic patient record#" or "general practice data" or "general practice record#" or "hospital data" or "hospital record#" or "hospitali#ation record#" or "inpatient data" or "national data#" or "outpatient data" or "primary care record#" or "primary care data" or "routine data" or "secondary care data" or "secondary care record#" or "registries" or "registry").mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kf, fx, dq, bt, nm, ox, px, rx, ui, sy, ux, mx] OR Routinely

	Collected Health Data/ OR Database/ or Databases, Pharmaceutical/ AND
Data quality	(quality or accuracy or agreement or completeness or concordance or kappa or negative predictive value or positive predictive value or npv or ppv or sensitivity or specificity or valid or validation or validity or treatment effect# or intervention effect# or matched or miscoding or validation studies as topic/ or validation study/ or data accuracy/) AND
Medications data	(medication or drug or pharmaceutical preparation or adherence or compliance or persistence or discontinuation or switching or "drug utilization review" or "active drug list" or "drug list" or "current medication list" or "active medication list" or "drug count" or "drug utilization reviews" or "active drug lists" or "drug lists" or (("current medication" or "active medication") and (list or list)) or "drug counts" or "drug utilization review" or "active drug list" or "drug list" or "current medication list" or "active medication list" or "drug count" or "drug utilization reviews" or "active drug lists" or "drug lists" or (("current medication" or "active medication") and (list or list)) or "drug counts" or Pharmaceutical Preparations/ or Patient Medication Knowledge/ or Medication Reconciliation/ or Medication Adherence/ or Medication Errors/ or Medication Review/ or Potentially Inappropriate Medication List/ or Drug utilization review/ or Drug utilization/ or Drug Prescriptions/ or Drug Utilisation Services/ or Drug Therapy/ or Prescriptions/ or Pharmacoepidemiology/ or Prescription Drugs/ or Pharmacovigilance/ or Electronic Prescribing/ or Drugs, Essential/ or Medication Systems/)

1

In this situation, and similarly to previous reports, simple chain searches (also termed footnote chasing) - wherein references from identified articles are collected, as well as articles that referenced those studies – were found to be the most informative and useful approach for identifying relevant publications.[196,203] In addition to this, the "Similar articles" feature in PubMed was used to retrieve other potentially relevant studies.

Based on the 97 relevant studies identified thus far, an initial list of 9,890 titles was therefore extracted using automated retrieval from the "Cited By" and "Similar Articles" features in PubMed. This list was deduplicated, and a sequential review of titles, abstracts, and full-texts was performed based on that list, with new relevant studies identified finally added to the original reference list. In an iterative approach, the same technique was then applied to new articles identified. Subsequent rounds were then initiated until no more relevant studies could be found. In total, 10 review rounds were performed, with the first on the 10th February 2023 and the last on the 10th March 2023. Finally, additional studies were added from manual review of reference lists from relevant articles identified in the footnote chasing round. All searches performed using this recursive approach were restricted to PubMed only.

2.3.3 Data collection

Details of relevant studies were collected using a single tool for all studies. Characteristics of interest were: country, study size, research setting (RCT or observational), population, medication group, RCD source, comparator source, reference source chosen, agreement statistic, drug exposure metric, and drug exposure derivation approach. Agreement results for each direct comparison were also collected, as well as the overall interpretation of the agreement as described by authors in each study.

2.3.4 Statistical analysis, data, and code

Study characteristics are presented using median and IQR or absolute and relative frequencies, as appropriate, and grouped by study type (RCT vs observational). Aggregate agreement results are presented using medians and IQRs and plotted using box plots, with box edges corresponding to quartiles 1 (Q1) and 3 (Q3) and whiskers spanning to 1.5 times the IQR below Q1 and above Q3. The association between study size and quantitative agreement results was assessed with Pearson's correlation. No formal hypothesis testing was performed to compare agreement results across the different study characteristics assessed. Data management, analysis, and plotting were performed using R v4.3.1. Data tables with individual study details and results, as well as all statistical code used, are available at https://github.com/gpessoaamorim/dphil-literature-review.

2.4 Results

2.4.1 Included studies – geographical settings, populations, and medications

In total, approximately 37,500 potentially relevant studies were screened (based on titles) over the multiple review rounds performed. This included approximately 21,000 from the free searches, and 16,423 from the subsequent footnote chain searches – Figure 2.1.

The final number of relevant studies identified was 192. This included 91 studies identified within the first set of free literature searches, 78 from the recursive footnote chasing rounds, and 23 from manual review of reference lists. Identified references included 18 manuscripts conducted within RCTs and 174 within observational studies, resulting in 173 distinct head-to-head comparisons within RCTs and 1277 within observational studies (as the same study could include comparisons on multiple drugs). Supplementary tables S2.1-S2.2 and the supplementary references list (included in the Appendix to this thesis) provide more details on the individual studies identified.

Figure 2.1 - Literature review flowchart

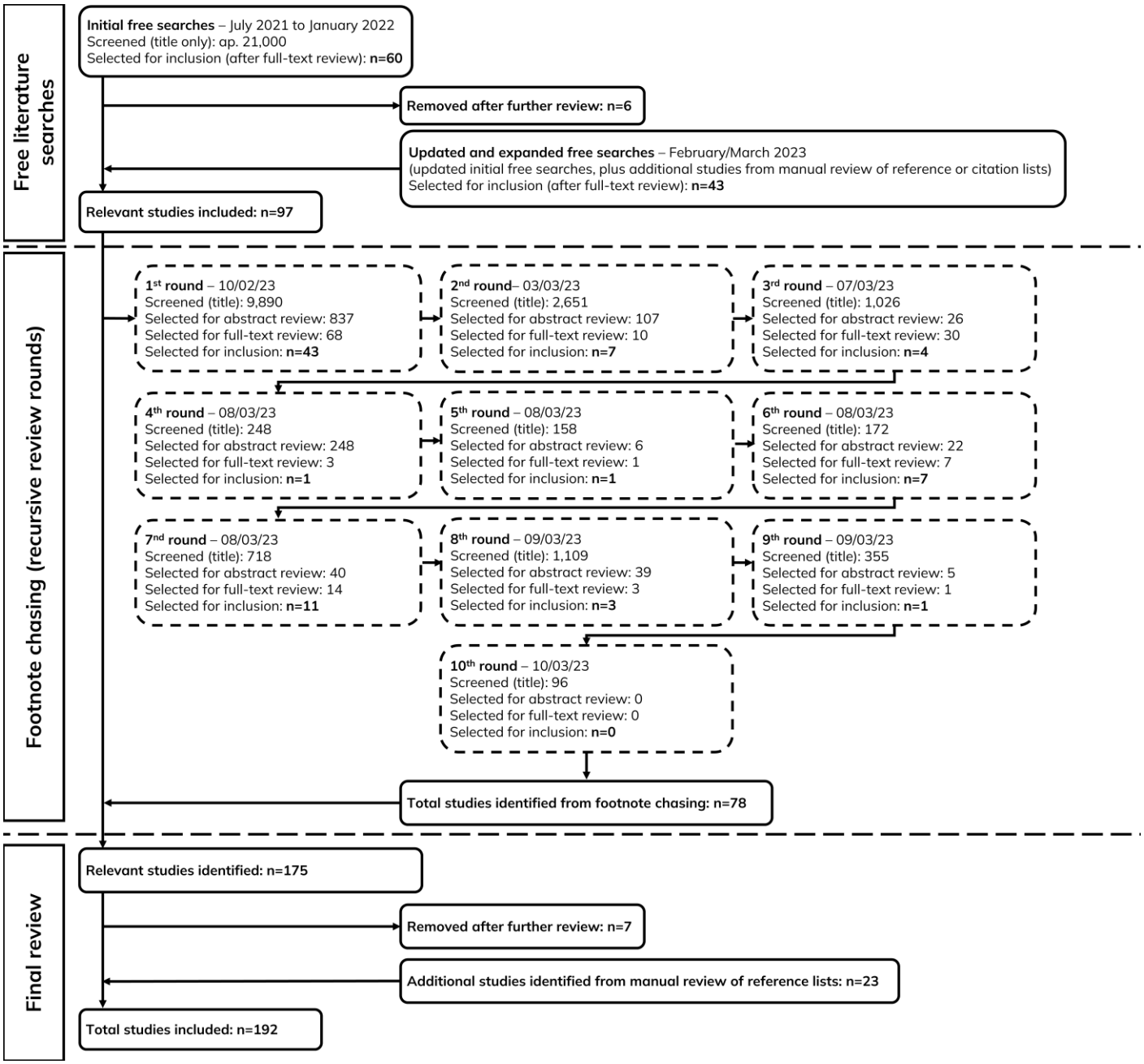


Figure legend: Footnote chasing refers to review of reference lists and the Referenced by and Similar Articles features in Pubmed

The median number of individuals included in the studies identified was 544 (164, 602) for RCTs and 620 (207, 2,472) for observational studies. Most studies were conducted in the United States (67% of RCTs and 37% and observational studies), Canada (5.6% and 11%), Scandinavia (11% and 21%), Australia (11% and 4.6%), and the Netherlands (0% and 7.5%), but comparisons within observational studies included a wide range of geographical locations (Table 2.3). The most common study populations considered were population cohorts (28% and 33%), followed by cardiovascular disease (39% and 14%) and oncology (5.6% and 13%) - Table 2.4. For medication groups, the most common were cardiovascular (22% and 17%), endocrinology (10% for both study types) and mental health medications (14% and 8.5%). Both RCTs and observational studies included comparisons focused on a diverse range of populations and medications.

1 Table 2.3 - Characteristics of included studies: size and geographical location

Characteristic	RCT, N=18	Observational study, N=174
Study size (individuals) – median (IQR)	544 (164, 602)	620 (207, 2,472)
Country		
US	12 (67%)	65 (37%)
Canada	1 (5.6%)	19 (11%)
Denmark	0 (0%)	18 (10%)
Netherlands	0 (0%)	13 (7.5%)
Australia	2 (11%)	8 (4.6%)
Finland	2 (11%)	6 (3.4%)
Sweden	0 (0%)	8 (4.6%)
Norway	0 (0%)	6 (3.4%)
Spain	0 (0%)	6 (3.4%)
France	0 (0%)	5 (2.9%)
Ireland	0 (0%)	4 (2.3%)
Japan	0 (0%)	3 (1.7%)
UK	0 (0%)	3 (1.7%)
Belgium	1 (5.6%)	1 (0.6%)
Israel	0 (0%)	2 (1.1%)
Austria	0 (0%)	1 (0.6%)
Germany	0 (0%)	1 (0.6%)
Italy	0 (0%)	1 (0.6%)
Mongolia	0 (0%)	1 (0.6%)
New Zealand	0 (0%)	1 (0.6%)
Poland	0 (0%)	1 (0.6%)
Taiwan	0 (0%)	1 (0.6%)

Rows are presented by descending order of overall count (regardless of study type). Proportions were calculated separately for each study type. Each study may be counted more than once if including comparisons in more than one location.

2

3

1 Table 2.4 - Characteristics of included studies: population and medication group

Characteristic	RCT, N=18	Observational study, N=174
Population		
Population cohort	5 (28%)	57 (33%)
Cardiovascular disease	7 (39%)	24 (14%)
Oncology	1 (5.6%)	23 (13%)
Hospitalised patients	1 (5.6%)	13 (7.5%)
Paediatrics	0 (0%)	14 (8.0%)
Respiratory disease	2 (11%)	11 (6.3%)
Gynaecology	0 (0%)	9 (5.2%)
Mental health	2 (11%)	3 (1.7%)
Chronic diseases	0 (0%)	3 (1.7%)
Endocrinology	0 (0%)	3 (1.7%)
Frailty	0 (0%)	3 (1.7%)
Gastrointestinal disease	0 (0%)	2 (1.1%)
Infectious diseases	0 (0%)	2 (1.1%)
Neurology	0 (0%)	2 (1.1%)
Rheumatology	0 (0%)	2 (1.1%)
Dermatology	0 (0%)	1 (0.6%)
Nephrology	0 (0%)	1 (0.6%)
Ophthalmology	0 (0%)	1 (0.6%)
Medication group		
Cardiovascular	11 (22%)	65 (17%)
Endocrinology	5 (10%)	39 (10%)
Mental health	7 (14%)	33 (8.5%)
All medications	2 (4.0%)	36 (9.3%)
Respiratory	6 (12%)	29 (7.5%)
Vaccinations	0 (0%)	27 (7.0%)
Analgesia	3 (6.0%)	23 (5.9%)
Gastrointestinal	5 (10%)	19 (4.9%)
Oncology	0 (0%)	22 (5.7%)
Infectious diseases	2 (4.0%)	17 (4.4%)
Neurology	2 (4.0%)	12 (3.1%)
Ear, eye, nose, and throat	2 (4.0%)	11 (2.8%)
Metabolic	3 (6.0%)	10 (2.6%)
Gynaecology	0 (0%)	11 (2.8%)
Rheumatology	0 (0%)	11 (2.8%)
Urology	1 (2.0%)	6 (1.5%)
Dermatology	0 (0%)	6 (1.5%)
Haematology	1 (2.0%)	4 (1.0%)
Musculo-skeletal	0 (0%)	5 (1.3%)
Alimentary tract/central nervous system	0 (0%)	1 (0.3%)
Ophthalmology	0 (0%)	1 (0.3%)

Rows are presented by descending order of overall count (regardless of study type). Proportions were calculated separately for each study type. Each study may be counted more than once if including comparisons of more than one population or medication group.

2

3

2.4.2 Included studies – RCD sources, comparator sources, and reference used

The most common RCD source used were claims/pharmacy data across both study types (100% of RCTs and 67% of observational studies), followed by prescription and vaccination registries (13% and 12%, within observational studies only) -Table 2.5. Less commonly used RCD sources included specific disease registries, inpatient registries, and primary care data.

The most common medication data source used for comparison with RCD was patient self-report across both study types (76% and 56%), followed by medical records (23% within observational studies). Less commonly used sources included parental reports (for studies in paediatric populations), medication reconciliation lists upon hospital admission, blood drug levels, pharmacy records, electronic monitoring (e.g. pill bottle opening or inhaler use), and pill counts. Of note, most sources of drug exposure considered were, like RCD, indirect measures (with the exception of blood drug levels). RCD sources were the most common reference used for comparisons both within RCTs (39%) and observational studies (37%), followed by comparisons with no specific reference (i.e. where both RCD and the comparator source were considered equally appropriate; 33% and 16%), medical records (18% within observational studies), and self-report (17% and 14%).

Table 2.5 - Characteristics of included studies: RCD source, comparator source, and reference used

Characteristic	RCT, N=18	Observational study, N=174
RCD source		
Claims/pharmacy data	18 (100%)	117 (67%)
Prescription registry	0 (0%)	22 (13%)
Vaccination registry	0 (0%)	21 (12%)
Disease registry	0 (0%)	9 (5.1%)
Inpatient registry	0 (0%)	4 (2.3%)
Primary care data	0 (0%)	2 (1.1%)
Comparator		
Self-report	16 (76%)	109 (56%)
Medical records	0 (0%)	45 (23%)
Parental report	0 (0%)	11 (5.7%)
Medication reconciliation lists	0 (0%)	7 (3.6%)
Drug levels	1 (4.8%)	5 (2.6%)
Pharmacy records	1 (4.8%)	5 (2.6%)
Electronic monitoring	2 (9.5%)	3 (1.5%)
Clinician report	0 (0%)	4 (2.1%)
Home inventory	0 (0%)	3 (1.5%)
Pill counts	1 (4.8%)	2 (1.0%)
Reference standard		
RCD	7 (39%)	66 (37%)
Not specified (sources considered equivalent)	6 (33%)	29 (16%)
Medical records	0 (0%)	32 (18%)
Self-report	3 (17%)	24 (14%)
Medication reconciliation lists	0 (0%)	7 (4.0%)
Drug levels	1 (5.6%)	5 (2.8%)
Electronic monitoring	1 (5.6%)	3 (1.7%)
Parental reports	0 (0%)	4 (2.3%)
Pharmacy records	0 (0%)	3 (1.7%)
Home inventory	0 (0%)	2 (1.1%)
Clinician report	0 (0%)	1 (0.6%)
Pill counts	0 (0%)	1 (0.6%)

Rows are presented by descending order of overall count (regardless of study type). Proportions were calculated separately for each study type. Each study may be counted more than once if including comparisons of more than one RCD source or comparator.

2.4.3 Included studies – drug exposure methodology

The most common drug exposure metric assessed was current medication use at a specific time point (43% within RCTs, 42% within observational studies), followed by use within a defined period in the past (14% and 20%), adherence to a medication over a period of time (38% and 12%), and ever use in the past (0% and 12%) -Table 2.6. Less commonly used drug exposure metrics included dose, date of exposure, date of initiation, number of doses, and treatment duration.

The most common analytical method used to extract drug exposure from RCD sources was the fixed window method (i.e. requiring one or more records within a specific past period to define exposure at a date of interest, as depicted in Figure 2.2; 33% within RCTs and 47% within observational studies). This was followed by the legend-time method (i.e. defining exposure at a date of interest if the patient was expected to have medication in their possession at that date, based on dispensing dates, dispensed quantities, and daily dose, as depicted in Figure 2.3; 62% and 23% respectively). Other methods used included direct extraction (24% within observational studies) and a combination of fixed window and legend-time (0.5% of observational studies). Approximately 5% of studies across RCTs and observational studies did not specify the method used.

Figure 2.2 - Schematic illustration of the fixed window method for deriving drug exposure using RCD

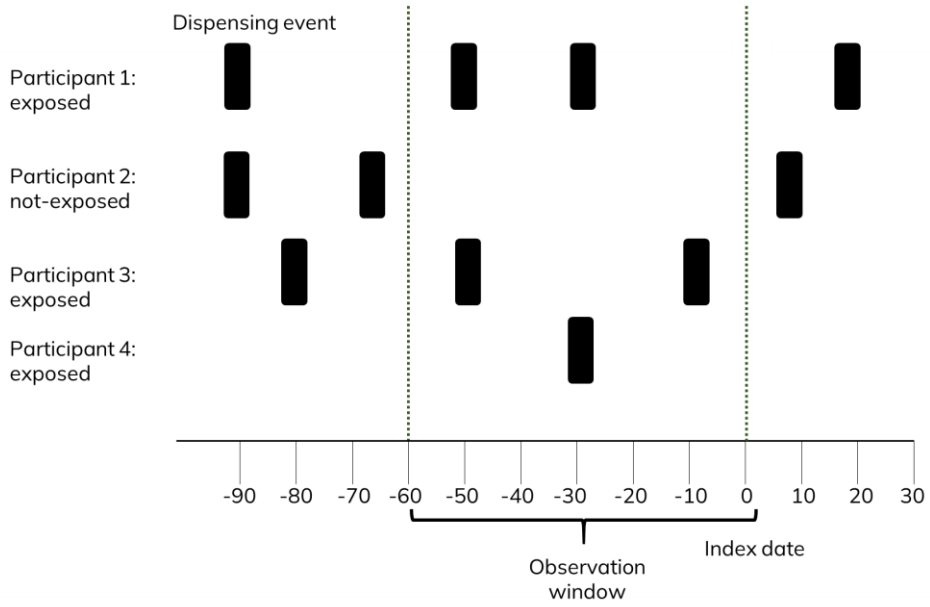


Figure legend: Each bar represents a dispensing event. In the fixed window method, any medication use events occurring within the observation window (in this example, dispensing events within the 60 day period preceding the index date) are used to define current exposure at the index date. Adapted from Anderson et al, Med Care, 2019

Figure 2.3 - Schematic illustration of the legend-time method for deriving drug exposure using RCD

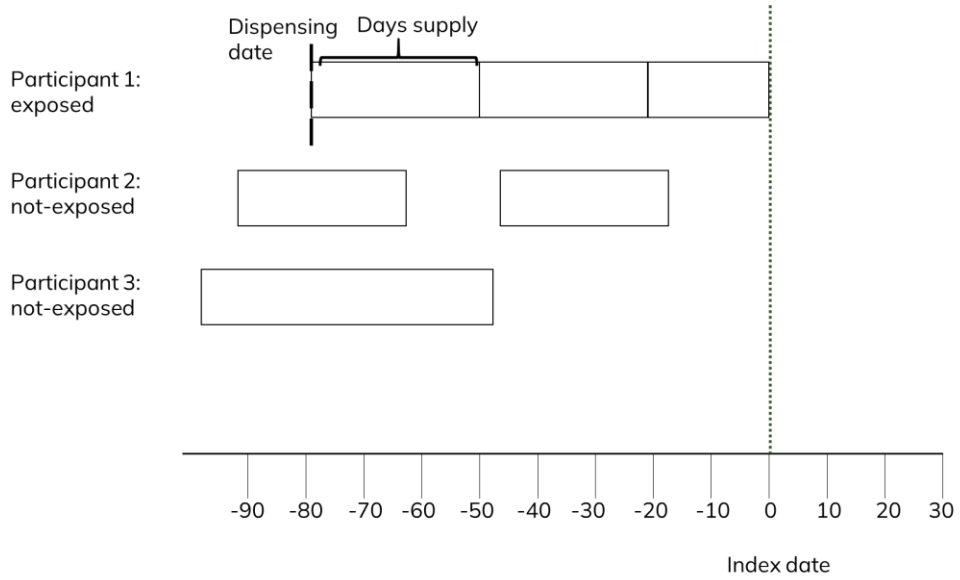


Figure legend: Each rectangle represents the number of medical days' supply issued at a particular dispensing date. An individual is considered to be exposed to medication at the index date if they are expected to be in possession of medication at that time, by taking into account the medication daily dose and days supply issued in preceding dispensing events. Adapted from Anderson et al, Med Care, 2019

Table 2.6 - Characteristics of included studies: drug exposure and analytical method

Characteristic	RCT, N=18	Observational study, N=174
Drug exposure metric		
Current use	9 (43%)	88 (42%)
Past use	3 (14%)	42 (20%)
Adherence	8 (38%)	26 (12%)
Ever use	0 (0%)	26 (12%)
Dose	1 (4.8%)	9 (4.3%)
Date of use	0 (0%)	4 (1.9%)
Initiation	0 (0%)	4 (1.9%)
Number of doses	0 (0%)	2 (1.0%)
Treatment duration	0 (0%)	2 (1.0%)
Medication name	0 (0%)	1 (0.5%)
Quantity	0 (0%)	1 (0.5%)
Supply duration	0 (0%)	1 (0.5%)
Switch	0 (0%)	1 (0.5%)
Time of use	0 (0%)	1 (0.5%)
Time since first use	0 (0%)	1 (0.5%)
Time since last use	0 (0%)	1 (0.5%)
Drug exposure derivation method		
Fixed window	7 (33%)	99 (47%)
Legend-time	13 (62%)	49 (23%)
Extracted directly	0 (0%)	50 (24%)
Not specified	1 (4.8%)	11 (5.2%)
Fixed window plus legend-time	0 (0%)	1 (0.5%)

Table legend: Rows are presented by descending order of overall count (regardless of study type). Proportions were calculated separately for each study type. Each study may be counted more than once if including comparisons of more than one drug exposure or derivation method.

2.4.4 Agreement metrics and results

Table 2.7 shows the agreement metrics used and the overall interpretation of agreement results in each study (as specified by the authors of each study). Multiple agreement metrics were used, with the most common ones being 1) Cohen's κ (which provides a coefficient describing the proportion of agreement observed beyond that expected by chance; 20% of RCTs and 26% of observational studies); 2) sensitivity and specificity (22% and 24%), 3) PPV/NPV; (18% and 20%), and 4) percent agreement (the proportion of cases where both sources agree; 22% and 16%). Less commonly used metrics included direct comparisons of counts or proportions identified (7.5% and 8.6%), correlation of numeric estimates such as adherence, dose, or number of doses (7.5% and 4.4%), and prevalence-and-bias-adjusted-kappa (PABAK, an alternative measure to Cohen's κ; 2.5% and 1.9%).

Supplemental Table S2.1 (in the Appendix to this thesis) shows the underlying formulae involved in calculating agreement metrics for binary exposures (yes/no) from two-by-two tables, and used to calculate sensitivity, specificity, PPV, NPV, percent agreement, κ, and PABAK.

The overall interpretation of agreement results was mostly positive (56% and 71% in RCTs and observational studies respectively), with negative/poor agreement results in 28% and 14%, and mixed results in 17% and 14%.

Table 2.7 - Characteristics of included studies: agreement metrics and interpretation

Characteristic	RCT, N=18	Observational study, N=174
Agreement metric		
κ	8 (20%)	81 (26%)
Sensitivity and specificity	9 (22%)	76 (24%)
Positive/negative predictive values	7 (18%)	62 (20%)
Percent agreement	9 (22%)	49 (16%)
Direct comparison (counts/proportions)	3 (7.5%)	27 (8.6%)
Correlation	3 (7.5%)	14 (4.4%)
PABAK	1 (2.5%)	6 (1.9%)
Conclusion		
Positive	10 (56%)	124 (71%)
Negative	5 (28%)	25 (14%)
Mixed	3 (17%)	25 (14%)

Table legend: Rows are presented by descending order of overall count (regardless of study type). Proportions were calculated separately for each study type. Each study may be counted more than once if including comparisons of more than one agreement metric.

Figure 2.4 shows aggregate agreement results. The median (IQR) percent agreement across all studies was 0.87 (0.7-0.95), κ 0.68 (0.46-0.82), PABAK 0.94 (0.84-0.99), sensitivity 0.79 (0.63-0.9), specificity 0.98 (0.92-0.99), PPV 0.82 (0.68-0.92), NPV 0.96 (0.88-0.99), and correlation 0.3 (0.17-0.72). Of note, comparisons performed within RCTs exhibited generally higher or similar agreement metrics when compared with observational studies. Studies reporting overall positive results had generally higher average agreement metrics resulting from individual comparisons between sources than

1 those reporting mixed and negative results (Figure 2.5). However, there was considerable
2 overlap in agreement results between studies included in the three interpretation
3 categories, as well as large dispersion of results from individual comparisons in studies
4 included in each category.

Figure 2.4 - Aggregated statistical metrics of agreement for comparisons on drug exposure assessments between RCD and comparator sources by study type in the literature review

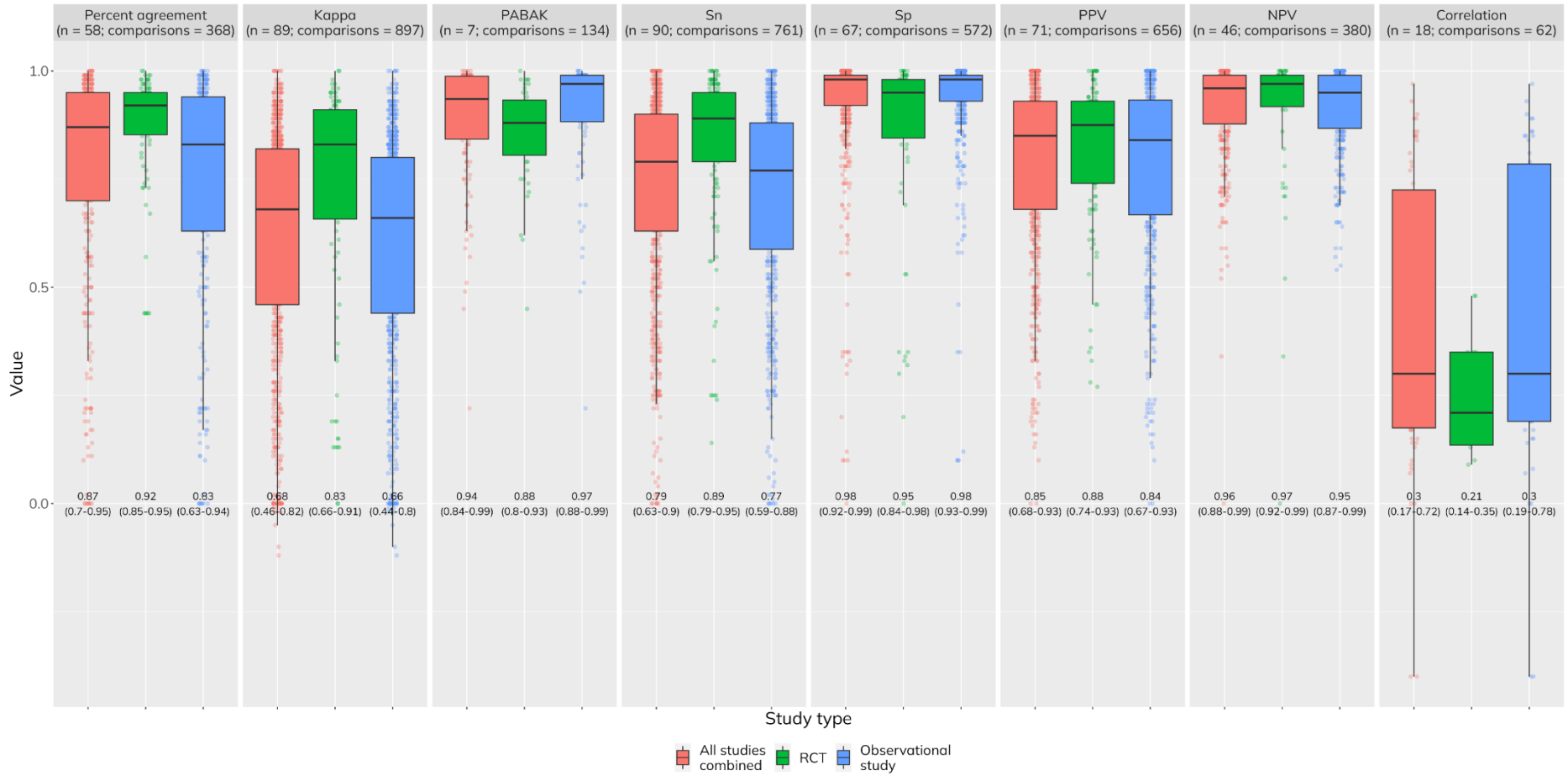


Figure legend: Each panel represents results for a different agreement metric, with the vertical axis depicting the value of that particular metric (with limits 0:1 for most metrics, and -1:1 for Kappa, PABAK, and correlation). Within each panel, comparisons are split by study type using different colour labels. Box plots are shown overlaid on dots representing individual results from each distinct comparison. Box plots span from Q1-Q3, with the middle black line showing median, and whiskers spanning to 1.5 times the IQR below Q1 or above Q3. The number of studies and individual comparisons considered in each panel are shown below the top label. Numerical values at the bottom depict median (IQR) for each agreement metric. A full list of abbreviations is included in page xxi.

Figure 2.5 – Study-specific interpretation of agreement metrics for comparisons on drug exposure assessments between RCD and comparator sources

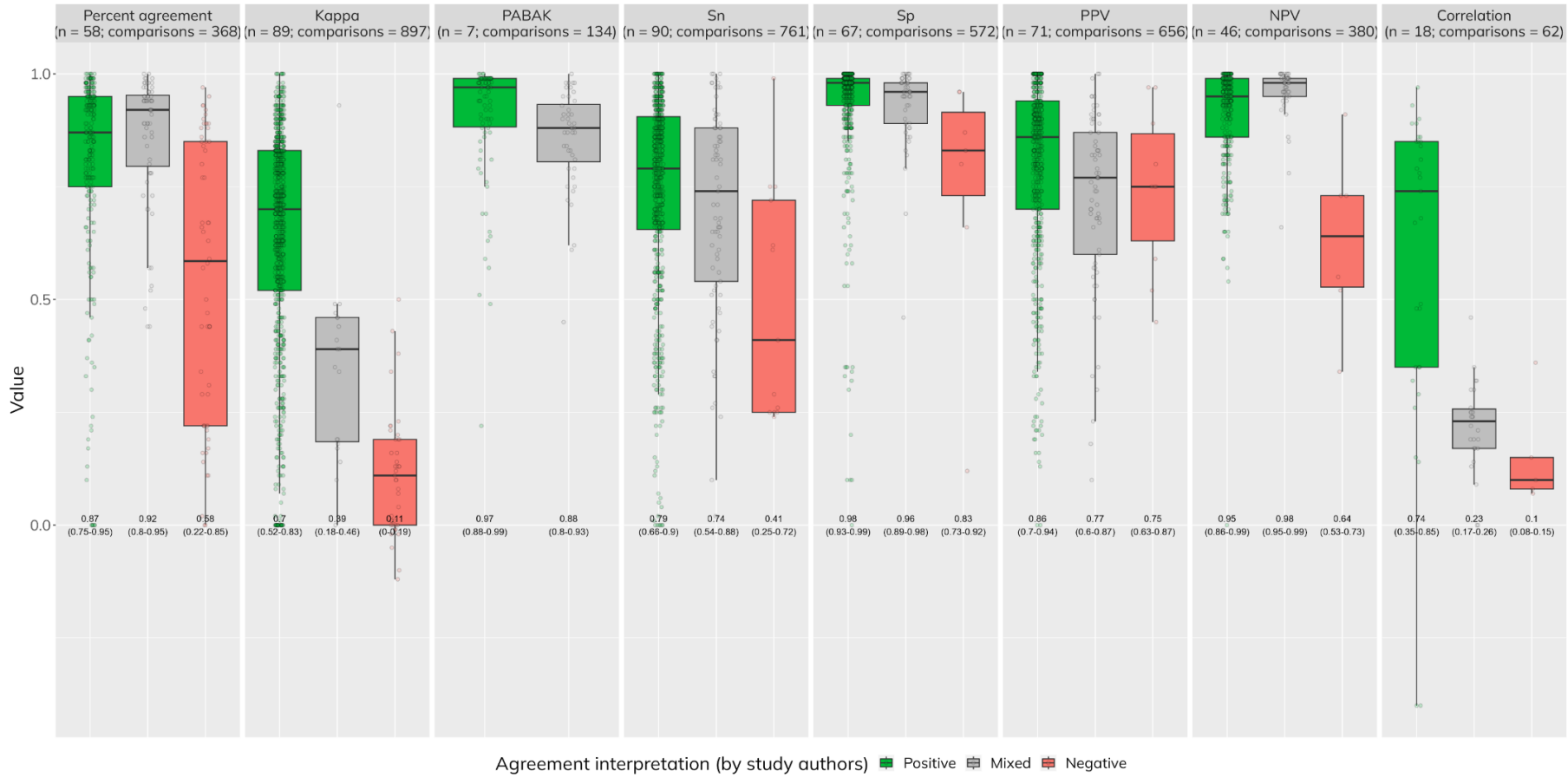


Figure legend: Each panel represents results for a different agreement metric, with the vertical axis depicting the value of that particular metric (with limits 0:1 for most metrics, and -1:1 for Kappa, PABAK, and correlation). Within each panel, comparisons are split by the overall interpretation of the agreement results seen in each study, as specified by their authors. Interpretations of agreement were extracted at the study level, rather than for individual comparisons. Box plots are shown overlaid on dots representing individual results from each distinct comparison. Box plots span from Q1-Q3, with the middle black line showing median, and whiskers spanning to 1.5 times the IQR below Q1 or above Q3. The number of studies and individual comparisons considered in each panel are shown below the top label. Numerical values at the bottom depict median (IQR) for each agreement metric. A full list of abbreviations is included in page xxi.

Agreement results were generally not associated with study size, with the exception of a positive correlation for percent agreement and a weak negative correlation for PPV (Figure 2.6).

Figure 2.6 – Linear correlation between statistical agreement metrics and study size (i.e. number of participants) for comparisons on drug exposure assessments between RCD and comparator sources

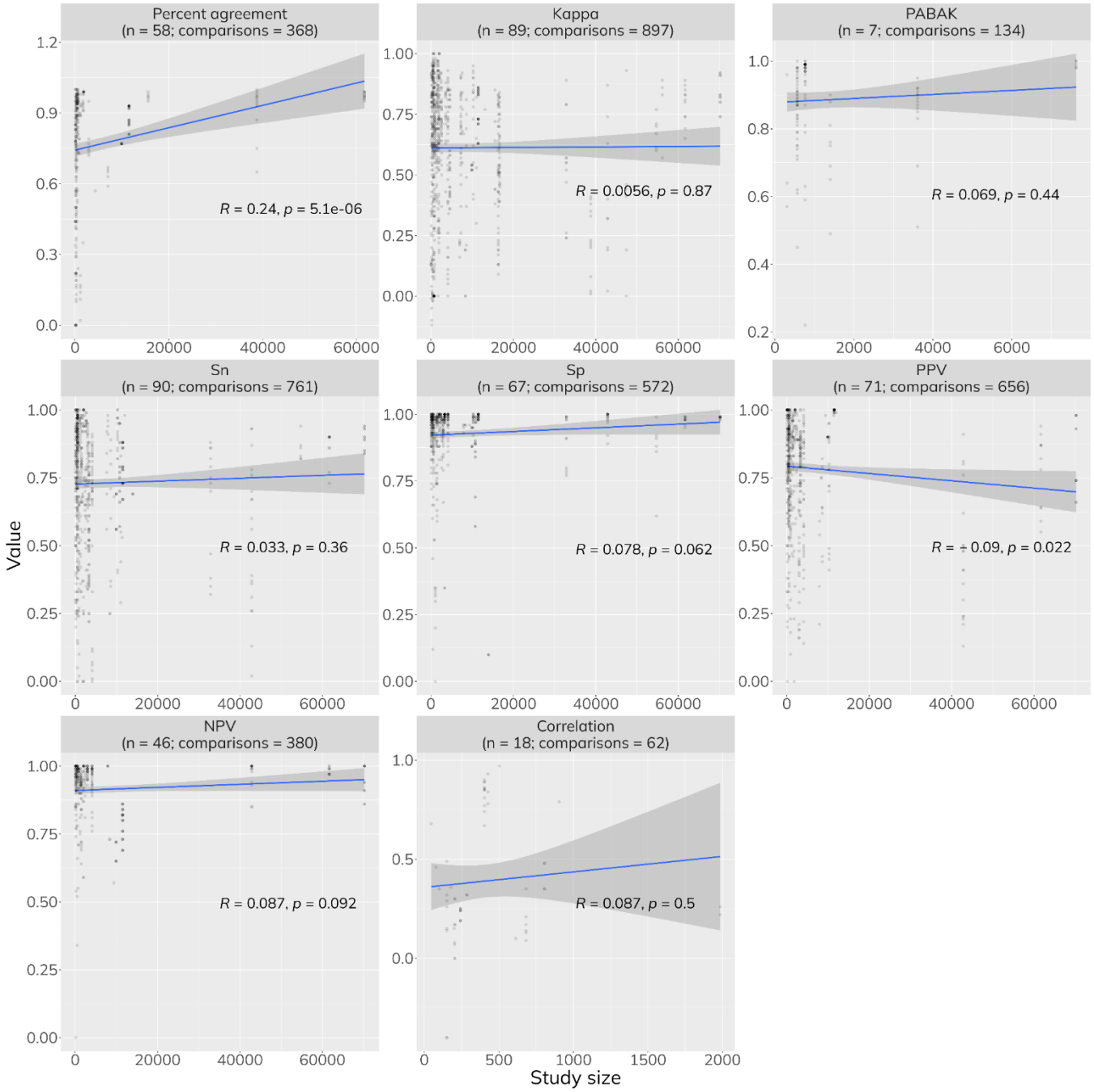


Figure legend: Each panel represents results for a different agreement metric, with the vertical axis depicting the value of that particular metric (with limits 0:1 for most metrics, and -1:1 for Kappa, PABAK, and correlation). Each dot shows one individual comparison. The number of studies and individual comparisons considered in each panel are shown below the top label. R is the Pearson correlation coefficient with p-values assessing for statistically significant evidence of linear correlation. Linear regression lines are shown alongside 95% CI margins. A full list of abbreviations is included in page xxi.

Agreement results varied greatly with the specific geographical location, population and medication group considered (Figures 2.7-2.9), particularly percent agreement, κ, sensitivity, PPV, and correlation coefficients. However, PABAK, specificity, and NPV were generally high (i.e. >0.85) across all settings considered, with the exceptions of PABAK in oncology (0.67 [0.60-0.80]); specificity in Israel (0.75), in hospitalised patients (0.32 [0.27-0.34]), and in ophthalmology (0.1); and NPV in New Zealand (0.77 [0.73-0.80]), in gastrointestinal disease (0.75 [0.64-0.80]) and in mental health (0.82), and for vaccinations (0.82 [0.72-0.86]).

Figure 2.7 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by geographical location

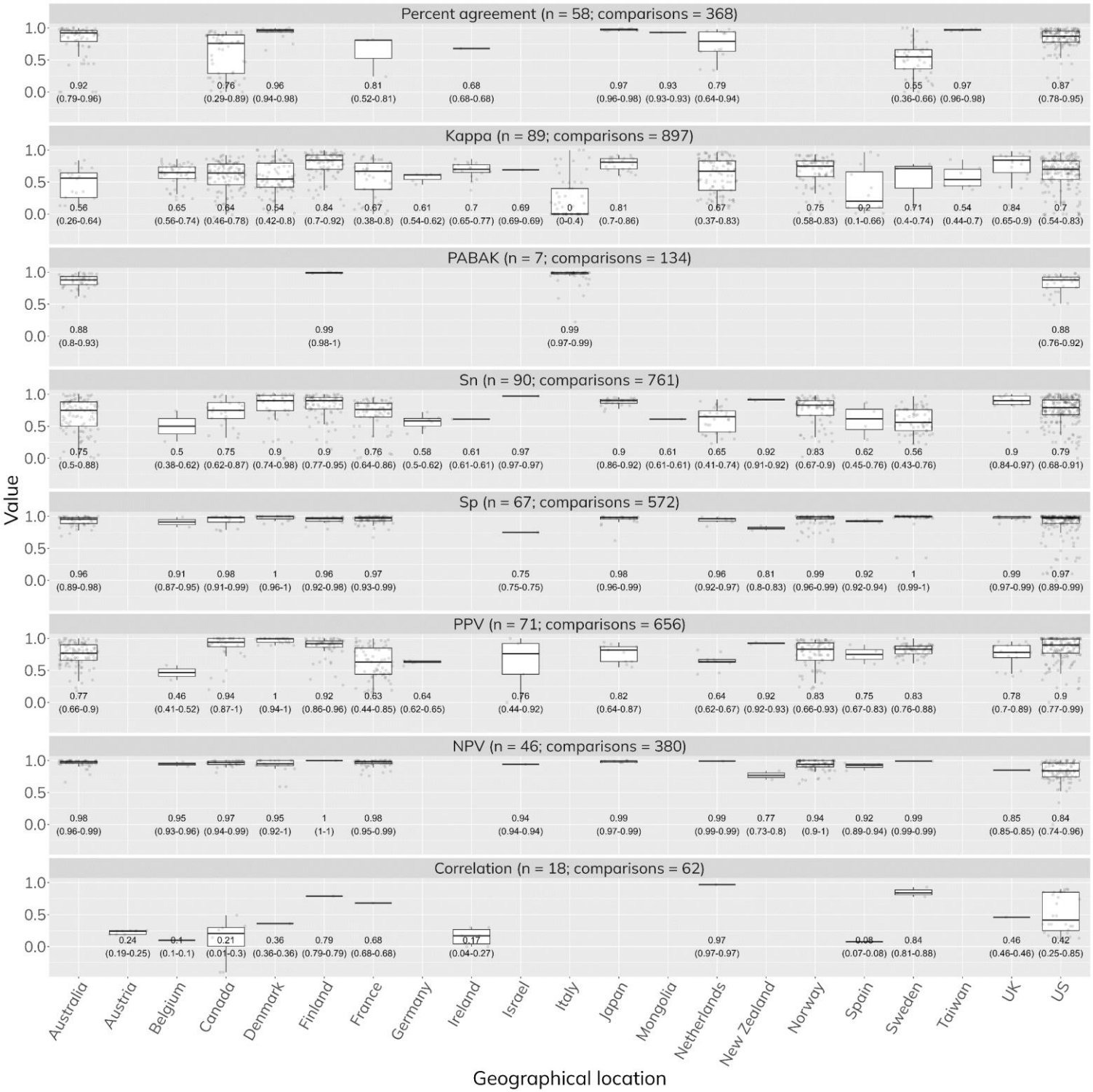


Figure legend: Each panel represents results for a different agreement metric, with the vertical axis depicting the value of that particular metric (with limits 0:1 for most metrics, and -1:1 for Kappa, PABAK, and correlation). Within each panel, the horizontal axis represents different countries. Box plots are shown overlaid on dots representing individual results from each distinct comparison. Box plots span from Q1-Q3, with the middle black line showing median, and whiskers spanning to 1.5 times the IQR below Q1 or above Q3. The number of studies and individual comparisons considered in each panel are shown below the top label. Numerical values at the bottom depict median (IQR) for each agreement metric. A full list of abbreviations is included in page xxi.

Figure 2.8 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by population group

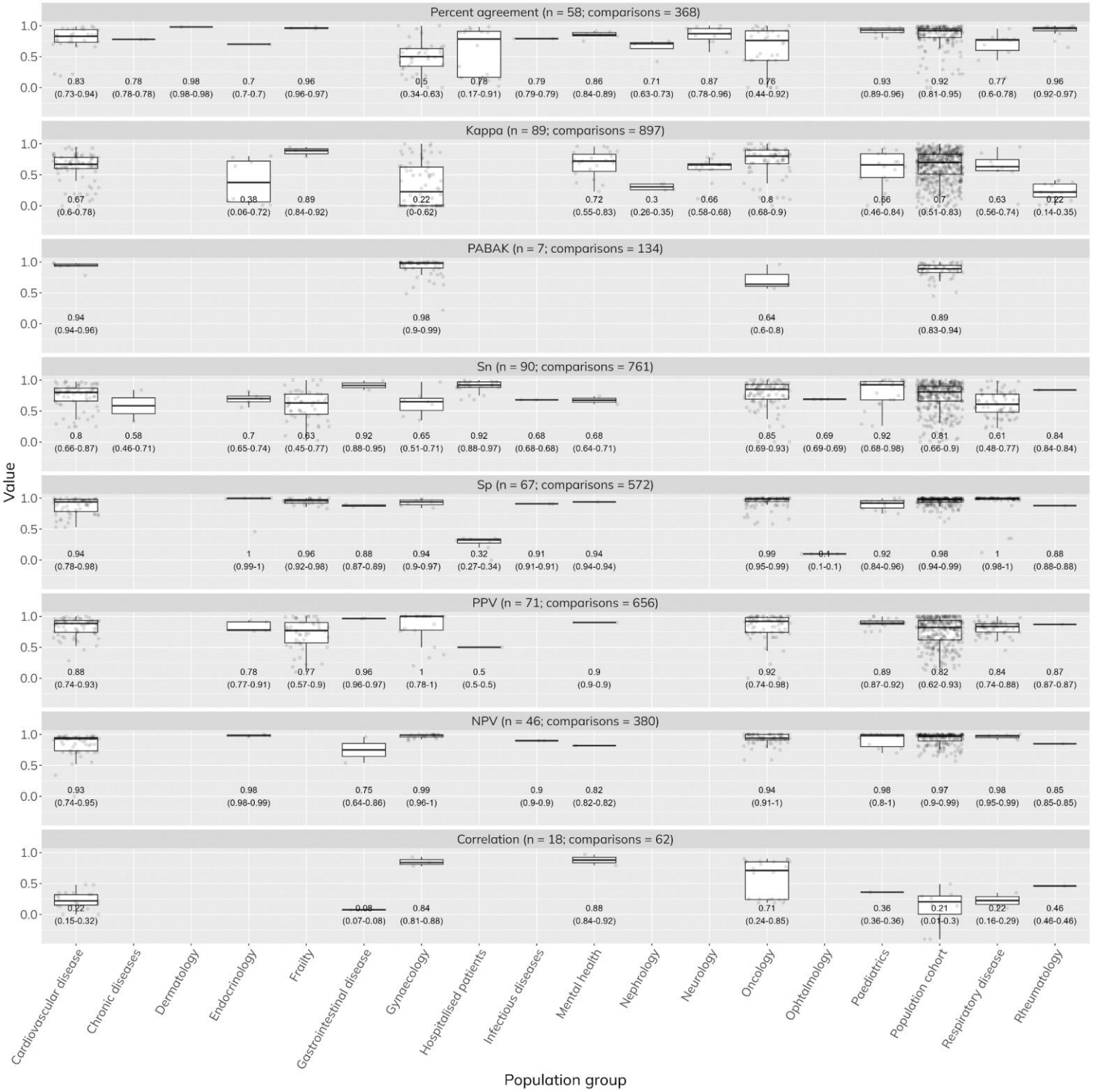


Figure legend: Each panel represents results for a different agreement metric, with the vertical axis depicting the value of that particular metric (with limits 0:1 for most metrics, and -1:1 for Kappa, PABAK, and correlation). Within each panel, the horizontal axis represents different population groups. Box plots are shown overlaid on dots representing individual results from each distinct comparison. Box plots span from Q1-Q3, with the middle black line showing median, and whiskers spanning to 1.5 times the IQR below Q1 or above Q3. The number of studies and individual comparisons considered in each panel are shown below the top label. Numerical values at the bottom depict median (IQR) for each agreement metric. A full list of abbreviations is included in page xxi.

Figure 2.9 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by medication group

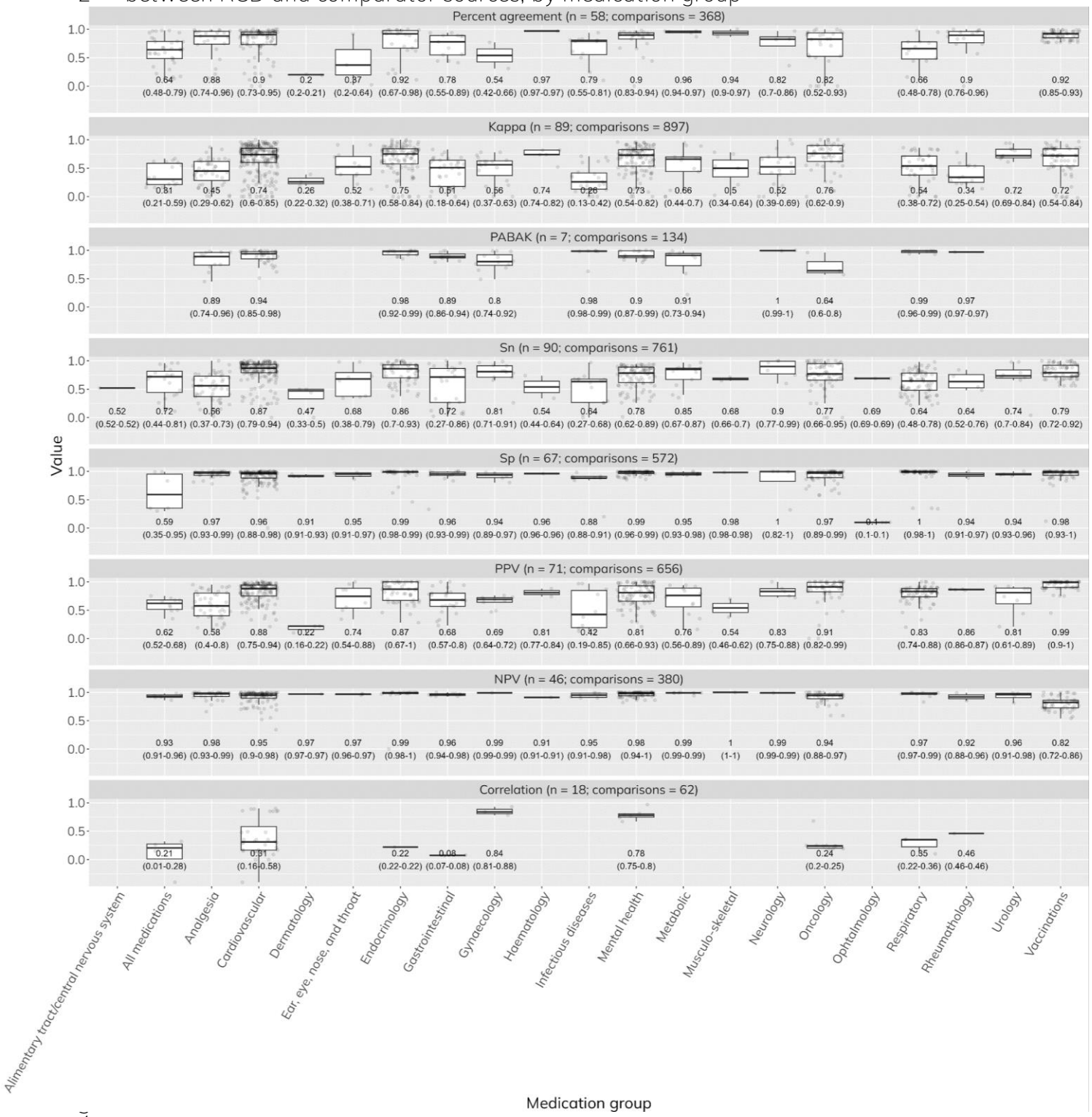


Figure legend: Each panel represents results for a different agreement metric, with the vertical axis depicting the value of that particular metric (with limits 0:1 for most metrics, and -1:1 for Kappa, PABAK, and correlation). Within each panel, the horizontal axis represents different medication groups. Box plots are shown overlaid on dots representing individual results from each distinct comparison. Box plots span from Q1-Q3, with the middle black line showing median, and whiskers spanning to 1.5 times the IQR below Q1 or above Q3. The number of studies and individual comparisons considered in each panel are shown below the top label. Numerical values at the bottom depict median (IQR) for each agreement metric. A full list of abbreviations is included in page xxi.

Agreement results exhibited generally low heterogeneity based on the RCD source used (as shown by closer median agreement metrics found across the different possible sources). Agreement was generally poorer if using prescription registries (percent agreement: 0.62 [0.41-0.87], κ: 0.54 [0.37-0.80], sensitivity: 0.68 [0.48-0.84], PPV: 0.81 [0.73-0.89]) and primary care data (percent agreement: 0.65 [0.63-0.66], κ: 0.22 [0.17-0.22], specificity: 0.88, PPV: 0.87, NPV: 0.85) when compared with other sources (Figure 2.10).

Figure 2.10 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by RCD source type

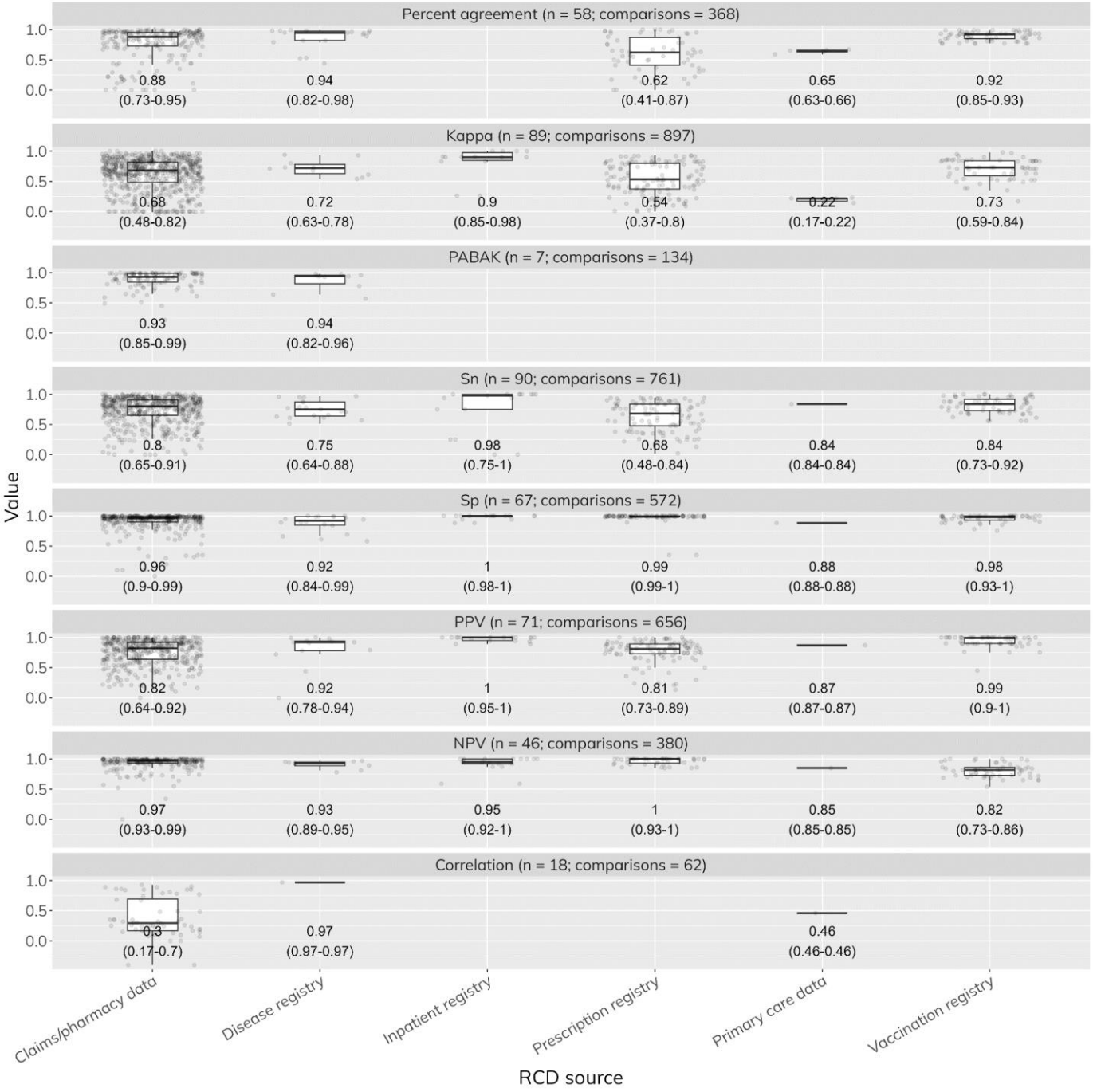


Figure legend: Each panel represents results for a different agreement metric, with the vertical axis depicting the value of that particular metric (with limits 0:1 for most metrics, and -1:1 for Kappa, PABAK, and correlation). Within each panel, the horizontal axis represents different RCD source types. Box plots are shown overlaid on dots representing individual results from each distinct comparison. Box plots span from Q1-Q3, with the middle black line showing median, and whiskers spanning to 1.5 times the IQR below Q1 or above Q3. The number of studies and individual comparisons considered in each panel are shown below the top label. Numerical values at the bottom depict median (IQR) for each agreement metric. A full list of abbreviations is included in page xxi.

When considering results based on the comparator source used, NPVs and correlation were similar across all comparators (NPVs >0.89, correlation approximately 0.20), except correlation with self-reported results (0.67 [0.24-0.85]). Other agreement metrics exhibited larger variations across the different comparators considered, with percent agreement lowest for parental report (0.16) and medication reconciliation lists (0.41 [0.16-0.89]), κ for clinician report (0.1) and pharmacy records (0.26 [0-0.9]), sensitivity for parental report (0.62 [0.48-0.78]), specificity for clinician report (0.52 [0.31-0.74]), and PPV for medication reconciliation lists (0.5).

Figure 2.11 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by comparator data source type

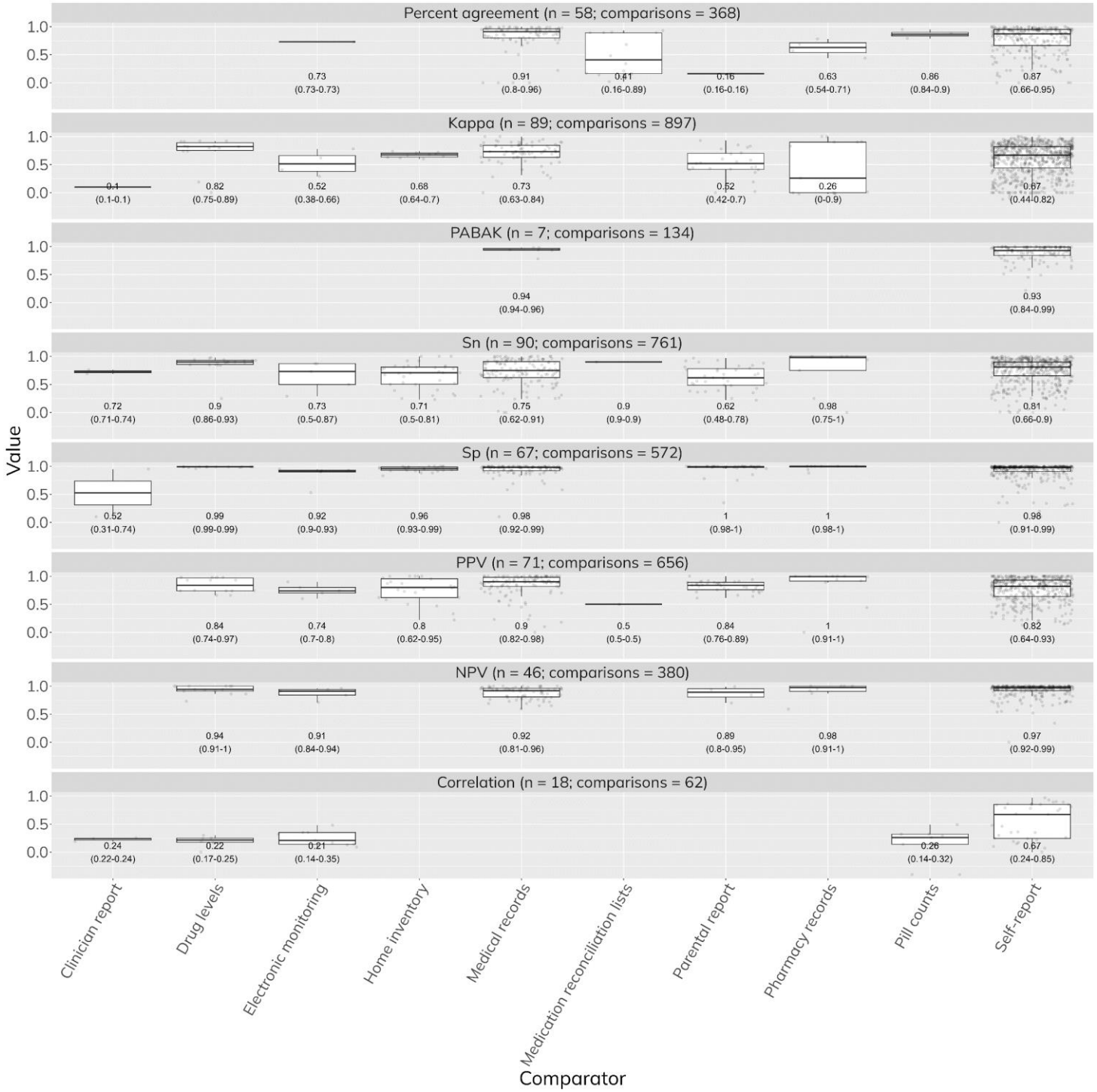


Figure legend: Each panel represents results for a different agreement metric, with the vertical axis depicting the value of that particular metric (with limits 0:1 for most metrics, and -1:1 for Kappa, PABAK, and correlation). Within each panel, the horizontal axis represents different comparator data source type (versus RCD). Box plots are shown overlaid on dots representing individual results from each distinct comparison. Box plots span from Q1-Q3, with the middle black line showing median, and whiskers spanning to 1.5 times the IQR below Q1 or above Q3. The number of studies and individual comparisons considered in each panel are shown below the top label. Numerical values at the bottom depict median (IQR) for each agreement metric. A full list of abbreviations is included in page xxi.

Agreement results were generally similar regardless of the analysis method used to derive drug exposure (legend-time vs fixed window) across the different drug exposure metrics where both methods could be used (adherence, current use, and past use) - Figure 2.12.

Figure 2.12 - Agreement results for comparisons on drug exposure assessments between RCD and comparator sources, by drug exposure outcome and analysis method

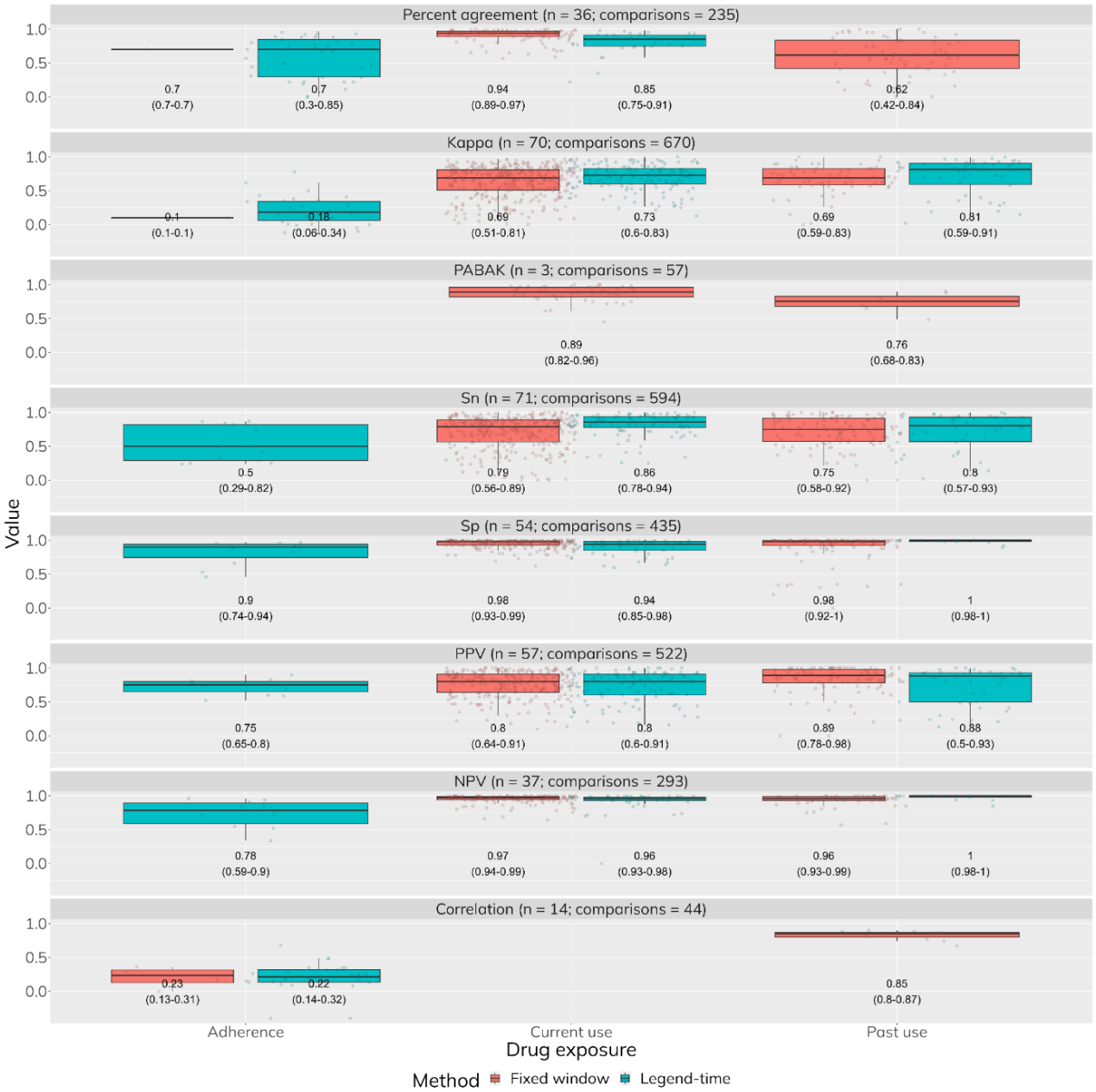


Figure legend: Each panel represents results for a different agreement metric, with the vertical axis depicting the value of that particular metric (with limits 0:1 for most metrics, and -1:1 for Kappa, PABAK, and correlation). Within each panel, the horizontal axis represents different drug exposure metric, further split by analysis method (along with different colour coding). Box plots are shown overlaid on dots representing individual results from each distinct comparison. Box plots span from Q1-Q3, with the middle black line showing median, and whiskers spanning to 1.5 times the IQR below Q1 or above Q3. The number of studies and individual comparisons considered in each panel are shown below the top label. Numerical values at the bottom depict median (IQR) for each agreement metric. A full list of abbreviations is included in page xxi.

2.5 Discussion

Repurposing of RCD for trials has been limited by concerns regarding data quality. Regulatory agencies in the UK and the US have issued specific recommendations on definitions and approaches for data quality assessment in clinical research, stressing the need for validation against existing reference sources in each specific context. This chapter presents results from a scoping review of the literature on comparisons of RCD on medications with other sources of information on medication use, within RCTs or observational studies. RCD have been compared to other sources of medications data across multiple geographical locations, populations, and drug groups. Claims/pharmacy data were the most common RCD source used, with RCD commonly being used as a reference standard themselves. Agreement between RCD and collateral sources was heterogeneous, with important variations across geographical settings, populations, and medication groups, as well as RCD sources, comparators, and statistical and data derivation approaches used. Nonetheless, in most cases the agreement observed between RCD and the comparator source was considered adequate/positive. Overall, this review has described and summarised a large body of research, identifying relevant methodological approaches and patterns to guide data analysis and interpretation in subsequent chapters of this thesis.

This review identified approximately 200 studies comparing RCD on medications with other sources. Most studies were performed in high-income countries such as the US, Canada, Australia, and several locations in Europe, but also low-and-middle-income countries (LMIC) such as Mongolia. Of note, other studies retrieved as part of this review (but not included as out of scope) also described assessments of EHR data quality in many other LMICs, including South America.[204] This shows that RCD on medications are available for research purposes across multiple locations globally, with an extensive body

of literature addressing their validation in this context. Similarly, the comparative studies included covered a wide range of patient populations and medication groups, highlighting the potential value of using RCD to extract information on exposure to multiple medications and in multiple clinical settings.

Claims and pharmacy data were the most common RCD source used. This likely reflects their widespread availability (namely in the US, across Europe, and in Asia),[178–181,205] the need for accurate data collection on medication purchase or dispensing for reimbursement purposes, and the generally automated nature of these data collections. Other sources such as prescription, vaccination, or disease registries were less commonly used. Less frequent use of these sources may be justified by a narrower scope in terms of medications included, the need to manually submit data for some of these collections, or a lack of universal coverage. Also, prescription registries are only available in a few countries with single-payer, centralised healthcare systems such as Scandinavia, the Netherlands, and Canada.[206–209] The most common comparator sources used were self-report and medical record review, possibly as these are the cheaper, more easily accessible, and most practical approaches to identify drug exposure.[137,210] More direct and objective measures such as blood drug levels, electronic intake monitoring, and pill counts were seldom used (although more commonly in RCTs than in observational studies), likely due to restraints with cost, practicality, or scalability which render them less attractive for large-scale studies. Moreover, these objective measures may still generate biased estimates (e.g. due to individual metabolism or "white-coat" adherence before study visit), and only provide a snapshot of patient adherence at a specific moment, which may not be reflective of patient behaviour over time (namely for long-term therapies in chronic diseases).[96] However, these methods may be of particular value in settings where close adherence monitoring may justify the cost and difficulty (e.g. phase I or II trials, and those in HIV or tuberculosis).[210,211]

Overall, these observations suggest that self-report (a subjective, indirect measure of drug

exposure) is widely used as the preferential method of medication data collection across trials and observational studies, and thus acceptable as reasonable reference standard against which RCD sources may be validated.

Nevertheless, RCD was more commonly used as the reference standard for validation than any other source, both within trials and observational studies, and much more frequently than either medical records or self-report. Although not widely used in trials, RCD are a practical, scalable, objective way to identify drug exposure. RCD are an established standard in observational research, where they are commonly used to validate subjective measures of drug exposure or adherence.[32,197,199] Previous studies recognised that drug exposure and adherence may be collected in a multitude of ways, without any single gold-standard. Some of the validation studies included here have also recognised this ambiguity by running alternative analyses using either RCD or a comparator source as reference.[212,213] Thus, the optimal approach for selecting data source, as highlighted by relevant FDA and MHRA guidance, is to consider the specific purpose being addressed, as well as the intricacies and characteristics of each possible method, so that the one that most closely aligns with the objective at hand can be selected.[211] For example, claims data may be more reliable than self-report to capture exposure to chronic medications for which reimbursement is made. On the other hand, self-report may be preferred to ascertain more granular medication intake patterns (e.g. how frequently doses are missed) or to capture exposure to medications not recorded in claims data. Altogether, these results indicate that RCD should also be considered a reasonable source of drug exposure information for use in trials, provided that data quality assessments focused on the specific use case are satisfactory.

Multiple drug exposure metrics were identified, with the most common ones being cross-sectional drug exposure at an index moment (current use) or during a period of time (past/ever use), or longitudinal exposure across a period of time (adherence). However,

RCD was also used to extract a variety of other drug use metrics such as dose or timing, highlighting its versatility. Also, when compared with other sources such as self-, parent, or clinician report, RCD may be considered a more reliable ascertainment source for detailed information (e.g. dose, time since first use, date of use, time of switch) in order to avoid incorrect reports, particularly for past exposure (i.e. recall bias). Conversely, self-report may be more accurate for very recent use, namely recent switches or discontinuations which may be missed if there are lags in RCD reporting or dissemination.[136,212,214–218]

Comparisons of RCD with other sources yielded mostly positive overall results (as interpreted by each study's authors) across the various agreement metrics used (e.g. median percent agreement: 0.87; κ: 0.68). However, overall sensitivity (which may be interpreted in terms of RCD completeness) was suboptimal (median 0.79), although with good PPVs (which may be interpreted as accuracy; median 0.85), and excellent specificity and NPV (median 0.97 and 0.96 respectively). These results suggest that, overall, RCD may be considered an accurate indicator of true drug exposure (with high PPV and specificity), but not a perfectly complete source (as shown by the lower sensitivity, albeit with very high NPV). Potential reasons include issues with coverage of claims data (e.g. age-based eligibility, medication dispensing at pharmacies not covered by that plan, use of non-reimbursed, over-the-counter, or as-need medications, out-of-pocket payments, or use of free drug samples), primary care data (e.g. medications prescribed or dispensed in the hospital setting), or disease registries (e.g. manual submission, focus on specific drugs).[176,211,211,211,219] This confirms previous reports that triangulation between different sources is likely to be the best approach to correctly identify true drug exposure, with distinct sources providing information on different drugs or complementary aspects of drug exposure and associated behaviours.[115,211,218,220–222] Importantly, overall agreement metrics were generally higher within RCTs than observational studies, or similar.

Assuming that higher agreement between parallel sources of medication exposure is generally indicative of improved data quality (of either source),[223] this observation further supports the use of RCD on medications within trials to a similar extent to that occurring in observational studies, particularly as data collected within trials is usually subject to stricter quality controls.

Agreement results did not vary greatly depending on the specific RCD source or comparator used, although they were generally better when RCD was compared against other objective sources (such as medical records, drug levels, or home inventory) versus comparisons with subjective measures (such as self-, clinician, or parent report). While this suggests that RCD (and other objective sources) may be considered more accurate than subjective sources, the latter may also reveal true medication use patterns not captured by objective data collection methods (e.g. non-adherence to prescribed and collected medications, or exposure to medications not captured by RCD). Likewise, agreement between RCD and other sources exhibited distinctive patterns depending on the type of comparator source used, which might point towards plausible reasons for disagreement. These included lower overall percent agreement and PPV for medication reconciliation records (possibly representing the impact of usual medication discontinuation upon hospital admission), lower sensitivity for parental reports (use of over-the-counter or non-reimbursed medications by their children), and lower specificity for clinician reports (due to recent medication stops).

Extraction of medication exposure from RCD may be performed directly (e.g. current medication lists), but generally requires the implementation of computer algorithms or similar analytical approaches. Algorithm selection and development is a particularly challenging and sensitive dimension of research using RCD, as different approaches may lead to distinct results, even if using the same source and population.[224,225] Two main analytical approaches to extract drug exposure information from RCD sources were

identified (legend-time, fixed-window), which yielded overlapping overall agreement metrics, similarly to results from a previous systematic review.[182] Of note, approximately 5% of all studies did not specify the methodology used, limiting interpretation and extrapolation of their results.[224,225]

There was great heterogeneity of agreement results across different locations, reflecting the impact of different healthcare systems and data collection purpose and pathways upon the resulting RCD sources.[123] There was also great variability across different populations, and medication groups. Of note, specificity was particularly poor for studies in hospitalised patients and those in ophthalmology, possibly as the result of stopping medications used in the community (captured by RCD) upon hospital admission, and of challenges with correctly identifying drug exposure with topical ocular formulations.[226,227]

For the most common medication groups considered (cardiovascular and oncology), all agreement results were generally high, but there was wide variability in the interpretation of similar agreement results by individual study authors, with considerable overlap between studies with overall positive, mixed, and negative results. Taken together, these results confirm that assessment of RCD quality is purpose- and context-specific, so that harnessing RCD requires appropriate and case-specific data quality validation and interpretation within a *fitness-for-purpose* framework.

2.5.1 Limitations

Several limitations should be discussed. First, this review did not follow a systematic framework. Hence, it will be challenging to repeat it later to assess changes over time, or for it to be replicated by other researchers. Some studies within the defined scope may have been missed by the approach used, as noted previously.[195] However, given the broad scope of this literature review, it would be unfeasible to find, screen, and describe all the relevant research based on the number of studies identified, nor was this the proposed

aim (as discussed by others).[201,202,228] The approach adopted resulted in a multi-pronged review of an extensive number of studies, including some not captured by previous reports (such as a 2019 systematic review on adherence measures using RCD sources,[182] a 2021 systematic review on comparisons of self-reported medication intake with other sources,[197] and a 2022 scoping review of adherence measures in patients with chronic diseases).[111] Also, rather than systematising results on a specific research question, the aim of this review was to describe a large body of literature to guide subsequent analyses and interpretation of results within this thesis.[195] Even if some studies were missed, this is unlikely to have meaningfully altered the overall interpretation of these results. Likewise, no formal hypothesis testing was performed to look for statistically significant differences in agreement results across the categories considered, as exploration of such hypotheses was not within the remit of this review. The extensive heterogeneity of methodological approaches, definitions of drug exposure and medication groups, data sources, as well as geographical settings and populations included would also prevent meaningful formal testing of association of agreement with study characteristics, even within the same medication class, as previously suggested in two recent reviews on similar topics.[111,197]

2.6 Conclusion

Results from this literature review show large heterogeneity within previous studies comparing RCD with other sources of medication exposure, in terms of geographical settings, populations, medications considered, RCD sources and comparators used, and methodological approaches. Assessment of RCD data quality is context dependent, and appropriate validation efforts should be undertaken in the specific settings where each particular data source may be used. Agreement between RCD and other sources was generally higher or similar in trials when compared with observational studies, suggesting

that RCD may also be considered a suitable option for collecting information on medication within trials.

CHAPTER 3: Routinely-collected community prescribing and dispensing data in England: validation and implementation in the RECOVERY trial of COVID-19 treatments among hospitalised patients

3.1 Chapter summary

Aims: To validate, develop, and test the impact of using two novel RCD sources on medication use in the community in England within a randomised controlled trial in acute COVID-19.

Methods: This chapter describes the validation and implementation of two nationwide RCD sources on medication use in the community setting within a subset of 36,202 participants recruited in English sites within the RECOVERY trial of treatments for hospitalised COVID-19 patients. RECOVERY collected information on drug exposure at randomisation for a limited number of medications using CRFs, based on medical records review. To complement this, two RCD sources were assessed: 1) the General Practice Extraction Service Data for Pandemic Planning and Research (GDPPR or "GP") dataset (a focused extract of primary care records), and 2) the Medicines Dispensed in the Community ("Dispensing") dataset (generated from all pharmacy reimbursement claims for medications prescribed in England). To validate use of these sources, data quality was assessed by evaluating completeness (i.e. capture of the study population and drug exposures using the 21 British National Formulary [BNF] chapters and 17 specific drug groups of interest, most of which not captured by the study CRF), and timeliness and stability (i.e. time lags between data recording and reception, and identification of discrepancies resulting from retrospective updates across consecutive data extracts).

Then, different possible algorithms for deriving drug exposure from RCD were developed and tested against CRF records (used as reference) for antiplatelets and oral anticoagulants (as these were the only two CRF-recorded drugs which are commonly used in the community setting). RCD algorithms tested included 1) different lookback observation windows to define drug exposure at randomisation, 2) approaches to handling missing exact dates in the Dispensing dataset (provided as month only), and 3) whether there is additional value in using both RCD sources combined vs only one. Finally, the resulting impact of implementing these RCD sources in possible use-case scenarios in RECOVERY was evaluated in comparison with CRF records (for antiplatelets and oral anticoagulants), and if using one RCD source versus the other. Agreement between sources for binary exposures was assessed using Cohen's κ and complemented with sensitivity, specificity, and PPV/NPV.

Results: Linkage coverage was high, with 91% (32,768/36,202) of the study population captured in both datasets and 98% (35,344/36,202) in at least one. Both RCD sources were able to capture exposure to most drugs of interest, but Dispensing data captured more participants and distinct medications across most drug groups. Dispensing data was more stable across consecutive extracts, and had a 2-month lag between an event and its inclusion in the disseminated data (contrary to fortnightly dissemination in the GP dataset). A 3-month lookback period to identify drug exposure at randomisation maximised agreement between CRF and RCD (antiplatelets: κ=0.76, percent agreement= 93.6%; oral anticoagulants: κ=0.67, percent agreement= 94.4%). This interval also replicates average underlying prescribing and dispensing intervals in each dataset (average 30 days between consecutive records for each participant across most drug groups). With regards to dealing with exact dates, imputing day 15 for Dispensing records combined with including records in the month of randomisation yielded the best agreement between CRF and RCD data for assessments of drug exposure at

randomisation. However, there were small differences in agreement across all rules tested. Compared with Dispensing data only, inclusion of GP data improved agreement with CRF marginally (κ=0.75 to 0.76 for antiplatelets; 0.68 to 0.69 for oral anticoagulants). RCD yielded lower estimates for drug exposure at randomisation versus CRF (antiplatelets: 14.6% RCD vs 16.3% CRF; oral anticoagulants: 8.0% RCD vs 9.9% CRF). Estimates of drug exposure at randomisation and post-randomisation initiation were similar between Dispensing and GP data across most medication groups of interest.

Conclusions: Two novel RCD sources concerning medication use in England provided almost complete coverage of participants in the RECOVERY trial of COVID-19 medications. RCD yielded good agreement versus CRF for drug exposure at randomisation at an individual participant level, and similar overall estimates, while contributing information on medications not captured in the CRF. These results support expanded use of RCD for assessment of drug exposure in the community setting in future RCTs in England.

3.2 Introduction

Collection of accurate information on concomitant medications taken by participants within trial is of paramount importance to assess study eligibility, describe the population, and interpret results.[20,229] While there is no accepted gold-standard for collecting information on concomitant medications, self-reporting has been estimated to be the method most commonly used to complete trial case report forms (as also suggested by the results described in Chapter 2), making it the de facto standard in this setting.[99] However, this approach is time-consuming, subject to variability depending on the interview method used, and may generate inaccurate estimates.[98,100–105]

RCD is that generated as a by-product of routine clinical care, such as administrative drug claims databases.[24,29] Two, nationwide, RCD sources on medication used in the community in England were made available for research purposes for the first time in 2020. These include prescriptions recorded in primary care (for use in COVID-19 research only), and separately medications dispensed in the community and claimed for reimbursement.[230,231] While other sources of data on medications were already available,[232,233] their value for RCTs was limited by incomplete population coverage or barriers to accessing identifiable data on an individual level. Moreover, concerns have been raised about the use of RCD within trials, particularly with respect to data quality, completeness, timeliness, and the need to develop specific methods to handle these data sources for a different purpose than the one they were collected for.[220,234–237]

The aims of this chapter are to 1) assess whether novel RCD sources on medications in England are of appropriate quality to identify drug exposure in the community setting for participants in a trial, 2) identify and develop optimal methods to derive drug exposure from these sources using data-driven procedures, and 3) test the impact of using these RCD sources to derive drug exposure in potential use-case scenarios.

3.3 Methods

3.3.1 Study population, design, and setting

RECOVERY is a RCT of medications for hospitalised COVID-19 patients.[51] The trial is recruiting in all acute hospitals in the UK NHS, and in several other countries. Recruitment started in March 2020. As of mid-2023, RECOVERY had recruited over 48,000 participants, most of which in England.

3.3.2 Data collection

3.3.2.1 Case report form data

Data collection by RECOVERY staff is streamlined, and performed for adult participants in the UK using two simple CRFs, one completed at randomisation and the other during follow-up (the latter to be completed at discharge, death, or 28-days after randomisation, whichever occurs first). The randomisation form records simple clinical and demographic information (such as demographics, comorbidities, and baseline clinical status), and medications being used upon study entry (i.e. at randomisation). This includes drugs commonly used in the community setting (i.e. antiplatelets, oral anticoagulation), which may be continued, discontinued, or initiated during the acute hospital admission, and medications used for in-hospital management of COVID-19 (e.g. corticosteroids, remdesivir, tocilizumab/sarilumab, macrolides, among others). These data were used to both describe baseline characteristics of the study population, and to support time critical eligibility assessments (such as background use of aspirin or oral anticoagulation for participants being considered for inclusion in the aspirin comparison). Of note, research staff completing the CRF had access to their hospital's local prescribing system.

As the established method of data collection on medication exposure in RECOVERY, CRF data were used as reference standard in this chapter (for comparisons versus RCD sources). Therefore, given the focus on RCD on medications in the community setting in

this chapter, CRF records were restricted to those completed upon randomisation, and for antiplatelets and oral anticoagulation only (as these were the only two groups for which reasonable comparisons could be made). Collection of information on these two drugs started on the 6th November 2020, and only complete answers to each question were retained for analysis (i.e. participants with missing records were excluded from comparisons with RCD for those drugs). Chapter 4 will complement the present chapter with a comparative analysis of CRF data versus RCD on inpatient medications in England, for all medications recorded in the RECOVERY randomisation and follow-up CRFs.

3.3.2.2 Routinely-collected data sources

RECOVERY CRF data is complemented with linkage to RCD sources in the UK. Amongst these are the GDPPR (hereafter "GP") dataset,[231] and the Medications Dispensed in Primary Care ("Dispensing") dataset.[230] Both sources are maintained by NHSE, and restricted to data recorded in England. Data linkage was based on NHS number, with subsequent checks for consistency in date-of-birth and sex performed between the CRF data and the received linkage data.

Given the limited scope of CRF data collection on background (i.e. chronic) medications, a list of other drugs of interest to RECOVERY, for which extraction of information from RCD was envisaged, was developed in collaboration with the trial investigators (Table 3.1).

Table 3.1 - Drug groups of interest to RECOVERY selected for analysis

Disease area	Drug
Cardiovascular drugs	Antiplatelets
	Oral anticoagulants
	Antiarrhythmics (class I or III)
	Antidyslipidemics
	Renin-angiotensin-aldosterone system (RAAS) inhibitors
	Sodium-glucose co-transporter-2 (SGLT2) inhibitors
	Insulin
	Antidiabetic drugs (excluding insulin)
Respiratory drugs	Inhaled corticosteroids
	Inhaled bronchodilators
Antimicrobials	Antibacterials
	Antifungals
	Antivirals
Immunosuppression	Systemic immunosuppressive drugs (excluding corticosteroids)
	Systemic corticosteroids
Mental health	Antidepressants
	Antipsychotics

3.3.2.2.1 GP dataset

The GP dataset is generated from fortnightly extractions of primary care data performed by NHSE via the General Practice Extraction Service (GPES). This extraction is based on lists of Systematized Nomenclature of Medicine Clinical Terms (SNOMED) codes of interest (termed clusters or "refsets"), assembled for specific public health purposes (such as surveillance or auditing). A complete list of codes used to generate the GP dataset is available in the NHSE website.[231] GPES data extraction began before the COVID-19 pandemic, and the resulting data were made available exclusively for COVID-19 research as the GP dataset. Some changes to the data collection procedure were implemented later (e.g. inclusion of codes for COVID-19 vaccination in 2021). While there are ongoing consultations on the release of similar nationwide primary care data for other research

purposes, at the moment of writing (early 2024) there are no published plans for release.[238] Among other information (such as demographics, clinical observations, diagnoses, examinations and procedures), GP data include data on medications, such as primary care prescriptions or lists of usual medications. However, GP data are not a complete copy of all recorded primary care data, as only records matching pre-specified clusters of interest are extracted. GPES covers 97.5% of English primary care practices and 93.5% of patients.[231] There is no publicly-available information on internal data quality assessments. The data include a SNOMED code for each record, the date in which a prescription was recorded in the system (record date), and the date to which the record refers (event date) - Table 3.2.

Table 3.2 - GP dataset structure

Field	Description
study_number	Unique RECOVERY participant ID
sex	Sex of patient: 0 Not Known; 1 Male; 2 Female; 9 Not specified
lsoa	Lower super output area of participant residence
date_of_death	Date of death
ethnic	Ethnic origin (as per the code set within the NHS Data Dictionary)
gp_system_supplier	Name of GP system supplier of the source patient record
processed_timestamp	Date that the extract was processed (by NHSE)
reporting_period_end_date	Date that the extract was performed on the source patient record
date	Date which the record refers to
record_date	Date in which the record was created
code	SNOMED code value indicating the nature of the characteristic, event or intervention recorded
sensitive_code	SNOMED code value, which is sensitive in nature, indicating the nature of the characteristic, event or intervention recorded
episode_condition	Whether first or subsequent episode of the coded condition. F First; N New; O Other; D Cause of Death
episode_prescription	Whether a prescription for the coded item is repeat or acute: A Acute (one-off issue); I Issue of repeat; R Repeat authorisation
value1_condition	First numeric value (in a record with CODE for investigation, measurement or result)
value2_condition	Second numeric value (in a record with CODE for blood pressure e.g. diastolic pressure)
value1_prescription	Amount prescribed as number of tablets, capsules, etc. or volume of liquid in mL
value2_prescription	Daily dose prescribed as number of tablets, capsules, etc. or volume of liquid in mL
links	A list of zero, one or many link numbers. A link number is shared by records for the same patient that have been explicitly linked in some way by the use. For example, problem-linkage, links between disorders and associated treatments and outcomes

3.3.2.2.2 Dispensing dataset

The Dispensing dataset is a data collection assembled by the NHS Business Services Authority (NHSBSA) containing all medications prescribed in the community setting in England, and submitted for reimbursement in England/Scotland/Wales/Isle of Man/Jersey/Guernsey/Alderney since April 2018.[230] Of note, submission to NHSBSA by pharmacies and other dispensing bodies is compulsory for reimbursement. The data only provide the month in which a dispensing event was submitted for reimbursement (i.e. there are no exact dates) - Table 3.3. It is not known why exact dates are not included, but in an

ongoing public consultation NHSE have indicated that this information may be included in future versions of the dataset. Each drug record includes both SNOMED and BNF codes. The data include separate medication name, code, quantity, strength, and formulation fields for the item prescribed and the item paid. “Paid” fields are the original information captured by NHSBSA (as the data collection is built from reimbursement claims), with “prescribed” fields extracted directly from the electronic prescription (or inferred in some circumstances for paper prescriptions). Paid and prescribed fields may occasionally differ due to specific handling of dispensed quantities, special containers, or reimbursement coding rules. NHSBSA continuously monitor data quality in this dataset, with coverage close to 100% and >99.9% data capture accuracy from original records.[239]

1 Table 3.3 - Dispensing dataset structure

Field	Value
study_number	Unique RECOVERY participant ID
bsa_prescription_id	NHSBSA identification number for a prescription (can be used to identify an item dispensed when used with 'prescription item identification number')
cost_centre_sub_type	Sub-type of prescribing organisation
cost_centre_type	Type of prescribing organisation
dispensed_pharmacy_type	Type of dispenser
eps_prescription_id	Unique reference for a prescription within the Electronic Prescription Service (EPS)
eps_prescription_indicator	Indicator showing whether the prescription was processed through EPS, (0 no, 1 yes)
item_id	NHSBSA identification number for an item on a prescription (can be used to identify an item dispensed when used with 'prescription identification number')
not_dispensed_indicator	Indicator showing whether the medicine or appliance was dispensed (N no, Y yes)
out_of_hours_indicator	Indicator showing whether the prescription was written out of hours
paid_bnf_code	BNF code of the medicines dispensed, as reimbursed by NHSBSA
paid_bnf_name	BNF name of the medicine dispensed, as reimbursed by NHSBSA
paid_drug_strength	Strength of the medicine dispensed, as reimbursed by NHSBSA
paid_formulation	Formulation of the medicine dispensed, as reimbursed by NHSBSA
paid_quantity	Quantity of medicine dispensed, as reimbursed by NHSBSA
paiddmd_code	dm+d code of the medicine or appliance dispensed, as reimbursed by NHSBSA
patient_age	Patient's age when medicine or appliance was dispensed (calculated value based on prescription or Personal Demographic Service value)
patient_gender	Patient's gender (0 not known / unrecorded, 1 male, 2 female, 9 not specified / indeterminate
prescribed_bnf_code	BNF code of the medicine prescribed
prescribed_bnf_name	Name of the prescribed medicine or appliance
prescribed_formulation	Formulation of the medicine prescribed
prescribed_medicine_strength	Strength of the medicine prescribed
prescribed_quantity	Quantity of medicine prescribed
prescribed_supplier_name	Name of the supplier of the prescribed medicine or appliance
prescribeddmd_code	Dictionary of medicines and devices (dm+d) code of the medicine or appliance prescribed
prescriber_type	Type of prescriber
private_prescription_indicator	Indicator showing whether the prescription was private
processed_period	Month and year prescriptions were submitted to NHSBSA for processing in the format 201807
processing_period_date	Month and year prescriptions were submitted to NHSBSA for processing, displayed as a first of the month date in the format 2018-07-01

3.3.2.2.3 Drug exposure derivation using RCD

Following preliminary data exploration and personal communication with NHSE, drug exposure was extracted using the *code* and *date* fields in the GP dataset and the *prescribed_bnf_code* and *processing_period_date* fields in the Dispensing dataset. For each drug group, two mirror codelists were built (one using SNOMED and the other BNF codes, for use in the GP and the Dispensing datasets respectively). Codelists were assembled using a combination of manual code selection from the GP dataset code clusters reference table,[231] and automatic retrieval via interrogation of the BNF hierarchy within the OpenCodelists platform.[240] Where possible, these lists were merged with relevant publicly-available codelists stored in OpenCodelists. All codelists were built by one clinician (the candidate), and manually reviewed and approved by a second one, and are available for inspection and reuse at https://github.com/gpessoaamorim/dphil-recovery-outpatient-data.

3.3.3 Data analysis

The main RCD extracts used in this analysis were received on the 27th July 2021, with earlier extracts used for data stability assessments (as detailed below). The analysis population was restricted to RECOVERY participants randomised in England ("England cohort") up until the 26th July 2021 (the most recent randomisation date for participants included in both linked data sources). Participants aged <16 years at randomisation were excluded (as RECOVERY did not request linkage data for children), as well as any participants who withdrew any consent for continued participation in the study.

3.3.3.1 Cohort characteristics

General baseline cohort characteristics were extracted from the randomisation CRF. Ethnicity was derived from either GP data or secondary care records from HES.[148,241]

3.3.3.2 Data quality assessment

Data quality for both RCD sources was assessed to evaluate whether they could be considered fit-for-purpose for the specific aim of identifying background exposure to drugs used in the community in the RECOVERY trial.[191,219,242,243] This conceptual validation approach was devised in line with relevant guidance issued by the MHRA, the FDA, and other regulators.[24,33,244] Data quality was considered in a contextual/external (i.e. whether the data are of sufficient quality for a particular intended purpose) vs internal dimension (whether the recorded data accurately represents the reality it aims to capture),[192,242] as it is assumed that publicly-available datasets published by NHSE should have undergone internal data quality checks.[239] As discussed in Chapter 2, multiple approaches have been proposed for RCD validation, but this usually involves comparison with an established standard.[192] Within trials, CRFs recorded by study staff (using either self-reported medications, or medical record review) are the most commonly used method of medication data collection (as outlined in Chapter 2). Moreover, CRF data entail a degree of clinical adjudication by the healthcare professional recording it. CRF records were therefore used as the reference standard for RCD validation and implementation in this analysis. This approach assumes that both RCD and CRF may be used to capture the underlying true drug exposure, that there should be reasonable agreement between both, and that the CRF is a reasonable reference standard in this context. Given its streamlined design, CRF data on use of outpatient medications at randomisation was scarce (restricted to antiplatelets and oral anticoagulants). Therefore, data quality assessments were based not only on comparisons between RCD and CRF records, but also on comparisons between the two RCD sources, with the aim of triangulating all the available data and generating the best possible estimates.

Three data quality features were assessed: 1) completeness/coverage (whether the intended population and drug exposures are captured adequately, and whether there are

any geographical or temporal data gaps), 2) timeliness and stability (how soon a data item is available for analysis after the respective event occurred, and whether the data provided is updated retrospectively – i.e. unstable), and 3) accuracy/validity/comparability – i.e. how well RCD sources agree with a reference standard for the exposure being assessed, and with a collateral RCD source.[24,245–247]

Data coverage and completeness were assessed by calculating the number of RECOVERY participants recruited in England (the England cohort, as defined above) which were captured by each RCD source. General baseline demographic and clinical characteristics were then computed for participants in the England cohort and those captured by each RCD source. Recruitment timing and location were also assessed based on the groups described above, to look for potential geographical gaps along border regions and to check for continuous and consistent data capture over time. Drug coverage was assessed at both broad and specific levels by calculating numbers of participants with a drug record in each data source across the 21 BNF chapters and 17 specific drug groups of interest to RECOVERY (as described in Table 3.1).

To assess data stability, three consecutive monthly extracts from both datasets were used (May 2021 to July 2021), and then restricted to participants included in the earliest extract (to avoid inclusion of data from newly randomised participants in the later extracts). Discrepancies across extracts were inspected by calculating monthly counts of distinct participants, records, and drug codes included in each extract, and within the same time period adequately covered by all extracts (i.e. restricted to the time period covered in the earliest one). Timeliness was assessed by evaluating the time lag between the most recent items recorded in each extract and data reception in each dataset.

Finally, data accuracy was assessed by evaluating the agreement observed between RCD and CRF records, and between both RCD sources. For this purpose, an appropriate

methodological approach needed to first be identified in order to extract drug exposure from both RCD sources, as detailed below.

3.3.3.3 Data-driven methodology development

To inform the implementation of these novel RCD sources within RECOVERY, different drug exposure derivation rules were identified and tested by assessing their impact on the resulting agreement between RCD and CRF records for antiplatelets and oral anticoagulants (and between RCD sources where relevant). The ones yielding the best possible agreement were selected for subsequent use. It should be noted that the RECOVERY CRF enquired about current prescription at the time of randomisation (during an in-hospital stay), while both RCD sources capture medication use in the community setting. However, in the absence of a perfect measure of drug exposure at randomisation, and of established guidance or experience with handling these novel RCD sources, this approach was chosen so that the use of RCD in this context would most closely approximate the only possible reference for drug exposure (the CRF).

Building from preliminary data exploration, the following questions were identified for evaluation:

1) Which lookback window should be used to define baseline drug exposure at randomisation;
2) How should Dispensing records with no exact dates be handled; and
3) Is there added value in combining the two RCD sources.

Using the best performing algorithm (following the three development steps mentioned above), the final agreement between RCD and CRF or between RCD sources was calculated.[100]

3.3.3.3.1 Methodology development – lookback windows

As described in Chapter 2, assessment of drug exposure at a particular time point from medication records is usually performed with either a fixed lookback window method (i.e.

using records within that window as evidence of drug exposure at a particular time point of interest at the end of that period, such as randomisation), or a legend-time/medication-in-hand method (i.e. considering exposure at a particular time point if a patient was expected to be in possession of medication at that date, based on the date they obtained the medication, the quantity prescribed/dispensed, and the daily dose).[96] The latter could not be employed here as the Dispensing dataset does not provide exact dates, and neither dataset provides exact dosages.

To guide algorithm choice, underlying medication use patterns were investigated in each RCD source by computing average time intervals between consecutive prescribing or dispensing records (i.e. interval between two records for the same drug code, for each participant and in each dataset), for participants with at least two records with distinct dates in each dataset, and then aggregated by drug group of interest. This followed a preliminary assessment exploring the number of distinct drug codes recorded for each participant and drug group, which showed that on average most participants only have one distinct drug code recorded for each group (Figure 3.1). Time intervals were calculated using exact record days in the GP dataset, or the first day of the month in the Dispensing dataset (where no exact dates are provided).

Figure 3.1 - Number of distinct drug codes per participant, split by drug group and dataset

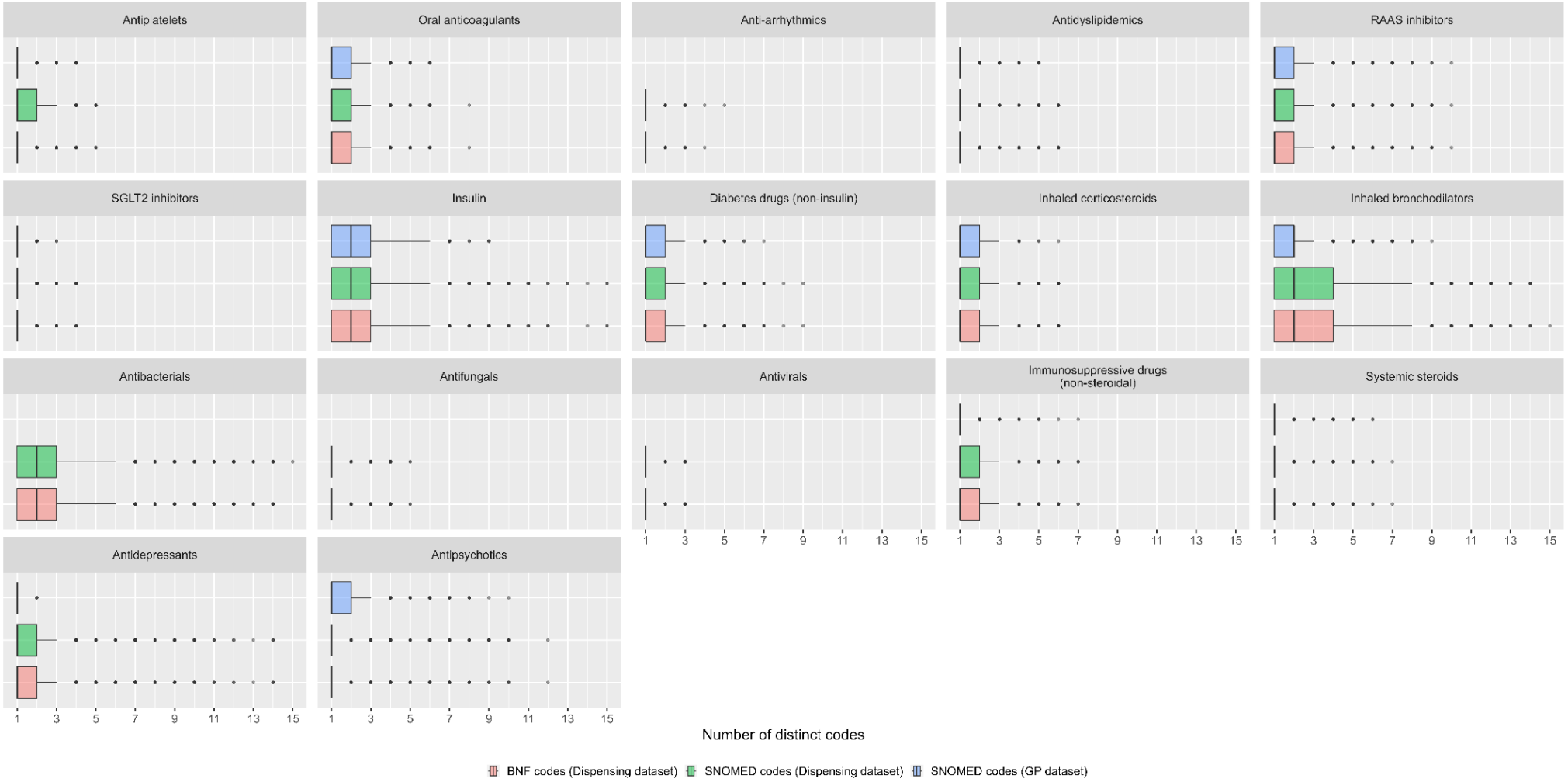


Figure legend: Box plots span from Q1-Q3, with middle black line showing median, and whiskers spanning to 1.5 times the IQR below Q1 or above Q3. Individual dots show outliers beyond whisker margins. Horizontal axes were capped at 15.

Then, for each drug group and dataset, the proportion of participants with average intervals falling within 1, 2, 3, 6, and 12 months was calculated (regardless of when randomisation occurred), as well as the number of participants identified as exposed at randomisation using each window (i.e. based on actual randomisation dates). Next, the agreement between RCD and CRF data using each lookback period was calculated (for antiplatelets and oral anticoagulants only). Based on these assessments, a lookback period was then selected and used in subsequent analyses.

3.3.3.3.2 Methodology development – dispensing date handling

To guide use of Dispensing data (where dates are provided as months only), a combination of different date handling rules were tested, namely: 1) imputing an exact date (day 1, day 15, or last day of the month), and 2) including or ignoring data in the month of randomisation (which, given the lack of exact dispensing dates, cannot be reliably ascertained as occurring before or after randomisation). Dispensing records were transformed according to each rule and then used to ascertain drug exposure as occurring pre- or post-randomisation. Estimates of drug exposure at randomisation resulting from each of those algorithms (using RCD) were then compared with those from CRF records (for antiplatelets and oral anticoagulants only).

3.3.3.3.3 Methodology development – combining collateral RCD sources

The value of using one or both RCD sources was then tested by calculating agreement between CRF records for antiplatelets and oral anticoagulants and the estimates generated using each of the RCD sources in isolation, and both RCD sources combined.

3.3.3.4 Use-case scenarios

Finally, the impact of incorporating the two RCD sources on possible use-case scenarios directly relevant to RECOVERY was assessed by calculating the agreement between CRF and RCD at an individual participant level, and aggregate estimates of proportions of

participants identified as exposed to antiplaletets or oral anticoagulants using CRF records versus each of the RCD sources. The same approach was applied to comparisons between both RCD sources for all medication groups of interest (i.e. including medications not captured by the RECOVERY CRF), for the assessment of both baseline exposure and post-randomisation initiation (including both fact and timing of drug initiation). The latter was defined on the basis of a relevant drug record captured after randomisation for people who were not on that drug at randomisation (as derived using the optimal drug exposure derivation methodology selected above).

3.3.3.5 Analysis periods

Following preliminary data explorations and timeliness/stability assessments (see Results section 3.4.1.3), RCD sources were restricted to June 2018 - May 2021 to account for different inception dates and data reporting lags. For all assessments of drug exposure post-randomisation, the analysis population was restricted to participants who could have completed 6 months (180 days) follow-up by the 31st May 2021, in line with the pre-specified timeframe for long-term outcome assessment in RECOVERY.[241]

3.3.3.6 Statistical analyses

Continuous variables are presented with mean and SD or median and IQR, as appropriate. Distribution normality was assessed visually. Agreement for binary drug exposure estimates (yes/no) was assessed using Cohen's κ, which represents the proportion of agreement observed beyond that expected by chance.[248] Percent agreement (i.e. the sum of true positives and true negatives over the total number of observations) was used to complement κ coefficients. For each comparison, counts and proportions of individuals identified in each source are also presented. For comparisons between RCD and CRF data, sensitivity, specificity, and PPV and NPV were also calculated, using CRF as reference. For assessments of time since randomisation (in post-randomisation drug initiation events),

agreement between RCD sources was evaluated by calculating mean differences between GP and Dispensing records (number of days after randomisation in the Dispensing dataset minus number of days after randomisation in the GP dataset), together with intraclass correlation coefficients (ICC). ICCs were computed using two-way mixed effects models, agreement between sources (vs measurement consistency), and single measurements as unit of analysis (vs average of several measurements), following previously published methodology.[135,214,249,250] κ coefficients and ICCs are presented along with the respective 95% confidence intervals (95% CI).

3.3.3.7 Data and code

R v4.2.1 was used for all data manipulation, analysis, and visualisation. All statistical programming code used in this chapter is available for inspection and reuse at https://github.com/gpessoaamorim/dphil-recovery-outpatient-data. Access to the RECOVERY trial data can be requested via the Infectious Diseases Data Observatory (https://www.iddo.org/covid19/data-sharing/accessing-data). Linkage to the RCD sources used here can be requested for other research purposes from NHSE at https://digital.nhs.uk/services/data-access-request-service-dars.

3.3.4 Ethics

The RECOVERY trial (ISRCTN number: 50189673, ClinicalTrials.gov number: NCT04381936) is sponsored by the University of Oxford and has been approved by the UK Medicines and Healthcare products Regulatory Agency and the Cambridge East Research Ethics Committee (reference 20/EE/0101). All RECOVERY participants provided written informed consent to take part in the trial, including for linkage to the RCD sources used here. Permission to use RECOVERY data for this study was given by the trial Chief Investigators and Steering Committee.

3.4 Results

3.4.1 Study population

The study population (England cohort) included 36,202 RECOVERY participants, mean age 63 years (SD 15), 37% female. Baseline characteristics of the study population are shown in Table 3.4.

3.4.2 Data quality assessment

3.4.2.1 Linkage coverage

Baseline characteristics of study participants captured by each RCD source are shown in Table 3.4.

Table 3.4 - Baseline cohort characteristics, by linkage status

	England cohort	GP dataset cohort	Dispensing dataset cohort	GP dataset cohort but not in Dispensing dataset cohort	Dispensing dataset cohort but not in GP dataset cohort	Non-linked cohort (Dispensing-/ GP-)
Characteristic	N = 36,202	N = 34,140	N = 33,972	N = 1,372	N = 1,204	N = 858
Age in years, mean (SD)	63 (15)	63 (15)	64 (15)	49 (12)	66 (15)	55 (16)
Female	13,341 (37%)	12,628 (37%)	12,865 (38%)	223 (16%)	460 (38%)	253 (29%)
Ethnicity[1]						
White	28,889 (82%)	27,570 (82%)	27,608 (83%)	911 (70%)	949 (86%)	370 (68%)
Black	1,338 (3.8%)	1,245 (3.7%)	1,236 (3.7%)	65 (5.0%)	56 (5.1%)	37 (6.8%)
Asian	3,717 (11%)	3,572 (11%)	3,421 (10%)	220 (17%)	69 (6.2%)	76 (14%)
Other	851 (2.4%)	778 (2.3%)	721 (2.2%)	80 (6.1%)	23 (2.1%)	50 (9.2%)
Mixed	440 (1.2%)	421 (1.3%)	403 (1.2%)	28 (2.1%)	10 (0.9%)	9 (1.7%)
Unknown	967 (2.7%)	554 (1.6%)	583 (1.7%)	68 (5.0%)	97 (8.1%)	316 (37%)
Diabetes	9,637 (27%)	9,167 (27%)	9,428 (28%)	50 (3.6%)	311 (26%)	159 (19%)
Chronic heart disease	8,497 (23%)	8,064 (24%)	8,345 (25%)	33 (2.4%)	314 (26%)	119 (14%)
Chronic lung disease	8,049 (22%)	7,612 (22%)	7,896 (23%)	40 (2.9%)	324 (27%)	113 (13%)
Severe renal impairment	2,159 (6.0%)	2,046 (6.0%)	2,104 (6.2%)	25 (1.8%)	83 (6.9%)	30 (3.5%)

Table legend: [1] Proportions calculated separately for people with known and unknown ethnicity

The GP dataset included 34,140 and the Dispensing dataset 33,972 participants. Figure 3.2 depicts coverage of the England cohort in each dataset: 32,768 participants (90.5% of the England cohort) were included in both linked datasets, and 858 (2.4%) in none.

Figure 3.2 – Study population linkage coverage and overlap between GP and Dispensing datasets

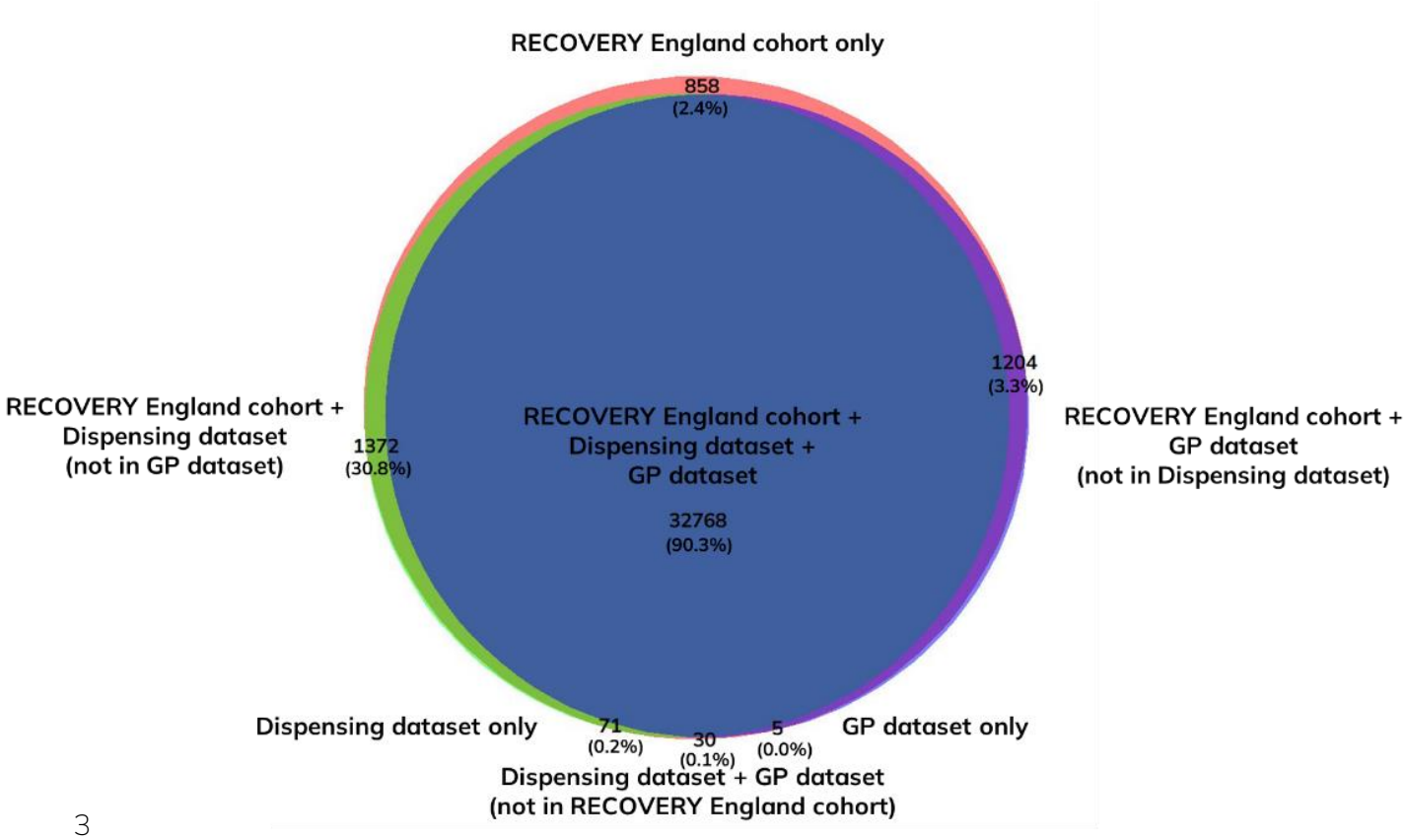


Figure legend: Areas are proportional to the number of participants included in each group. Proportions shown are calculated based on the total number of participants in either the England cohort (n=36,202), the GP dataset cohort (n=34,140), or the Dispensing dataset cohort (n=33,972); total n=36,560. The "Dispensing dataset only" and "GP dataset only" cohorts represent participants included in each of those datasets but not in the other neither in the RECOVERY England cohort (i.e. RECOVERY participants recruited elsewhere in the UK who have available GP or Dispensing data).

Non-linked participants were younger (mean age 55 years, SD 16), more ethnically diverse, and had fewer comorbidities than those with linkage data available (Table 3.1). Participants in this group were recruited in RECOVERY sites spread across England (Figure 3.3), and throughout the duration of the study (Figure 3.4).

Figure 3.3 - Geographical distribution of RECOVERY participants, by linkage status

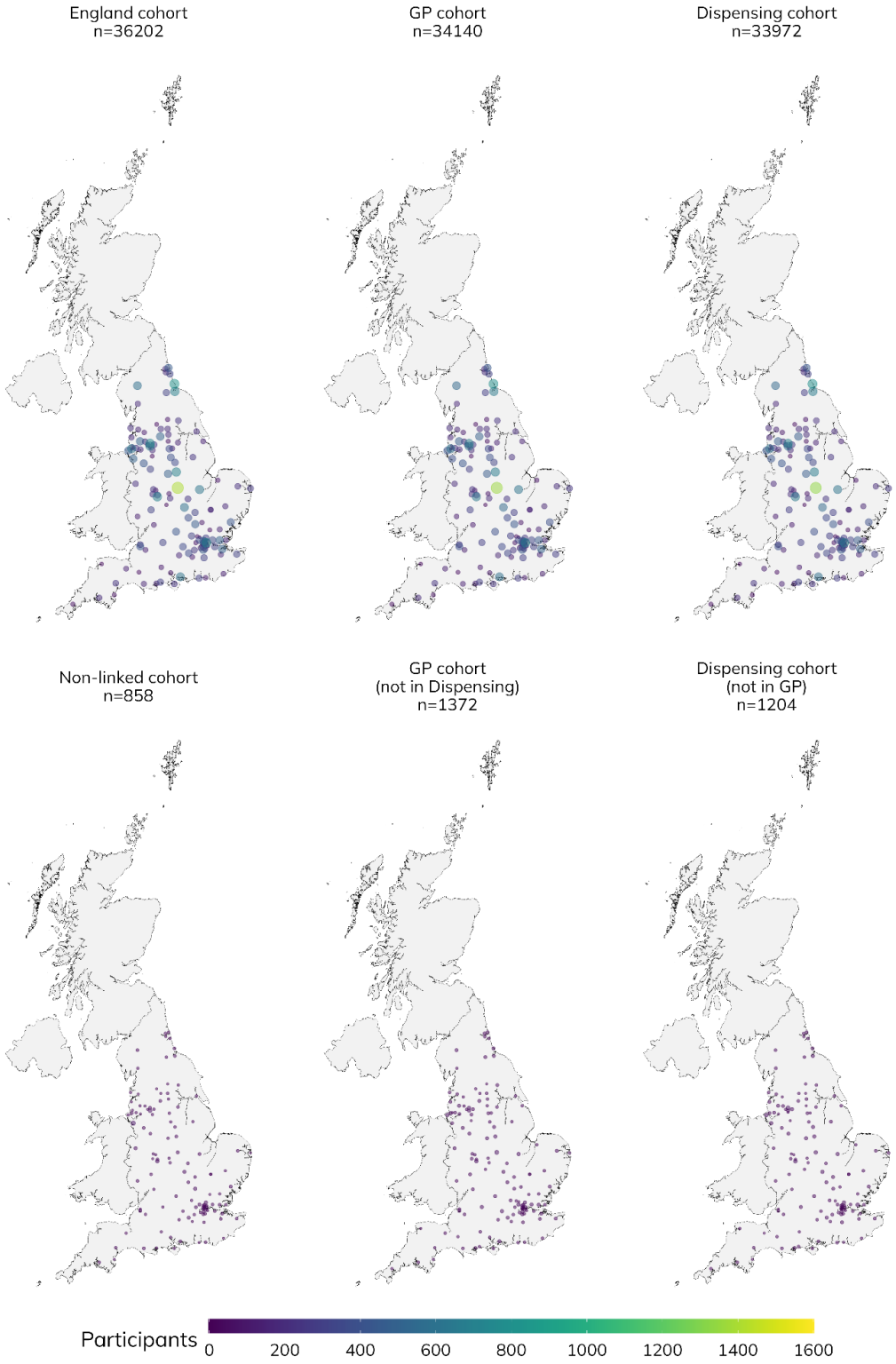
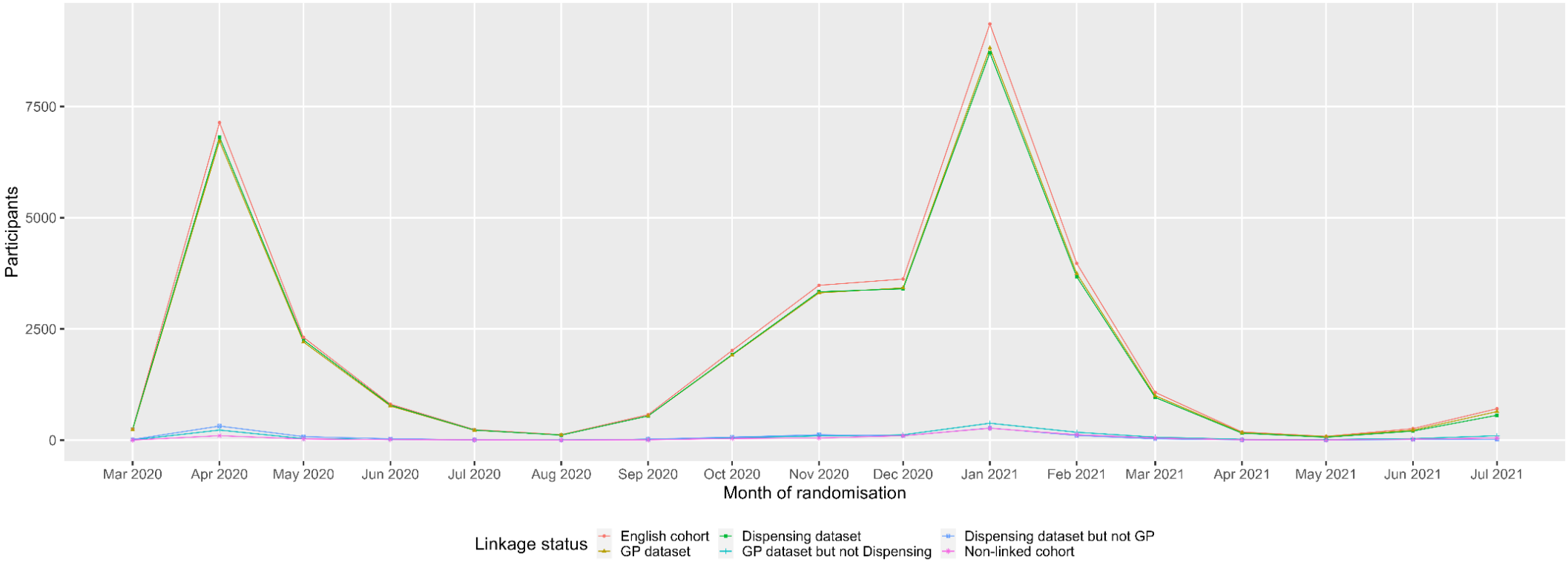


Figure legend: Dots placed at the geographical location of each RECOVERY recruiting hospital in England

Figure 3.4 - Monthly randomisation counts of RECOVERY participants, split by dataset linkage status

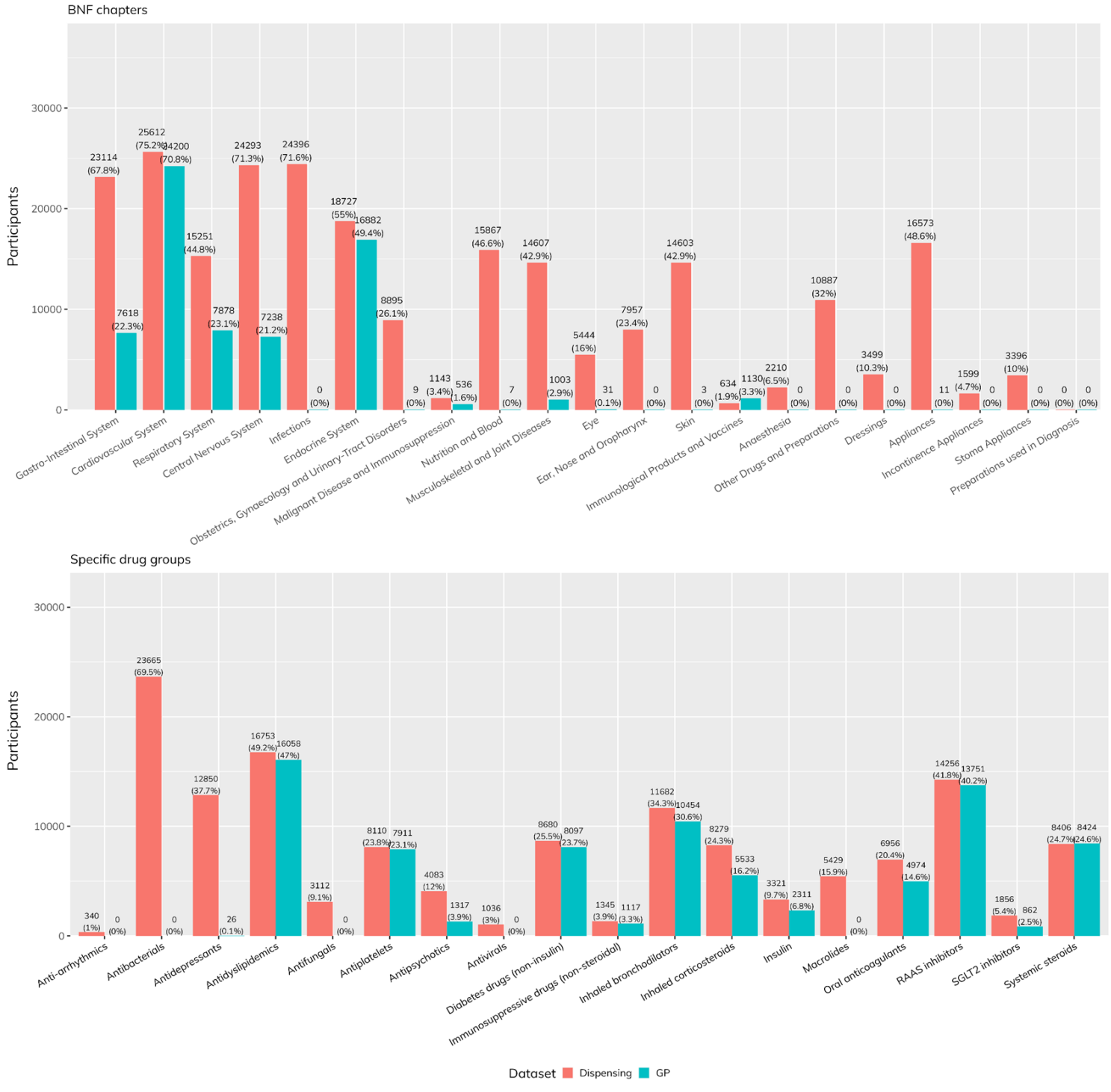


People captured in the GP dataset but not in the Dispensing dataset (n=1,372) were also younger and had lower prevalence of comorbidities than those in the overall England cohort. Manual review of GP data within this group showed mostly routine clinical checks, blood tests, and vaccinations, with a minority of diagnostic or medication codes (mainly for asthma or COVID-19). Conversely, people captured in the Dispensing dataset but not the GP dataset (n=1,204) were older and had more comorbidities than those in the overall England cohort, but a similar ethnic breakdown. Dispensing records for these participants included codes for protein-pump inhibitors (which are not captured by the GP dataset), but also some medications which were expected to be captured in this dataset (based on the scope of the GPES data extraction; e.g. antiplatelets, diuretics, calcium-channel blockers),[231] along with diagnostic primary care records for the relevant conditions. Both linkage datasets also included RECOVERY participants recruited in the UK but not in England.

3.4.2.2 Drug coverage

GP data captured mostly drugs from the Cardiovascular, Respiratory, and Endocrine disease BNF chapters (the latter including Diabetes) - Figure 3.5. Of note, GP data did not include any records for some chapters, including Infections. Dispensing data included records for all BNF chapters except “Preparations used in Diagnosis”, and included more participants across most chapters than Dispensing data, with the exception of “Immunological Products and Vaccines”. Dispensing data also captured more participants across all specific drug groups assessed. Importantly, some medications of interest such as antimicrobials (all types), anti-arrhythmics, and mental health drugs were poorly or not captured entirely by GP data. There were no obvious gaps in drug capture over time in either data source (data not shown).

Figure 3.5 - Capture of broad (BNF chapters) and specific drug groups in the linkage datasets



3.4.2.3 Timeliness and stability

The Dispensing dataset included records from April 2018 to May 2021 (Figure 3.6). The GP dataset included records occurring before the year 2000, with the majority of records occurring between June 2018 and June 2021. Figure 3.6 depicts monthly counts of individual participants, entries, and distinct drug codes recorded within 2021 in 3 consecutive data extracts for each dataset (May, June, and July 2021). In the Dispensing dataset, the use of updated data extracts did not change data retrospectively (i.e. monthly counts for the period included in the earliest extract remain the same). However, there was a 2-month lag from the occurrence of a dispensing event to it becoming available in the disseminated extract (e.g. the most recent dispensing records included in the May extract occurred in March).

Conversely, for GP data there was a shorter lag from an event occurring to its dissemination. However, the use of newer extracts could lead to a retrospective increase in data recorded up to the point of dissemination in the earlier extract (e.g. using the June 2021 extract would increase the number of distinct entries and drug codes in April 2021 vs the May 2021 extract). Of note, the retrospective increase in distinct entries and drug codes could span back several years, up to mid-2019. NHSE have informed, by personal communication, that these observations could be due to changes in the data extraction procedures (i.e. inclusion or exclusion of clinical codes from the clusters/refsets if deemed clinically appropriate).

Figure 3.6 - Monthly counts of distinct participants, entries, and drug codes per dataset across consecutive data extracts

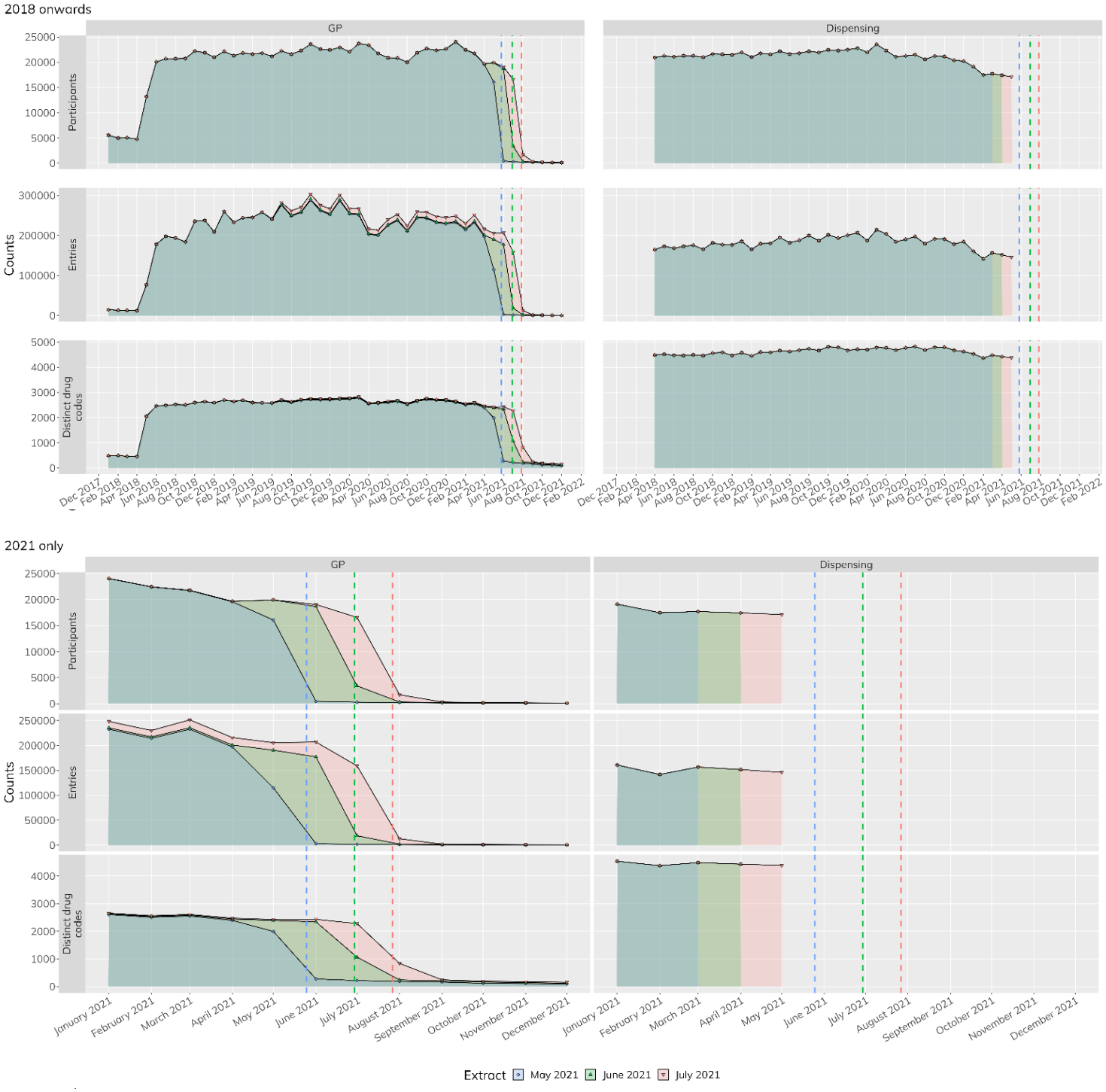


Figure legend: The area in each horizontal panel depicts monthly counts for a different metric (total participants, total entries, or distinct drug codes), split by monthly extract (colour codes). Individual panels are presented for each metric and dataset. The vertical distance between dots in a single month represents records added retrospectively in subsequent extracts. Vertical dashed lines show the date in which each extract was received.

3.4.3 Data-driven methodology development

3.4.3.1 Lookback windows

To guide the use of the two novel RCD sources, different data-driven algorithms were developed and tested, starting with the use of different lookback periods to define drug exposure at randomisation. Figure 3.7 shows average intervals between consecutive prescribing or dispensing records (within the GP or Dispensing datasets respectively), for each dataset and drug group of interest. For most drugs, average intervals were approximately 30 days in both datasets, except for all antimicrobial groups and systemic steroids (with larger intervals of 40-90 days).

Figure 3.7 - Average time interval between consecutive drugs records in each dataset

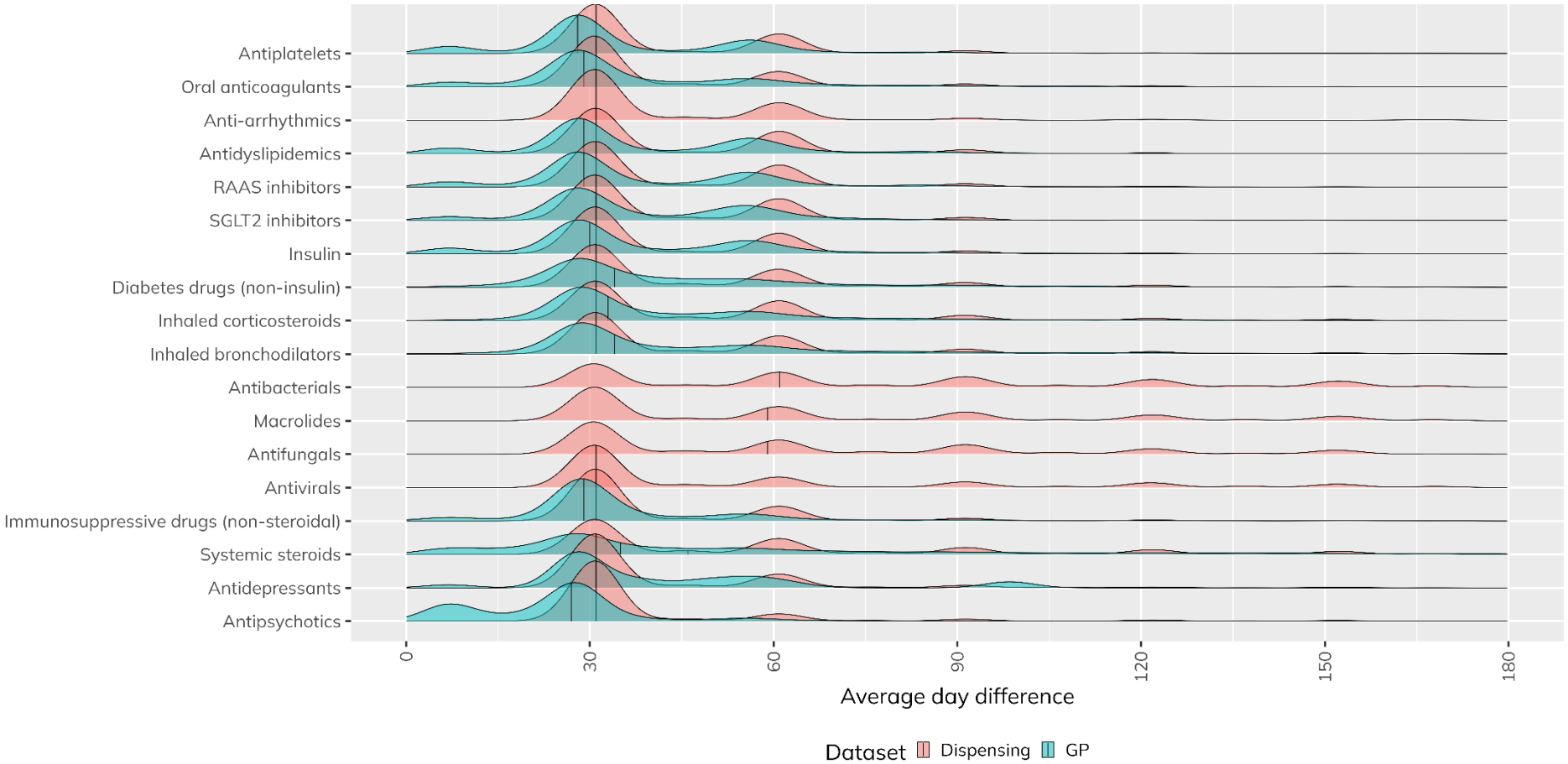


Figure legend: Average day differences calculated for each distinct code, date, and participant, and shown as medians aggregated per drug group and dataset. Vertical black lines show medians for each group. Horizontal axis capped at 180.

Based on these intervals, approximately 90% of all participants would have a repeat prescribing or dispensing event recorded within a 90-day period for most medications which are commonly used on a repeat basis (Figure 3.8). This proportion was lower for drugs used acutely or sporadically, such as antimicrobials or systemic steroids (i.e. those with larger average intervals between consecutive records, as shown above).

Figure 3.8 – Cumulative proportions of participants with different average time intervals between consecutive drug records, by dataset

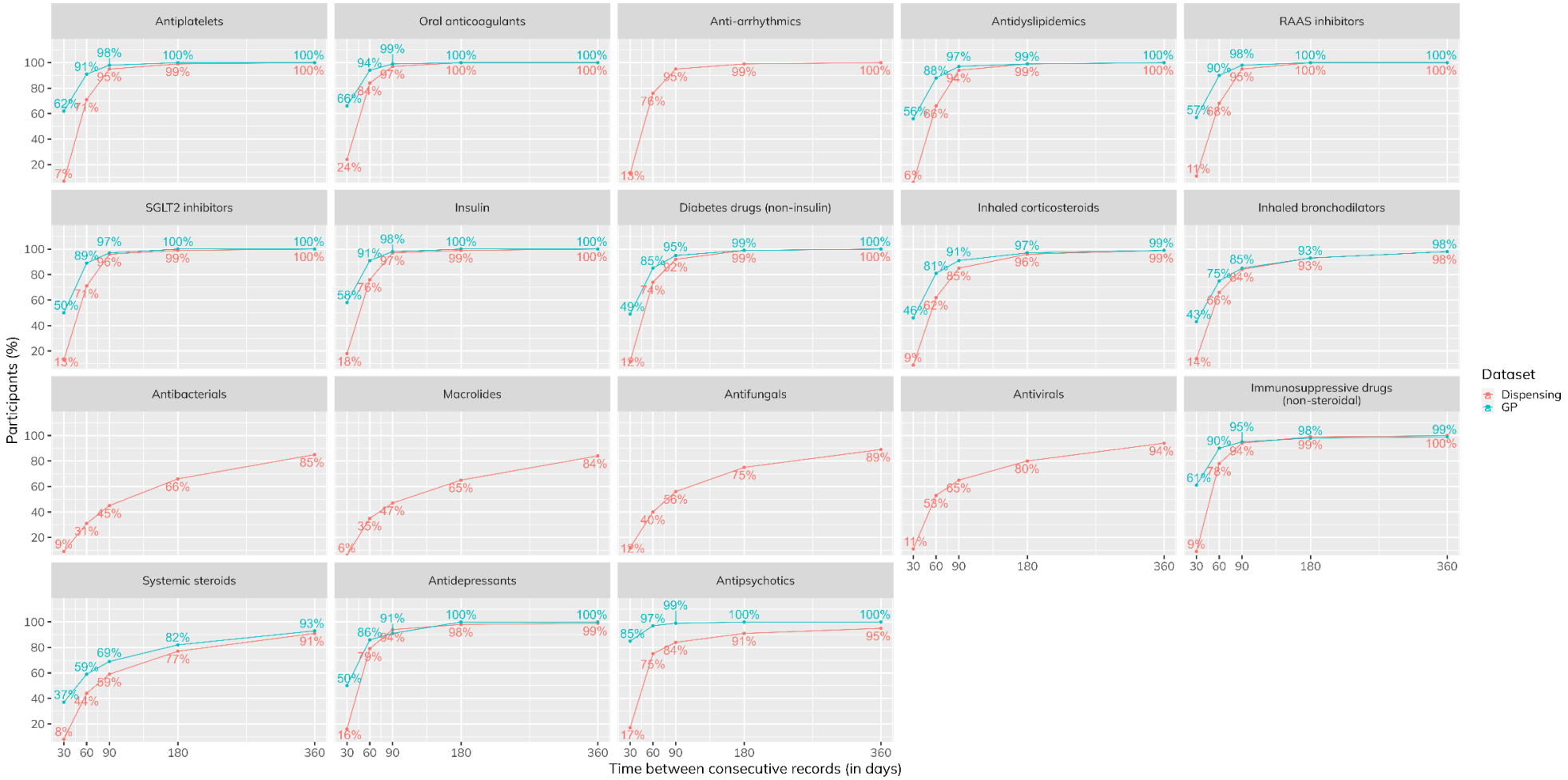


Figure legend: Each point shows the cumulative proportion of participants for whom the average time interval between consecutive drug records for the same drug code is less than 30, 60, 90, 180, or 360 days, split by drug group and dataset

CRF records for antiplatelets and oral anticoagulants were then used as reference to guide choice of a lookback period (using data from 22,295 and 22,382 participants with the relevant question answered in the CRF for antiplatelets and oral anticoagulants, respectively). The use of a 3-month lookback period maximised the observed agreement between CRF and both RCD sources combined (antiplatelets: κ=0.67, percent agreement=93.6%; oral anticoagulants: κ=0.76, percent agreement=94.4%; Table 3.5). For both drugs, sensitivity was lower than specificity (antiplatelets: sensitivity 76.3% vs specificity 97%; oral anticoagulants: sensitivity 66.1% vs specificity 97.6%), and NPVs larger than PPVs (antiplatelets: PPV 83.3% vs NPV 95.4%; oral anticoagulants: PPV 75.2% vs NPV 96.3%). Longer lookback periods yielded higher sensitivity at the expense of lower specificity. Based on these observations, a 3 month lookback period was used for subsequent derivation of baseline exposure using the two RCD sources.

Table 3.5 - Agreement for baseline drug exposure between case report form (CRF) and routinely-collected data (RCD) using different lookback periods before randomisation

		Participants in each source (%)				Agreement metrics				
Drug group	Period	CRF and RCD	CRF only	RCD only	Neither	Sn	Sp	PPV	NPV	K
Antiplatelets	1 month	2,459 (11%)	1,180 (5.3%)	448 (2%)	18,208 (81.7%)	67.6%	97.6%	84.6%	93.9%	0.71 (0.70 : 0.72)
	2 months	2,712 (12.2%)	927 (4.2%)	515 (2.3%)	18,141 (81.4%)	74.5%	97.2%	84%	95.1%	0.75 (0.74 : 0.76)
	3 months	2,776 (12.5%)	863 (3.9%)	556 (2.5%)	18,100 (81.2%)	76.3%	97%	83.3%	95.4%	0.76 (0.75 : 0.77)
	6 months	2,902 (13%)	737 (3.3%)	1,118 (5%)	17,538 (78.7%)	79.7%	94%	72.2%	96%	0.71 (0.69 : 0.72)
	12 months	2,906 (13%)	733 (3.3%)	1,124 (5%)	17,532 (78.6%)	79.9%	94%	72.1%	96%	0.71 (0.69 : 0.72)
Oral anticoagulants	1 month	1,307 (5.8%)	917 (4.1%)	412 (1.8%)	19,746 (88.2%)	58.8%	98%	76%	95.6%	0.63 (0.61 : 0.65)
	2 months	1,445 (6.5%)	779 (3.5%)	461 (2.1%)	19,697 (88%)	65%	97.7%	75.8%	96.2%	0.67 (0.65 : 0.69)
	3 months	1,470 (6.6%)	754 (3.4%)	486 (2.2%)	19,672 (87.9%)	66.1%	97.6%	75.2%	96.3%	0.67 (0.66 : 0.69)
	6 months	1,518 (6.8%)	706 (3.2%)	699 (3.1%)	19,459 (86.9%)	68.3%	96.5%	68.5%	96.5%	0.65 (0.63 : 0.67)
	12 months	1,519 (6.8%)	705 (3.1%)	702 (3.1%)	19,456 (86.9%)	68.3%	96.5%	68.4%	96.5%	0.65 (0.63 : 0.67)

Table legend: Analysis restricted to participants with at least one RCD source available and with the corresponding question (antiplatelets or anticoagulants) recorded in the randomisation CRF. Drug exposure collected from either RCD sources (GP and Dispensing) and calculated using exact dates in the GP dataset and calendar months in the Dispensing dataset. Dispensing records in the month of randomisation were excluded.

3.4.3.2 Handling Dispensing record dates

In the absence of exact record dates for Dispensing records (provided only as monthly periods), three possible rules were tested to guide imputing of exact dates (i.e. first day of the month, day 15, or last day of the month). Simultaneously, two possible rules for handling records in the month of randomisation were tested (i.e. including or excluding these records from drug exposure assessments). Table 3.6 shows the impact of using each rule combination on the observed agreement between CRF and Dispensing records for antiplatelet and oral anticoagulant exposure at randomisation. All rules yielded similar agreement between sources (antiplatelets: κ=0.74 to 0.76; oral anticoagulants: κ=0.67 to 0.69). Estimates for drug exposure at randomisation were similar regardless of the derivation rule chosen (antiplatelets: 13.9% to 14.9%; oral anticoagulants: 7.4% to 8.1%).

Table 3.6 - Agreement for baseline drug exposure between case report form (CRF) and routinely-collected data (RCD) using different lookback periods before randomisation

	Derivation rules		Participants in each source (%)				Agreement				
Drug group	Month of randomisation	Dispensing date	CRF and RCD	CRF only	RCD only	Neither	Sn	Sp	PPV	NPV	κ (95% CI)
Antiplatelets	Ignore	Day 1	2,625 (11.8%)	1,014 (4.5%)	476 (2.1%)	18,180 (81.5%)	72.1%	97.4%	84.7%	94.7%	0.74 (0.73 : 0.75)
		Day 15	2,711 (12.2%)	928 (4.2%)	507 (2.3%)	18,149 (81.4%)	74.5%	97.3%	84.2%	95.1%	0.75 (0.74 : 0.77)
		Last day of month	2,773 (12.4%)	866 (3.9%)	534 (2.4%)	18,122 (81.3%)	76.2%	97.1%	83.9%	95.4%	0.76 (0.75 : 0.77)
	Include	Day 1	2,740 (12.3%)	899 (4%)	513 (2.3%)	18,143 (81.4%)	75.3%	97.3%	84.2%	95.3%	0.76 (0.75 : 0.77)
		Day 15	2,766 (12.4%)	873 (3.9%)	521 (2.3%)	18,135 (81.3%)	76%	97.2%	84.1%	95.4%	0.76 (0.75 : 0.77)
		Last day of month	2,781 (12.5%)	858 (3.8%)	535 (2.4%)	18,121 (81.3%)	76.4%	97.1%	83.9%	95.5%	0.76 (0.75 : 0.77)
Oral anticoagulants	Ignore	Day 1	1,368 (6.1%)	856 (3.8%)	296 (1.3%)	19,862 (88.7%)	61.5%	98.5%	82.2%	95.9%	0.68 (0.66 : 0.69)
		Day 15	1,396 (6.2%)	828 (3.7%)	313 (1.4%)	19,845 (88.7%)	62.8%	98.4%	81.7%	96%	0.68 (0.66 : 0.70)
		Last day of month	1,424 (6.4%)	800 (3.6%)	325 (1.5%)	19,833 (88.6%)	64%	98.4%	81.4%	96.1%	0.69 (0.67 : 0.71)
	Include	Day 1	1,461 (6.5%)	763 (3.4%)	484 (2.2%)	19,674 (87.9%)	65.7%	97.6%	75.1%	96.3%	0.67 (0.65 : 0.69)
		Day 15	1,447 (6.5%)	777 (3.5%)	354 (1.6%)	19,804 (88.5%)	65.1%	98.2%	80.3%	96.2%	0.69 (0.67 : 0.71)
		Last day of month	1,425 (6.4%)	799 (3.6%)	325 (1.5%)	19,833 (88.6%)	64.1%	98.4%	81.4%	96.1%	0.69 (0.67 : 0.71)

Table legend: Routinely-collected data includes Dispensing data only. Dispensing records extracted from 3 months pre-randomisation. Dispensing records assigned to the day of randomisation counted as baseline exposure (pre-randomisation)

Based on these assessments and the small differences found, imputing day 15 and including records in the month of randomisation in drug exposure assessments was selected as a pragmatic approach to handling Dispensing records. The final algorithm selected for defining drug exposure was therefore the following:

- All Dispensing records imputed as occurring on day 15
- Dispensing records in the month of randomisation included in drug exposure assessments if the imputed date falls after the date of randomisation (for identification of both baseline and drug initiation)
- Any records occurring within 90 days before randomisation considered as baseline exposure (in both GP and Dispensing records)

3.4.3.3 Combining different RCD sources

Table 3.7 shows the impact of using the two available RCD sources on the agreement between RCD and CRF data for baseline exposure to antiplatelets and oral anticoagulants, using the algorithm specified above. Agreement between GP data and CRF records was lower than that for Dispensing data vs CRF (κ: 0.73 GP vs 0.75 Dispensing for antiplatelets, and 0.52 vs 0.68 for anticoagulants respectively). When compared with the use of Dispensing data alone, the inclusion of GP data yielded only marginal improvements in agreement between RCD and CRF (increase in κ of 0.01 for both drugs). Using the algorithm which maximises agreement between RCD and CRF records (i.e. combining both RCD sources), the overall κ coefficient was 0.76 for antiplatelets, with sensitivity 71%, specificity 94%, PPV 72%, and NPV 94%. For oral anticoagulants, κ was 0.69, sensitivity 62%, specificity 97%, PPV 68%, and NPV 96%.

Table 3.7 - Impact of using different data sources on the agreement for baseline antiplatelet and oral anticoagulant exposure between case report form (CRF) and routinely-collected data (RCD)

		Participants in each source (%)				Agreement				
Drug group	Dataset	CRF and RCD	CRF only	RCD only	Neither	Sn	Sp	PPV	NPV	κ (95% CI)
Antiplatelets	GP	2,736 (12.3%)	1,732 (7.8%)	676 (3%)	17,151 (76.9%)	61.2%	96.2%	80.2%	90.8%	0.73 (0.72 : 0.75)
	Dispensing	2,893 (13%)	1,235 (5.5%)	1,104 (5%)	17,063 (76.5%)	70.1%	93.9%	72.4%	93.3%	0.75 (0.74 : 0.76)
	Either	2,906 (13%)	1,180 (5.3%)	1,124 (5%)	17,085 (76.6%)	71.1%	93.8%	72.1%	93.5%	0.76 (0.74 : 0.77)
Oral anticoagulants	GP	1,000 (4.5%)	1,538 (6.9%)	261 (1.2%)	19,583 (87.5%)	39.4%	98.7%	79.3%	92.7%	0.52 (0.50 : 0.55)
	Dispensing	1,510 (6.7%)	939 (4.2%)	695 (3.1%)	19,238 (86%)	61.7%	96.5%	68.5%	95.3%	0.68 (0.66 : 0.70)
	Either	1,519 (6.8%)	917 (4.1%)	702 (3.1%)	19,244 (86%)	62.4%	96.5%	68.4%	95.5%	0.69 (0.67 : 0.70)

3.4.4 Use-case scenarios

If used to ascertain drug exposure at randomisation, RCD would lead to similar, but lower aggregate estimates when compared with CRF records for antiplatelets (14.6% RCD vs 16.3% CRF) and oral anticoagulants (8.0% RCD vs 9.9% CRF)- Figure 3.9. However, comparisons at an individual participant level showed that using RCD would also lead to important discrepancies for exposure to both drugs at randomisation (Figure 3.10). For example, 524 participants would be considered exposed to aspirin (14.3% of the total identified from the CRF) and 351 (15.8%) to oral anticoagulants, when no exposure was recorded in the CRF (i.e. false-positives).

For broader drug groups of interest (including those not recorded in the RECOVERY CRF), using Dispensing data alone to identify drug exposure at randomisation would lead to similar aggregate estimates when compared with the inclusion of GP data for most medication groups (i.e. a relative increase of <5%; Figure 3.11). Exceptions were inhaled corticosteroids (579 additional participants out of 6,299 potentially identified if using both data sources combined, or 9.2% relative increase), non-steroidal immunosuppressive drugs (72/1069, 6.7%), and systemic steroids (497/4193, 11.9%). Agreement between RCD sources was high (i.e. κ >0.60) across all drug groups, except inhaled corticosteroids and antidepressants (κ=0.58 and 0.00, respectively; Figure 3.11 and Table 3.8). Similar results were found for drug initiation within 180 days after randomisation (Figure 3.12). However, here the inclusion of GP data would have a larger impact, namely the inclusion of 5.7% additional participants initiating antiplatelets (23/401), 6.4% inhaled corticosteroids (33/497), 37.5% non-steroidal immunosuppressive drugs (27/72), and 8.4% systemic steroids (58/678). Agreement between sources was high for most drugs (i.e. κ>0.60; Table 3.9), except inhaled corticosteroids, non-steroidal immunosuppressive

drugs, and antipsychotics (κ=0.56, 0.42, and 0.32 respectively), with very close estimates for exact timing of drug initiation (ICC=0.93 to 0.99).

Figure 3.9 – Aggregate estimates for drug exposure at randomisation by ascertainment source (antiplatelets and oral anticoagulants)

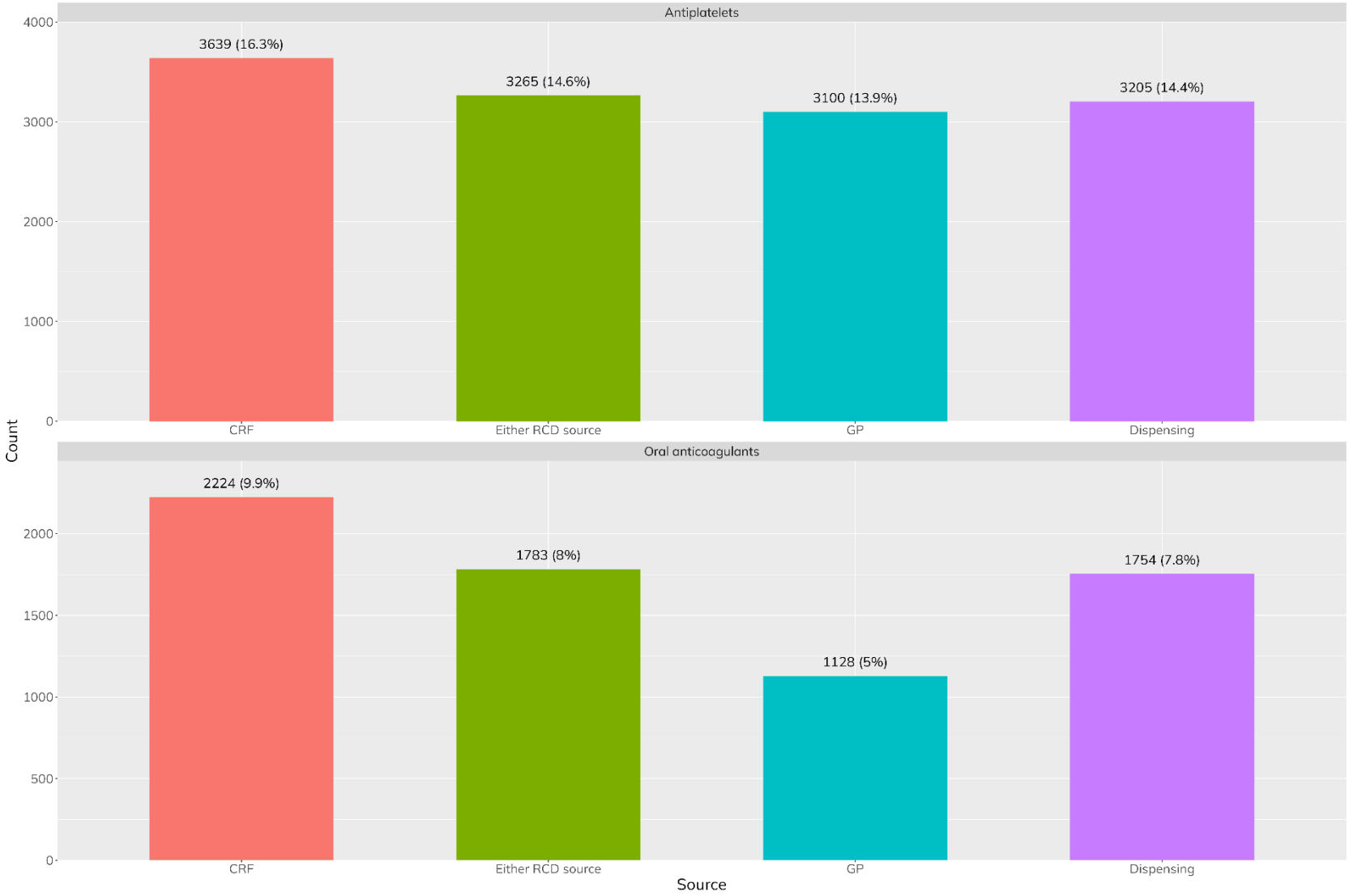


Figure 3.10 - Overlap between case report form (CRF) and routinely-collected data (RCD) sources for antiplatelet and oral anticoagulant exposure at randomisation at an individual participant level

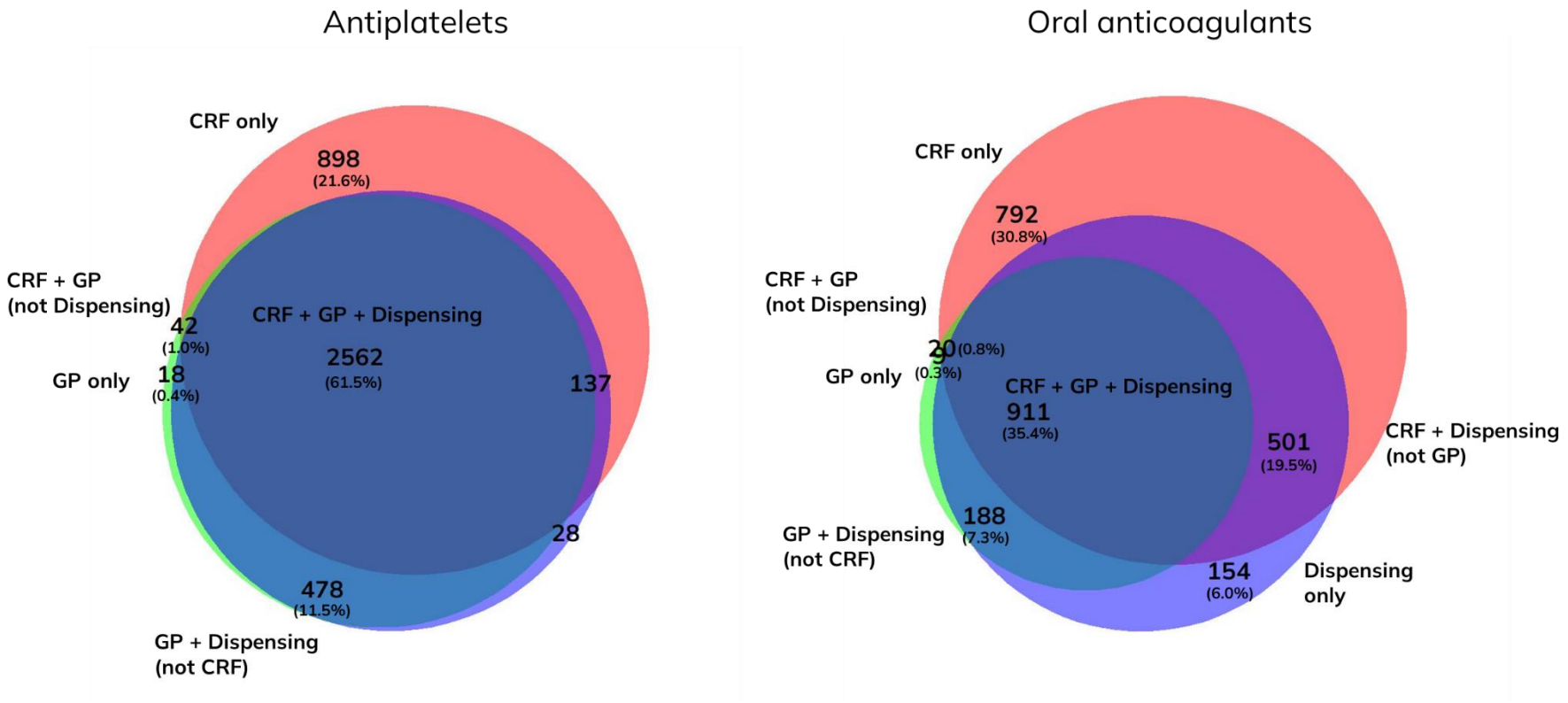
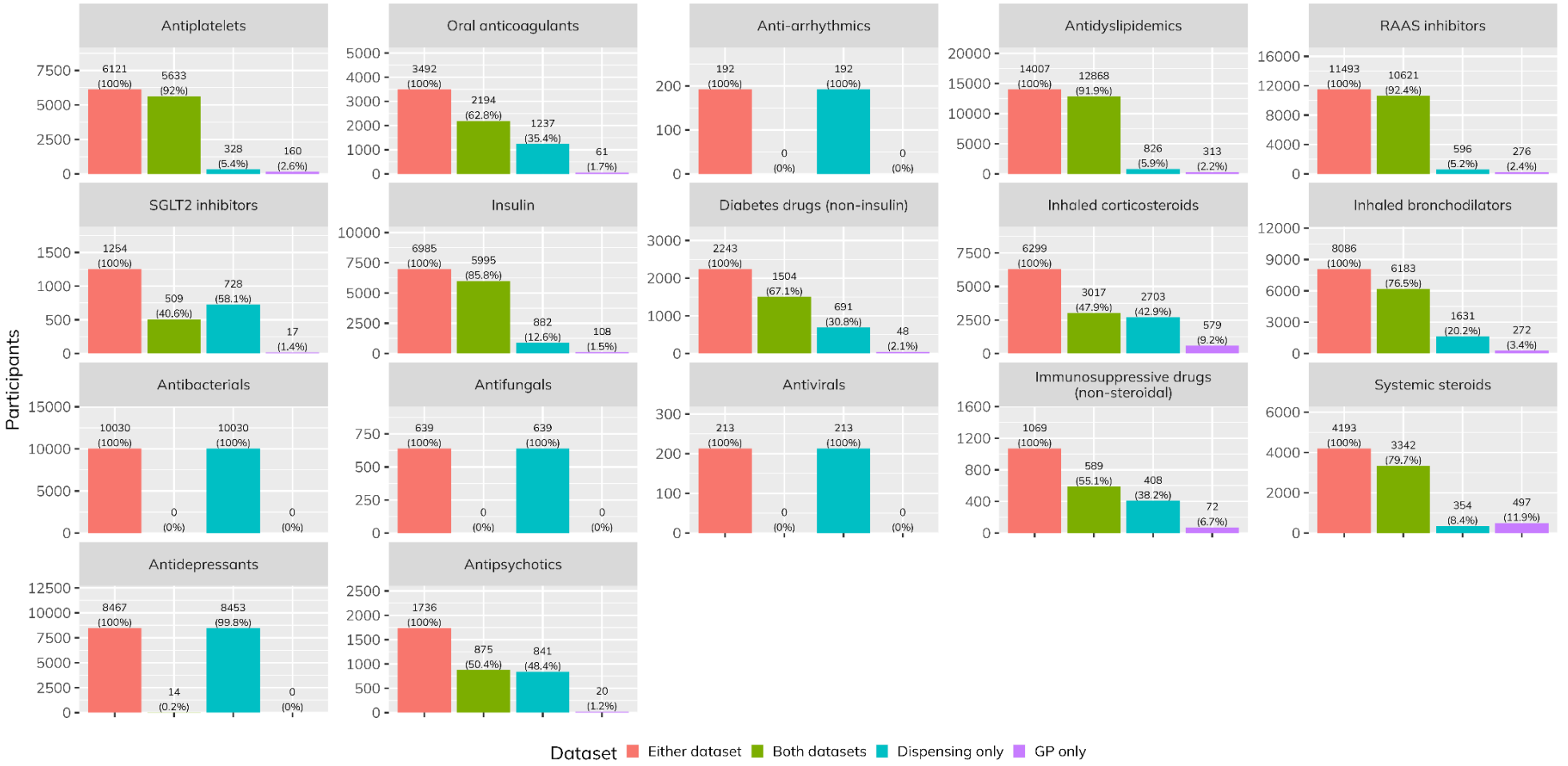


Figure legend: Areas are proportional to the number of participants included in each group. The proportions shown were calculated using the total number of participants identified in either data source.

Figure 3.11 - Aggregate estimates for different drug exposures at randomisation by routinely-collected data source



1 Table 3.8- Aggregate estimates for different drug exposures at randomisation by RCD source

	Participants in each source (%)					Agreement
Drug group	Either (%)[1]	Both (%)[2]	GP dataset only (%)[2]	Dispensing dataset only (%)[2]	Neither dataset (%)[2]	κ (95% CI)
Antiplatelets	6,121 (23.1%)	5,633 (21.2%)	160 (0.6%)	328 (1.2%)	20,395 (76.9%)	0.95 (0.94 : 0.95)
Oral anticoagulants	3,492 (13.2%)	2,194 (8.3%)	61 (0.23%)	1,237 (4.7%)	23,024 (86.8%)	0.75 (0.73 : 0.76)
Anti-arrhythmics	192 (0.7%)	0 (0%)	0 (0%)	192 (0.7%)	26,324 (99.3%)	0.00
Antidyslipidemics	14,007 (52.8%)	12,868 (48.5%)	313 (1.18%)	826 (3.1%)	12,509 (47.2%)	0.91 (0.91 : 0.92)
RAAS inhibitors	11,493 (43.3%)	10,621 (40.1%)	276 (1.04%)	596 (2.2%)	15,023 (56.7%)	0.93 (0.93 : 0.94)
SGLT2 inhibitors	1,254 (4.7%)	509 (1.9%)	17 (0.06%)	728 (2.7%)	25,262 (95.3%)	0.57 (0.53 : 0.60)
Insulin	6,985 (26.3%)	5,995 (22.6%)	108 (0.41%)	882 (3.3%)	19,531 (73.7%)	0.90 (0.89 : 0.91)
Diabetes drugs (non-insulin)	2,243 (8.5%)	1,504 (5.7%)	48 (0.18%)	691 (2.6%)	24,273 (91.5%)	0.79 (0.77 : 0.80)
Inhaled corticosteroids	6,299 (23.8%)	3,017 (11.4%)	579 (2.18%)	2,703 (10.2%)	20,217 (76.2%)	0.58 (0.56 : 0.59)
Inhaled bronchodilators	8,086 (30.5%)	6,183 (23.3%)	272 (1.03%)	1,631 (6.2%)	18,430 (69.5%)	0.82 (0.81 : 0.83)
Antibacterials	10,030 (37.8%)	0 (0%)	0 (0%)	10,030 (37.8%)	16,486 (62.2%)	0.00
Antifungals	639 (2.4%)	0 (0%)	0 (0%)	639 (2.4%)	25,877 (97.6%)	0.00
Antivirals	213 (0.8%)	0 (0%)	0 (0%)	213 (0.8%)	26,303 (99.2%)	0.00
Immunosuppressive drugs (non-steroidal)	1069 (4%)	589 (2.2%)	72 (0.27%)	408 (1.5%)	25,447 (96%)	0.70 (0.68 : 0.73)
Systemic steroids	4,193 (15.8%)	3,342 (12.6%)	497 (1.87%)	354 (1.3%)	22,323 (84.2%)	0.87 (0.86 : 0.88)
Antidepressants	8,467 (31.9%)	14 (0.1%)	0 (0%)	8,453 (31.9%)	18,049 (68.1%)	0.00 (-0.02 : 0.02)
Antipsychotics	1,736 (6.5%)	875 (3.3%)	20 (0.08%)	841 (3.2%)	24,780 (93.5%)	0.65 (0.63 : 0.68)

2 *Table legend: [1]Proportion calculated using all participants as denominator (separately from the subsequent columns); [2]Proportion calculated using all participants as*
3 *denominator (separately from the 'Either' column)*

Figure 3.12 - Aggregate estimates for drug initiation after randomisation, by drug group and RCD source



1 Table 3.9 - Agreement metrics for different drug initiations after randomisation between the GP and Dispensing datasets

2

	Participants in each source (%)					Agreement (yes/no)	Agreement (time after randomisation)
Drug group	Either dataset[1]	Both datasets[2]	GP dataset only[2]	Dispensing dataset only[2]	Neither dataset[2]	κ (95% CI)	ICC (95% CI)
Antiplatelets	403 (2.1%)	362 (1.9%)	23 (0.12%)	18 (0.1%)	19,139 (97.9%)	0.95 (0.93 : 0.96)	0.97 (0.96 : 0.97)
Oral anticoagulants	1,505 (7.7%)	934 (4.8%)	63 (0.32%)	508 (2.6%)	18,037 (92.3%)	0.75 (0.73 : 0.77)	0.94 (0.92 : 0.95)
Anti-arrhythmics	37 (0.2%)	0 (0%)	0 (0%)	37 (0.2%)	19,505 (99.8%)	0.00	N/A
Antidyslipidemics	528 (2.7%)	473 (2.4%)	17 (0.09%)	38 (0.2%)	19,014 (97.3%)	0.94 (0.93 : 0.96)	0.97 (0.96 : 0.98)
RAAS inhibitors	466 (2.4%)	415 (2.1%)	17 (0.09%)	34 (0.2%)	19,076 (97.6%)	0.94 (0.92 : 0.96)	0.98 (0.97 : 0.98)
SGLT2 inhibitors	82 (0.4%)	36 (0.2%)	0 (0%)	46 (0.2%)	19,460 (99.6%)	0.61 (0.50 : 0.72)	0.99 (0.98 : 1.00)
Insulin	397 (2%)	329 (1.7%)	9 (0.05%)	59 (0.3%)	19,145 (98%)	0.90 (0.88 : 0.93)	0.94 (0.93 : 0.95)
Diabetes drugs (non-insulin)	353 (1.8%)	183 (0.9%)	5 (0.03%)	165 (0.8%)	19,189 (98.2%)	0.68 (0.63 : 0.73)	0.94 (0.92 : 0.95)
Inhaled corticosteroids	516 (2.6%)	205 (1%)	33 (0.17%)	278 (1.4%)	19,026 (97.4%)	0.56 (0.51 : 0.61)	0.93 (0.91 : 0.95)
Inhaled bronchodilators	794 (4.1%)	488 (2.5%)	31 (0.16%)	275 (1.4%)	18,748 (95.9%)	0.75 (0.73 : 0.78)	0.97 (0.96 : 0.98)
Antibacterials	2,743 (14%)	0 (0%)	0 (0%)	2,743 (14%)	16,799 (86%)	0.00	N/A
Antifungals	408 (2.1%)	0 (0%)	0 (0%)	408 (2.1%)	19,134 (97.9%)	0.00	N/A
Antivirals	110 (0.6%)	0 (0%)	0 (0%)	110 (0.6%)	19,432 (99.4%)	0.00	N/A
Immunosuppressive drugs (non-steroidal)	72 (0.4%)	19 (0.1%)	27 (0.14%)	26 (0.1%)	19,470 (99.6%)	0.42 (0.26 : 0.57)	0.94 (0.78 : 0.98)
Systemic steroids	694 (3.6%)	562 (2.9%)	58 (0.3%)	74 (0.4%)	18,848 (96.4%)	0.89 (0.87 : 0.91)	0.98 (0.97 : 0.98)
Antidepressants	873 (4.5%)	0 (0%)	0 (0%)	873 (4.5%)	18,669 (95.5%)	0.00	N/A
Antipsychotics	550 (2.8%)	108 (0.6%)	9 (0.05%)	433 (2.2%)	18,992 (97.2%)	0.32 (0.26 : 0.38)	0.94 (0.92 : 0.96)

Table legend: [1]Proportion calculated using all participants as denominator (separately from the subsequent columns); [2]Proportion calculated using all participants as denominator (separately from the 'Either' column)

3.5 Discussion

This chapter describes the first reported use of two novel RCD sources on medication use in the community in England within a randomised trial. The two datasets provided extensive coverage of the study population, and captured information on medication use across 17 specific drug groups of interest. Both sources showed high levels of agreement with CRF records for antiplatelet and oral anticoagulant use at randomisation at the individual participant level, while generating similar aggregate drug exposure estimates when compared with CRF data. Overall, these results show that GP and Dispensing data are valuable sources for identifying baseline community drug exposure within a trial of in-hospital COVID-19 treatments, and support their potential use in future trials.

Capture of RECOVERY participants in the two RCD sources was very high, with 91% of the study population captured in both datasets and 98% in at least one. These figures confirm the high levels of coverage in the GP (93.5% of the population in England) and the Dispensing datasets (100% of reimbursed prescriptions in the community setting in England), as expected in a unified healthcare system such as the UK NHS. Participants not included in any of the linkage datasets were generally younger and healthier than those with no available linked data, possibly suggesting a true absence of relevant records in these datasets. Those included in the GP but not the Dispensing dataset were also younger and had little evidence of chronic conditions or prescriptions in primary care records, suggesting the absence of Dispensing data may reflect a true lack of reimbursed prescriptions in the community (either because these were not prescribed, or not collected).[251] Alternatively, this could be due to over-the-counter dispensing of medications (such as aspirin or low-dose statins) or secondary care dispensing. In some countries, the absence of dispensing records could also reflect coverage through alternative reimbursement schemes, use of samples, or out-of-pocket payments,[221,252] although this is

unlikely to be the case here. People included in the Dispensing but not the GP dataset had similar age and comorbidity burden as the overall cohort. Dispensing data for these participants included prescriptions for medications expected to be prescribed through primary care and also captured in the GP dataset extraction, suggesting these individuals might represent the minority of patients and GP practices not included in the GPES extraction. Overall, these findings confirm that both datasets are able to provide important information on medication use in the community in the large majority of people in England. However, GP data lacked any records for medication groups of important public health interest, such as antimicrobials, and had limited capture of mental health drugs. This likely results from the targeted (and evolving) nature of the GPES data extraction, and is an important limitation to broad use of this dataset for clinical research. Moreover, Dispensing data captured more distinct drug groups and more participants in each drug group than GP data. Thus, GP data (as available in GDPPR) should currently be considered less reliable than Dispensing data for drug exposure derivation, although the nature of this data collection may change in the future.

The timeliness analysis performed showed that the GP dataset had a two week, and the Dispensing dataset a two month delay between occurrence of a prescription/dispensing event and its inclusion in the disseminated data. These lags are shorter than those reported for other RCD sources in the UK and the US, and likely to be suitable for most clinical trial reporting schedules.[247] Nonetheless, they would prevent use of these data within fast analysis and reporting pipelines (such as those employed in RECOVERY at the early stages of the COVID-19 pandemic; i.e. weekly unblinded data reviews), and for real time safety monitoring.[51] Nonetheless, the limitations found may still allow use of these RCD sources in cases where fast production of results is not as paramount.

Taken together, the data quality assessments performed suggest that the Dispensing dataset is superior to the GP dataset, and should be the preferred source of information on community drug exposure. This is further confirmed by the lower levels of agreement found between GP data and CRF records when compared with Dispensing data, and the marginal improvement in agreement obtained by adding GP data to Dispensing data. However, this assessment was restricted to antiplatelets and oral anticoagulants, which were found to be captured to similar levels in both RCD sources. Moreover, the Dispensing dataset captures information on medications dispensed, rather than the issue of prescriptions or recording of usual medication lists (as in the GP dataset). While not a perfect measure of medication intake, dispensing is a closer surrogate than prescription. The improved agreement seen with Dispensing data might also result from different data collection processes: administrative and billing purposes in the Dispensing dataset (i.e. pharmacies are not reimbursed for prescriptions if not included in this data submission), versus recording of clinical encounters in the GP dataset. Of note, the value of claims-based data is highly-context specific and dependent on reimbursement rules. For example, novel oral anticoagulants are better represented in prescribing versus claims data in Finland, as electronic prescriptions are mandatory but these medications are not universally reimbursed.[253] Similar results have been described in other settings (e.g. contraceptives, benzodiazepines, and vaccinations in Denmark or the Netherlands).[254,255] Likewise, GP data identified more participants for medications included in the BNF chapter for vaccinations, and important numbers of additional participants for immunosuppressive and respiratory drugs. Thus, the value of each data source is not absolute, and depends on the specific research question. This further supports the idea that, when studying drug exposure, no source can be considered a true gold standard. Triangulation between different ascertainment sources is therefore likely to the best approach for optimal

estimates.[95,110,119–121] Overall, these findings highlight the importance of a thorough understanding of the purpose and underlying practice patterns involved in the generation of RCD sources to guide their appropriate use in research.[123]

Drug exposure can be identified from RCD sources using a multitude of analytical approaches.[256–258] Previous reports have found that the methodology used for this purpose is often not adequately reported, or that the impacts of choosing one method over the other are not appropriately discussed.[96,259] Due to the novelty of these two data sources, and the lack of guidance or previous experience to direct their implementation, a detailed, data-driven methodology development exercise was employed to guide their use here. This approach simultaneously harnessed the existence of a reference standard for drug exposure (the RECOVERY CRF), and of two collateral RCD sources (from which underlying medication use patterns could also be extrapolated). The 3-month lookback period selected to identify drug exposure at randomisation seemed to represent an inflection point, which maximised capture of participants who might have recently initiated a medication (and would be missed by using of shorter intervals), without overly extending the observation period and therefore including older records (representing discontinued medications). This is in keeping with the optimal agreement observed using this interval when compared with the CRF (although similar across all options considered), and with previous direct comparisons between drug exposure derivation methods.[258] Next, in the absence of exact Dispensing dates, the methodology development exercise performed showed very similar agreement results between Dispensing and CRF records across all possible date handling rules tested. This suggests that some flexibility may be allowed in the choice of a date imputing algorithm, should exact dates be required. However, the absence of a clear trend favouring any of the possible options also suggests that imputing of exact dates for trial analyses should be avoided, unless strictly necessary.

Using the selected algorithm, agreement between RCD and CRF for antiplatelets and oral anticoagulants was high (κ=0.69 and 0.76, percent agreement=93.7% and 95%, respectively), with RCD generating similar aggregate estimates of drug exposure at randomisation as the CRF (absolute differences <2%). Discrepancies between RCD and CRF could be due to participants being started on antiplatelets or anticoagulants upon admission (both of which were investigated as possible therapies in COVID-19), or having regular medications discontinued upon admission (as shown in studies of medication reconciliation errors upon hospital admission).[206,260,261] This was reflected in suboptimal sensitivity and PPVs (both in the range of 60-70%), although both specificity and NPVs were high (>94%). Consequently, in this context (i.e. a trial recruiting participants in-hospital), RCD may be considered a reasonable source of either aggregate estimates of in-hospital drug exposure at randomisation (e.g. to describe general characteristics of the study population, or risk modelling), or of individual estimates of drug exposure before hospital admission. However, it is unlikely to be suitable as a source of estimates of current treatment at the moment of randomisation on an individual basis (e.g. to confirm eligibility, safety assessments, or subgroup analyses), such as with the case illustrated here (i.e. assessing eligibility to the aspirin comparison).

Taken together, these findings suggest that RCD may be used to complement CRF records using a hybrid approach, with the CRF instantaneously capturing exposures of critical relevance in which small errors may be highly consequential (e.g. contraindicated drugs), and RCD used to gather information on multiple possible drug exposures of interest for non-time critical assessments and for analyses where small errors may be more acceptable. Incorporating RCD also offers the advantage of allowing retrospective, non-pre-specified analyses, without the need to create extensive CRFs upfront. This is particularly relevant in settings where streamlining may be paramount to successful study

delivery, as was the case in RECOVERY. Overall, these results suggest that RCD on medication use in the community in England can provide complete, accurate, and valuable information on drug exposure to either replace or complement traditional, manual data collection methods in randomised trials.

3.5.1 Limitations

Some limitations should be discussed. First, while the RECOVERY randomisation CRF was used as a reference standard to guide and validate the use of the two RCD sources, these data collection methods are not capturing the exact same medication exposure (medication use in the community from RCD, and medication use in-hospital at the time of randomisation from CRF), and thus some disagreement was to be expected. In this context, the use of best agreement with CRF as reference to guide implementation of RCD is a suboptimal approach, as it is possible that algorithms which provided best agreement could simply have artificially minimised discrepancies between the two sources (rather than improving identification of true drug exposure). Algorithm choice could also have been performed by replicating previously reported approaches using other datasets or in other clinical settings, but it is likely that specific patterns of healthcare use in this population (and therefore of RCD generation) could be different from those seen elsewhere. Therefore, in the absence of existing guidance or experience on how to implement these sources for the specific purpose considered, using CRF records as the closest reference is expected to be the best possible framework in which to anchor the empirical validation and development of the use of RCD in this context. Nonetheless, validation studies using RCD sources focused on in-hospital medication use are also needed. This will be the focus of the next chapter of this thesis, which will describe the validation of RCD on secondary care medication use within RECOVERY.

Second, both the methodology development exercise and interpretation of agreement found between CRF and RCD are limited by the small number of drugs recorded in the CRF which were suitable for comparison with the RCD sources used here (and restricted to two cardiovascular drugs used in similar settings). Therefore, no direct comparison with the study CRF was possible for most drug groups of interest. It is possible that comparisons involving other drugs could have led to the identification of alternative optimal drug exposure algorithms, or resulted in different levels of agreement between CRF and RCD. Further studies are therefore warranted to validate and guide the use of these RCD sources for other medication types. This will be addressed in part in chapter 6 of this thesis, where the Dispensing dataset will be compared with a larger number of cardiovascular drugs collected using CRFs in the ORION-4 trial. Nonetheless, the high levels of agreement found for most medication groups of interest between both RCD sources (both expected to be of reasonable quality) help to support their use in RECOVERY.

Third, given the specific nature of data collection in each RCD source, and the characteristics of each healthcare system, the observations made here may not be translatable to trials in other countries, as well as in other UK nations. Chapter 4 of this thesis will also complement the present chapter by focusing on the validation of RCD in Scotland within the LENS trial.

Finally, κ coefficients were used as the main statistic to assess agreement between RCD and CRF records, as the methodology proposed required the use of a single statistic to encompass the concept of "best agreement". In opposition to simple percent agreement, κ coefficients provide an estimate of the agreement observed beyond that expected by chance alone.[248] This is particularly relevant in scenarios where most individuals will not exposed to each drug, so that it is expected that agreement may be very high (as seen in most comparisons presented here), but by chance alone. However, κ coefficients have

known limitations (namely being influenced by properties other than agreement alone), are more difficult to interpret than percent agreement, and were not developed for the specific purpose of comparing two sources of data, so that their interpretation may be challenging.[262,263] To address this, comparisons between sources were complemented with other statistics, including absolute counts and percent agreement, as well as sensitivity, specificity, and PPV/NPV to investigate underlying reasons for disagreement. Moreover, as discussed above, the level of agreement between data sources required to allow reasonable use of RCD may depend on the specific scenario or intended use of the data (e.g. eligibility at an individual level vs characterisation of the overall cohort). Thus, the definition of acceptable agreement and conclusions on adequacy of RCD will necessarily be context specific.

3.6 Conclusion

Two novel RCD sources of medications in England provided excellent coverage of participants in the RECOVERY trial of COVID-19 medications. RCD yielded similar drug exposure estimates to those captured in the study CRF (when restricted to drugs expected to be commonly used in the community setting). RCD also contributed information on exposure to drug groups not captured in the CRF, with good agreement between both RCD sources. Dispensing data showed superior data quality features when compared with GP data, with the latter contributing small additional numbers of participants exposed to some drugs of interest. These results suggest that RCD on medications is a valuable and reliable source for assessment of drug exposure in the community for future trials in England. However, medication data collection using CRF records may still be required when highly accurate and time critical information is required.

CHAPTER 4: Routinely-collected nationwide inpatient medications data in England – validation and implementation within the RECOVERY trial of COVID-19 treatments

4.1 Chapter summary

Aims: To develop and validate the use of RCD on inpatient medications in England within the RECOVERY trial of treatments for hospitalised COVID-19 patients.

Methods: In RECOVERY, inpatient drug exposure is recorded in two CRFs completed by trial staff (randomisation and follow-up). In England, this is complemented via linkage to the Electronic Prescribing and Medicines Administration (EPMA) dataset. EPMA data coverage and quality were assessed by calculating counts and describing characteristics of RECOVERY participants who were included in the dataset, and verifying timeliness and data stability over time. Following preliminary data exploration, a data-driven methodology development exercise was performed by testing different possible algorithms to derive drug exposure, using randomisation CRF records as reference. Based on this exploration, agreement between EPMA and CRF data was calculated for baseline and follow-up drug exposure for multiple drugs. Agreement for binary drug exposure events (yes/no) was assessed using κ statistics and PABAK, along with sensitivity, specificity, and PPV/NPV. Agreement for quantitative drug exposure during follow-up was assessed using ICC and average differences for numerical estimates (i.e. number of days a drug was administered, as collected in the RECOVERY CRF for baricitinib, colchicine, corticosteroids, hydroxycloroquine, lopinavir-ritonavir, and remdesivir), and κ and PABAK for categorical estimates (i.e. proportion of days aspirin was administered during the

admission, categorised in the CRF as 0-50%, ≥50-90%, or ≥90%; or number of doses for tocilizumab/sarilumab, categorised as 1 or >1 doses).

Results: EPMA data captured a representative sample of 4,313 RECOVERY participants recruited in England (10.6%). Data had good stability and were provided in near real-time, but there was a two month lag between the end of analysis period and complete data reception (due to longer data submission intervals in some hospitals). EPMA data were provided in 6 different tables and 79 data fields, but most derivation rules tested yielded similar agreement with reference CRF data. Overall agreement between EPMA and the CRF was high (κ: 0.72, PABAK: 0.76, PPV: 75%, NPV: 95% for baseline exposure; κ: 0.77, PABAK: 0.85, PPV: 85%, NPV: 95% for follow-up exposure). Both sources yielded similar overall drug exposure prevalence estimates. Low sensitivity was the major contributor for discrepancies across most drugs (baseline: 74%-92%; follow-up: 19%-92%), except for corticosteroids (low specificity; baseline: 55%, follow-up: 68%), with important variations over time and by respiratory status at baseline. Agreement for number of days a drug was administered was moderate (overall ICC: 0.64), but with small overall absolute differences between sources (0.1 days). Similar results were found for proportion of days covered for aspirin (κ: 0.25; PABAK: 0.11) and number of doses for tocilizumab/sarilumab (κ: 0.24, PABAK: 0.37). EPMA data could be used feasibly for the different case-scenarios tested, but the limitations demonstrated hinder broader implementation within RECOVERY.

Conclusions: EPMA is a valuable source of inpatient drug data, with high overall agreement with trial CRF data. However, limited coverage and suboptimal sensitivity and specificity in some cases limit their use for drug data collection in trials. Nonetheless, these results favour using EPMA data as a valid epidemiological research tool.

4.2 Introduction

Information on inpatient medications is of vital importance in both acute care and chronic disease trials, to adequately characterise concomitant medication exposure and quantify study treatment adherence (if used in the inpatient setting). Contrary to outpatient settings, where prescription records may not be readily accessible, in the inpatient setting this information is usually easily available to research staff at site. However, manual data abstraction from medical records can be time-consuming and is prone to error or missing information, especially given the complexity of inpatient drug charts.[264]

The increased use of electronic prescription management systems has facilitated management, recording, and retrieving of inpatient prescriptions. However, and contrary to outpatient settings, access to RCD on inpatient medications for research purposes has been limited.[180,196,265] Methods to extrapolate and account for inpatient drug exposure have also been a source of discussion and variability in pharmacoepidemiology studies.[266,267]

Following expanded data release directions from the UK government as part of the COVID-19 response, in 2020 NHSE have made available a new dataset containing secondary care medicines collected from Electronic Prescribing and Medicines Administration (EPMA) systems in 26 hospital trusts in England.[158] These data offer immense promise for clinical research. Nonetheless, and similarly to other RCD sources, their widespread use requires development of dataset-specific algorithms, appropriate validation, and feasibility assessments.

The aims of this chapter are to: 1) assess whether EPMA data are of adequate quality to derive inpatient drug exposure in RCTs; 2) develop the methodology needed to implement these data for that purpose; and 3) test the implementation of these data in possible use-case scenarios within a randomised trial.

4.3 Methods

4.3.1 Study population, design, and setting

As described in more detail in Chapter 3, RECOVERY is a large, streamlined, platform, open-label, randomised controlled trial of treatments for patients hospitalised with COVID-19.[51] Data collection by trial staff is minimal, and restricted to essential information. This is complemented by linkage to RCD sources, including a subset of secondary care prescriptions in England from the EPMA dataset.[158]

The study cohort for this analysis included RECOVERY trial participants aged 16+ years at the time of randomisation, recruited in England up until the 31st May 2022 (inclusive), who had not withdrawn study consent, and who had linked EPMA data available during the admission during which they were randomised. All RECOVERY participants have provided written informed consent, including for linkage to RCD sources in the UK.

4.3.2 Data collection

4.3.2.1 RECOVERY case report form data

For this analysis, data from two distinct RECOVERY CRFs were used: 1) the randomisation form, completed just prior to randomisation, which gathers information on exposure to specific drugs at baseline (i.e. being used from hospital admission up to the moment of randomisation), and 2) the follow-up form, scheduled for completion upon death, discharge, or 28 days after randomisation (whichever occurred earlier), which captures exposure to specific drugs during this period. Available CRF data for this analysis included 41,214 RECOVERY participants aged 16+ years and recruited in England up until the 12th December 2022. Drugs included in this analysis were restricted to those evaluated in treatment comparisons completed by November 2022, or others drugs of interest to RECOVERY but which were not included in randomised comparisons (Table 4.1).

CRFs were completed by RECOVERY site staff (mostly research nurses) by reviewing available documentation locally (which would include electronic health records). Drug exposure was collected as a dichotomous measure (yes or no) in the randomisation and follow-up forms, and also as a quantitative measure for some drugs during follow-up (Table 4.1-4.2). For low-molecular weight heparins (LMWH), participants recorded as having standard and higher dose LMWH prescribed were combined. Quantitative exposure was collected for most drugs as number of days a drug was given during admission (up to a maximum of 10 days total, in line with the study protocol). The exceptions were tocilizumab/sarilumab (to be given as one or two doses, and captured accordingly in the CRF) and aspirin (to be given up until discharge, and recorded in the CRF as proportion of days drug was given during admission). Participants with a question left unanswered were excluded from comparisons for that particular drug within this analysis. An internal review of drug recording in the follow-up CRF (performed at 5 RECOVERY sites for 449 participants taking part in the tocilizumab comparison) found that the CRF had high accuracy when compared with source documentation in each hospital record (89% sensitivity, 97% specificity, 94% PPV and 94% NPV across all drugs recorded).

1 Table 4.1 - RECOVERY case report form drug exposure collection (for drugs used in this
2 analysis)

Event	Drug	Question
Baseline	Remdesivir	*Is the patient currently prescribed remdesivir?*
	Corticosteroids	*Is the patient currently prescribed systemic corticosteroids (dexamethasone, prednisolone, hydrocortisone, methylprednisolone)? (please do not include topical or inhaled treatments)*
	Baricitinib	*Is the patient currently prescribed baricitinib (or other JAK inhibitor)?*
	Tocilizumab/sarilumab	*Has the patient received tocilizumab or sailumab therapy during this admission?*
	Macrolides	*(Does the patient have any current comorbidities or other medical problems?) - current treatment with macrolide antibiotics which are to continue*
	Antiplatelets	*(Does the patient have any current comorbidities or other medical problems?) - antiplatelet therapy*
	Oral anticoagulants	*Is the patient currently on warfarin or a direct oral anticoagulant?*
	LMWH	*What venous thromboembolism prophylaxis is the patient receiving? (none, standard low molecular weight heparin, higher dose low molecular weight heparin)*
	Oxygen	*Does the patient require oxygen?*
Follow-up *("which of the following treatment(s) did the patient definitely receive as part of their hospital admission after randomisation? NB include RECOVERY study-allocated drug, only if given, PLUS any of the other treatments if received as standard hospital care"*)	Corticosteroids	Corticosteroid (dexamethasone, prednisolone, hydrocortisone or methylprednisolone)
	Tocilizumab/sarilumab	Tocilizumab or sarilumab
	Macrolides	Azithromycin or other macrolide (eg, clarithromycin, erythromycin)
	Remdesivir	Remdesivir
	Hydroxychloroquine	Hydroxychloroquine
	Lopinavir-ritonavir	Lopinavir-ritonavir
	Colchicine	Colchicine
	Baricitinib	Baricitinib
	Aspirin	Aspirin
	REGN antibodies	Synthetic monoclonal antibodies (REGN10933+REGN10987)
Follow-up	Tocilizumab/sarilumab	*Please select number of doses of tocilizumab or sarilumab the patient received (1 or >1)*
	Aspirin	*Please select the proportion of days the patient received aspirin during the first 28 days after randomisation (or from randomisation to date of discharge if this is sooner):* - *Most days (≥90%)* - *Some days (≥50% <90%)* - *Few days (<50% of days, but not zero)* - *None*
	All other drugs	*Please select the number of days the patient received this drug (1 to 10)*

3

Table 4.2 - Drugs and exposure metrics selected for analysis

Drug	Baseline exposure	Post-randomisation exposure	
	Binary		Quantitative
Systemic corticosteroids (dexamethasone, prednisolone, hydrocortisone or methylprednisolone)	X	X	Number of days
Tocilizumab or sarilumab	X	X	Number of doses
Azithromycin or other macrolide (azithromycin, clarithromycin, erythromycin)	X	X	Number of days
Remdesivir	X	X	Number of days
Oxygen	X		
Oral anticoagulants	X		
LMWH	X		
Antiplatelets	X		
Oxygen	X		
Hydroxychloroquine		X	Number of days
Lopinavir-ritonavir		X	Number of days
Colchicine		X	Number of days
Baricitinib		X	Number of days
Aspirin		X	Proportion of days (0-50%; ≥50-90%; ≥90%)
REGN antibodies (casivirimab, imdevimab)		X	

The randomisation CRF was used to collect general demographic and clinical characteristics (age, sex, race, level of respiratory support, days since hospitalisation, comorbidities, recruiting location) and admission date. The follow-up CRF was used to collect date of death or discharge, or the date when the form was completed (if none of the above occurred within 28 days after randomisation), with the earliest of these used to define the end of the follow-up analysis period (i.e. right-censoring). Algorithms to integrate CRF and linkage data to define date of admission, discharge, and death were defined by the RECOVERY study team (blind to treatment allocations or outcomes), and the same rules applied here.[148,241]

4.3.2.2 EPMA dataset

EPMA data contain secondary care prescriptions collated from 26 NHSE Trusts using electronic management systems provided by CareFlow Medicines Management.[158] This includes two separate systems (Cerner and WellSky), encompassing approximately 15% of total prescribing activity in England. Dissemination of this dataset started in December 2020, but it includes historic data starting from January 2019, and is only available for COVID-19 research. The data include inpatient but also outpatient prescribing and administration (if occurring in secondary care), and discharge prescriptions. The available documentation acknowledges that some specialties (such as intensive care, oncology, and paediatrics) or medications (such as warfarin or intravenous infusions) may not be adequately captured as some Trusts may use other systems for these purposes. Moreover, different Trusts provide data to the central EPMA system at different time points (either daily, weekly, or monthly). Data recorded each day is then extracted by NHSE from the EPMA system in the early morning of the following day (for Trusts that provide data on a daily basis). Retrospective updates to the EPMA data are never included in the disseminated data. Data retrieved by NHSE are then made available to third parties immediately. Therefore, the latest data provided may vary across different Trusts based on their specific data extraction cycle.[158]

The raw EPMA data is provided as a set of 6 tables, split into prescription and administration events: *prescription*, *prescription*_dmd (dm+d – Dictionary of Medicine and Devices, the standard NHS controlled terminology for drugs), and *prescription_dosage*; *administration*, *administration_dmd*, and *administration_dosage* (Table 4.3). In total, 79 data fields were provided (Table 4.4).

1 Table 4.3 - Number of participants and records included in each EPMA data table

Data table	Number of distinct participants	Number of distinct records
Prescription	5125	408,574
Prescription dmd	4716	337,090
Prescription dose	5125	409,998
Administration	4943	2,372,641
Administration dmd	4496	2,056,430
Administration dose	4943	2,372,959

2

3 Table 4.4 - Overview of EPMA data tables

Table	*Prescription*	*Prescription dmd*	*Prescription dose*	*Administration*	*Administration dmd*	*Administration dose*
Demographics	X			X		
Location	X			X		
Medication	Name	- dmd+d/ SNOMED code - dmd Mapping source		Name	-dmd+d/ SNOMED code - dmd Mapping source	
	Dose (free text)		Dose (coded): - weight/volume - number of tablets/equivalent - "as needed" flag			"As needed" flag
	Form (tablets, injections, etc)					
Prescription	Type (inpatient, discharge, outpatient, short-term leave)					
	Status (active, inactive, unused)					
Temporal data	Prescription authorised Data extraction		Prescription start Prescription end	Scheduled Administered Recorded Last updated		

4

Prescription and *administration* tables provide the exact medication name entered by the staff who generated each record (using character strings), while *prescription_dmd* and *administration_dmd* include SNOMED medication codes (from those included in the dmd terminology).[268,269] Following initial data exploration assessments, a decision was made to identify relevant records for each drug of interest from the medication name provided in the *prescription* or *administration* tables, as many participants had incorrect or missing dmd/SNOMED codes. These included medical devices or interventions, but also drugs. Search terms were identified after manual review of the BNF classification[270] to select all relevant drugs, and identification of all medication names available in the UK for that particular drug group (including brand names) at https://www.medicines.org.uk/emc/. Conversely, all distinct medication names provided in the raw EPMA data (~13,000) were also manually reviewed. This exercise showed that brand names are generally not used for prescription (except in combination with generic name). The exception (for drugs of interest in this analysis) were the REGN antibodies, which could be prescribed as either "REGN10933 & REGN10987", "Ronapreve", or Casirivimab and Imdevimab". Only the *prescription* and *administration* tables were used to extract drug exposure. The data fields selected for use in each table are detailed in Table 4.5.[123] Prescription status and type were extracted from the relevant coded fields. Relevant prescription formulations were identified by manual review and categorisation of all distinct free-text formulations provided (~600) into systemic, non-systemic (i.e. topical), inhaled, and any other formulation.

1 Table 4.5 - EPMA data tables and fields selected for analysis

Data field		Description and use	Completeness rate (%) – proportion of records with available information for each field
Prescription table	*Administration* table		
PRESCRIPTION_ID_HASHED	PRESCRIPTION_ID_HASHED	Unique prescription identifier; link prescription and administration records	100
	ADMINISTRATION_ID_HASHED	Unique administration identifier	100
MEDICATION_NAME	MEDICATION_NAME	Identify drug records	100
TYPE		Prescription type (inpatient/ outpatient/ discharge/ short-term leave)	100
STATUS		Prescription status (active/ inactive/ unused); active - this is the current dose for the Order and is currently being given to the patient; inactive - this is a previous dose for the Order OR the patient has stopped taking the medicine; unused - this was prescribed for the patient but the dose was never administered i.e. it was changed before a dose was given	100
FORM		Prescription formulation (free text)	95.4
INITIAL_AUTHORISED_DATE_TIME		Date and time prescription was authorised	100
	SCHEDULED_DATE_TIME	Date and time drug was scheduled	100
	ADMINISTERED_DATE_TIME	Date and time drug was administered	77.3
	RECORDED_DATE_TIME	Date and time drug administration was recorded	100
	LAST_UPDATED_DATE_TIME	Date and time administration record was last updated	100

2

EPMA data used in this analysis were received on the 30th September 2022, and included 5,116 participants (5,091 in the *prescription* and 4,911 in the *administration* tables). Linkage was based on NHS number, with subsequent checks for consistency in date-of-birth and sex performed between the CRF data and the received linkage data. A previous extract (June 2022) was used for data timeliness and stability assessments. Subsequent analyses were further restricted to participants with available EPMA data during the hospital admission during which they were randomised (i.e. if they had at least one record on or before date of admission and on or before randomisation date, and one record on or after randomisation and preceding the censoring date).

Baseline exposure was defined as a relevant EPMA record occurring on or after the admission date (i.e. the earliest of admission date provided in the CRF, admission date for the HES episode straddling randomisation date, or the randomisation date), and on or before randomisation date. A sensitivity analysis was also performed by excluding records occurring on the date of randomisation. Post-randomisation exposure was defined as a relevant EPMA record occurring after the randomisation date and before the right-censoring date (as defined above).

For quantitative exposure measures, the total number of days a drug was received (or doses for tocilizumab/sarilumab) was calculated as the number of distinct days with at least one relevant record within a period starting from the day after randomisation to the day of completion of the follow-up CRF form (inclusive). A value of 10 days was assigned if the calculated number was >10 (for alignment with the RECOVERY CRF data collection structure, as described in Table 4.1). For aspirin, the number of days the drug was received was transformed into proportion of days during the admission period (also to align with the structure of the RECOVERY follow-up CRF for that particular drug).

4.3.3 Data analysis

First, an assessment was made as to whether EPMA data are appropriate for the specific aim of identifying inpatient drug exposure in RECOVERY trial participants. As described in Chapter 3, this follows recommendations from previous reports on clinical data quality[33,192,244] and relevant guidance from the UK MHRA and the US FDA.[24,36] Three data quality features were assessed: 1) completeness/coverage (whether the intended population and drug exposure are captured adequately, and whether there are any temporal or geographical gaps in the data); 2) timeliness and stability (how soon a data item is available for analysis after the respective event occurred, and whether reported data may change retrospectively); and 3) accuracy/validity/comparability (how well RCD agree with a reference standard for the exposure being assessed).[123,192,245–247]

Similarly to Chapter 3, data quality was considered in a contextual/external (i.e. whether the data is of sufficient quality for a particular intended purpose) vs internal dimension (whether the recorded data accurately represents the reality it aims to capture),[192,242] as it is assumed publicly-available datasets published by NHSE have undergone internal data quality checks. However, contrary to other data products published by NHSE (such as the Medicines dataset),[230,239] there is no publicly-available information on data quality assessments for the EPMA dataset.

To guide the implementation of EPMA data, potential rules to derive drug exposure were identified and assessed against the CRF as the *de facto* reference standard.[100] Finally, implementation of EPMA data was tested using possible use-case scenarios of direct interest to the RECOVERY trial.

4.3.3.1 Data quality – completeness/coverage

Data completeness was assessed by first calculating the number of RECOVERY participants recruited in England with available EPMA data spanning the admission during

which they were randomised (i.e. the study cohort, as described above). General demographic and clinical characteristics were also computed for these participants versus those with no available data. Recruitment location was plotted in both groups to look for potential geographical gaps (by calculating the proportion of participants in each RECOVERY site in England who had available EPMA data, and plotting this over the map of England), and recruitment timing to check for continuous and consistent data capture over time (by calculating counts of participants with vs without available EPMA data over time, based on date of randomisation). Then, drug capture was evaluated at both broad and specific levels by calculating drug exposure estimates for each BNF chapter (using publicly-available BNF to dm+d mappings to identify relevant dm+d codes)[271] and for drugs of interest to RECOVERY (Table 4.2).

4.3.3.2 Data quality – stability and timeliness

Data stability was assessed by calculating counts of distinct participants, data items/rows/entries, and drugs recorded within a common time period included in two consecutive data extracts, restricted to participants who could have been included in the earliest extract based on their randomisation date. Data timeliness was evaluated by assessing the interval between the last data item recorded, the reporting date (i.e. the date in which data were retrieved by NHSE), and the timing of data reception. Separately, the lag between the most recent time point where data were found to be stable (therefore suitable for inclusion in study analyses) and data reception was also assessed.

4.3.3.3 Data methodology development

An algorithm was initially developed to extract drug exposure from EPMA data based on published user guidance[158] and making use of potentially relevant data fields (i.e. considering relevant prescription formulations, type, and status, and restricting to administration records with available administration dates). However, manual

investigation of discrepant records between CRF data and EPMA records extracted using this preliminary algorithm showed that many discrepancies were due to unexpected issues with the data provided (e.g. missing administration dates, prescription records without administration records, missing prescription statuses and types). Therefore a data-driven method was implemented, in which the resulting agreement between CRF and EPMA data was used in a series of sensitivity analyses to define the best performing algorithm for this particular setting. The following derivation options were tested:

1. Different temporal data fields in the a*dministration* table: "administered" vs "scheduled" vs "recorded" vs "last updated" vs any (if missing data in any of these fields, and extracted in hierarchical fashion in the order mentioned). Of note, no further definition of the expected usage or interpretation of these data fields is provided by NHSE
2. Data tables: *prescription* vs *administration* vs any
3. Prescription type: inpatient vs any
4. Prescription status: active vs any
5. Prescription formulations: systemic vs systemic or unknown vs any (for all drugs except oxygen), or inhaled vs inhaled or unknown vs any (for oxygen)
6. Handling records on the day of randomisation (for baseline assessments): include vs exclude

4.3.3.4 Agreement between EPMA and CRF data

Based on the sensitivity analyses performed, a final set of rules was chosen to extract drug exposure from EPMA data. The resulting algorithm and then used to calculate agreement between EPMA and CRF data for baseline exposure and exposure during the follow-up period (i.e. accuracy of EPMA data vs CRF standard).

4.3.3.5 Use-case scenarios

Finally, two use-case scenarios were used to test the feasibility of implementing these data in potential RCT settings, with the aim of fostering future adoption:[33]

1. Study monitoring/auditing: variations in agreement for follow-up drug exposure between EPMA and CRF data across study sites (as a method to identify potential data recording issues);
2. Characterisation of standard-of-care patterns: changes in in-hospital medication use patterns over time identified using either EPMA or CRF data for select drugs (corticosteroids, hydroxychloroquine, tocilizumab/sarilumab, and bariticitinib).

4.3.3.6 Statistical analyses

To assess agreement between sources, the number of participants considered exposed to a particular drug according to each data source (expressed as participant-drug pairs) was calculated initially, separately for baseline and follow-up exposure. Since data collection from the CRF evolved over time (as treatment comparisons were completed or started), each specific drug agreement calculation was restricted to participants for whom the relevant question was answered in the CRF form.

Agreement for categorical drug exposure measures was first assessed using Cohen's κ, a commonly used agreement metric which facilitates comparison with other studies.[248] As κ may underestimate actual agreement when event prevalence is very high (e.g. corticosteroids in RECOVERY participants), PABAK values[263] were also calculated, similarly to other RCD validation studies.[272–274]

For each drug comparison, sensitivity, specificity, PPV, and NPV were also calculated, using CRF as standard, in order to complement agreement interpretation and investigate underlying reasons for disagreement. Sensitivity and specificity were used to assess performance characteristics of EPMA data (i.e. probability of a present or absent CRF

record being correctly identified in EPMA), and PPV/NPV as measures of accuracy if EPMA data were to be used (i.e. probability that a positive or negative EPMA record is true).[219,246,247,275] As PPV and NPV are influenced by the underlying event rate,[276] sensitivity and specificity were used as the main parameters to investigate variations in agreement across different drugs.

Agreement estimates are presented combined (across all drugs), and separately by drug. Agreement was then stratified by respiratory status at baseline to assess the influence of clinical status (and likely care setting), using 3 groups: 1) no oxygen/ventilation; 2) oxygen requirement or non-invasive ventilation (NIV); and 3) invasive mechanical ventilation (IMV) or extracorporeal membrane oxygenation (ECMO), in line with the pre-specified subgroup analyses performed in the RECOVERY trial. To further explore some of the observed agreement patterns by drug, sensitivity and specificity were also plotted over time using 3 month periods. Finally, to assess the impact of using EPMA vs CRF data in specific analyses, numbers of participants that would be considered exposed at baseline or during follow-up to each drug were calculated if only one source had been used, or either of them.

For discrete quantitative measures of drug exposure during follow-up (i.e. number of days exposed), ICCs were calculated using two-way mixed effects models, assessing agreement between sources (vs measurement consistency), and using single measurements as unit of analysis (vs average of several measurements).[135,214,249,250,277] Average differences were also calculated for number of exposure days between EPMA and CRF data as the numbers of days from CRF data minus the number of days from EPMA data. ICC and average differences (in days) were restricted to participants with calculated exposure >0 for both EPMA and CRF. For categorical drug exposure measurements (aspirin - proportion of days exposed; tocilizumab/sarilumab – number of doses received), κ and PABAK were also calculated (as described above).

In the use-case scenario of CRF drug reporting monitoring using EPMA data, κ and the proportions of missing records (EPMA events not captured by the CRF) and unconfirmed records (CRF events not captured by EPMA) were calculated for baseline drug exposure across all drugs. This exploratory assessment was restricted to sites with >100 participants to prevent the occurrence of extreme κ values if the number of included events was very low. For the use-case scenario of characterising treatment patterns over time, follow-up CRF data was used (rather than randomisation form data) as this assessment covers more drugs of interest and could track drug exposure both during and after that drug was being studied in RECOVERY.

Continuous parameters are presented as mean with SD, or median with IQR, as appropriate (with distribution normality assessed visually), and frequency counts and percentage distribution for categorical parameters. κ, PABAK, and ICC values are presented along with 95% CI, as well as proportions of missing/unconfirmed records and drug exposure over time (in the use-case scenarios). Average differences (for quantitative exposures) are presented alongside 95% limits of agreement (LOA).

4.3.3.7 Data and code

R v4.2.1 was used for data transformation, statistical analyses, and plotting. The source code used in this work can be accessed at https://github.com/gpessoaamorim/dphil-recovery-inpatient-data. Access to the RECOVERY trial data can be requested via the Infectious Diseases Data Observatory (https://www.iddo.org/covid19/data-sharing/accessing-data). Linkage to the RCD sources used here can be requested for other research purposes from NHSE at https://digital.nhs.uk/services/data-access-request-service-dars.

4.3.4 Ethics

The RECOVERY trial is sponsored by the University of Oxford, and has been approved by the UK Medicines and Healthcare products Regulatory Agency and the Cambridge East Research Ethics Committee (reference 20/EE/0101). All RECOVERY participants provided written informed consent to take part in the trial, including for linkage to the RCD datasets used here; participants who withdrew consent for any reason were excluded from this analysis. Permission to use the data for this study was given by the Chief Investigators and the trial Steering Committee.

4.4 Results

4.4.1 Data quality assessments

4.4.1.1 Study cohort

Of 40,860 RECOVERY participants aged 16+ years recruited in England up until the 31st May 2022 (inclusive), 4,313 (10.6%) had available EPMA data spanning the admission in which they were randomised, and were selected as the study population. Table 4.6 shows baseline characteristics of the study population versus participants recruited in England during the same period but with no EPMA data. Participants included in the study population had similar age, sex, comorbidity burden, and baseline disease severity to those without EPMA data, but were more frequently of White ethnicity (90% for participants with available data vs 80% not available).

1 Table 4.6 - Baseline cohort characteristics for participants with vs without EPMA data

Characteristic	EPMA data availability	
	Available, N = 4,313[1]	Not available, N = 36,547[1]
Age, years	63 (15)	63 (15)
Respiratory support at randomisation		
No oxygen or ventilation	545 (13%)	4,301 (12%)
Oxygen requirement or non-invasive ventilation	3,533 (82%)	29,806 (82%)
Mechanical ventilation or ECMO	235 (5.4%)	2,440 (6.7%)
Female	1,652 (38%)	13,305 (36%)
Race		
White	3,865 (90%)	29,267 (80%)
Asian	150 (3.5%)	3,786 (10%)
Black	87 (2.0%)	1,414 (3.9%)
Other	87 (2.0%)	842 (2.3%)
Mixed	47 (1.1%)	461 (1.3%)
Unknown	77 (1.8%)	775 (2.1%)
Days since hospitalisation	2.0 (1.0, 3.0)	2.0 (1.0, 3.0)
Baseline diabetes	1,130 (26%)	9,627 (26%)
Baseline heart disease	951 (22%)	8,521 (23%)
Baseline severe liver disease	51 (1.2%)	437 (1.2%)
Baseline severe kidney disease	186 (4.3%)	2,150 (5.9%)
Baseline lung disease	1,034 (24%)	8,197 (22%)
Baseline renal replacement therapy	56 (1.3%)	600 (1.6%)

Table legend: [1]Mean (SD); n (%); Median (IQR)

2

3 Recruitment patterns over time were similar in both groups (Figure 4.1). Participants with

4 available EPMA data were recruited from 26 RECOVERY sites spread across England

5 (19% of total) (Figure 4.2).

Figure 4.1 - Recruitment timeline for monthly counts of RECOVERY participants with vs without available EPMA data

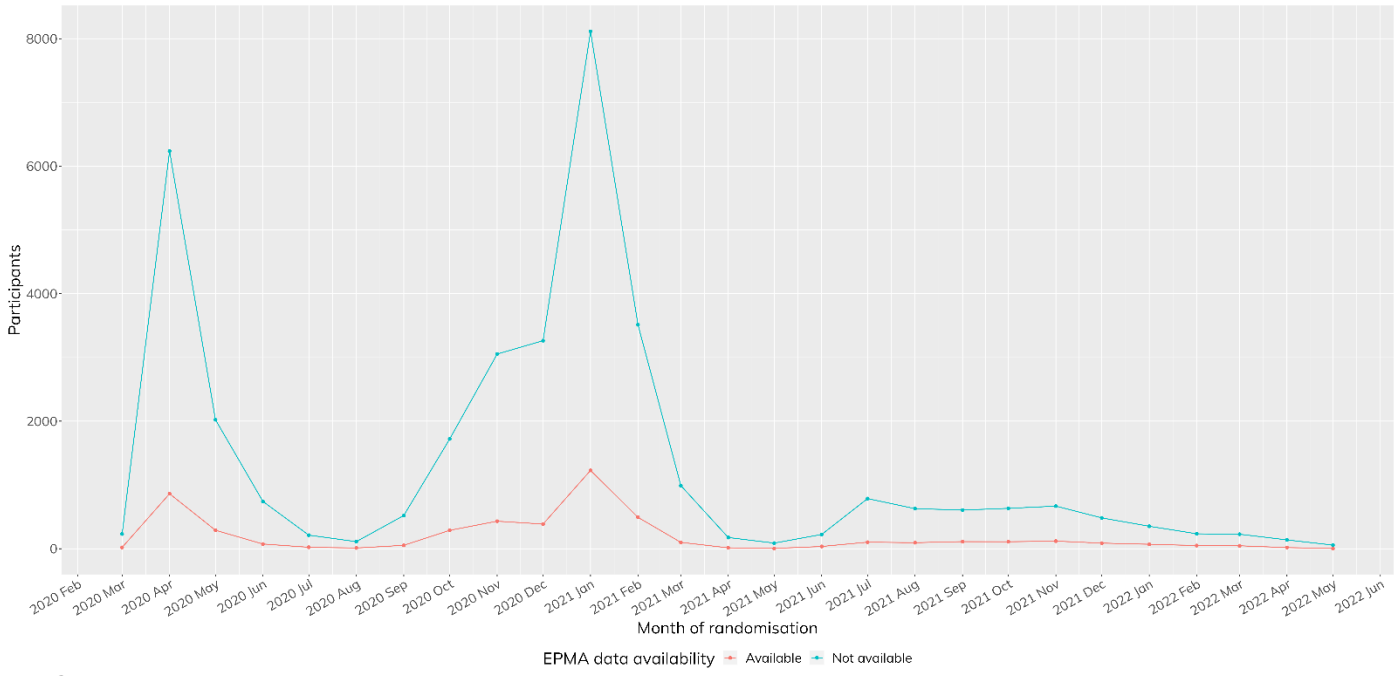


Figure legend: Analysis restricted to RECOVERY participants aged 16+ years and recruited in England up until the 31st of March 2022 (inclusive).

Figure 4.2 - Proportion of participants with available EPMA data within each RECOVERY site

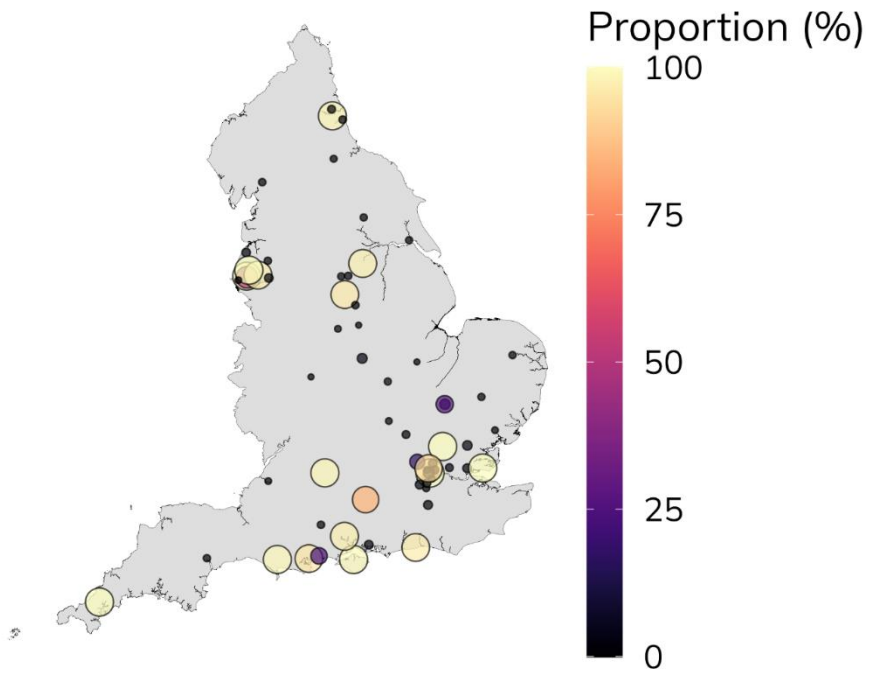


Figure legend: Circle sizes (together with colour) represent the proportion of participants in each site. Only sites with at least one participant with EPMA data available are shown

4.4.1.2 Drug coverage

EPMA data included drug records spanning all BNF chapters, with *prescription/administration* tables capturing similar numbers of participants (Figure 4.3). Of note, 99% of RECOVERY participants had at least one drug record for which no BNF chapter could be assigned. Manual review highlighted the presence of records not related to drugs or appliances (for example medication review alerts), but also medications with missing or incorrect dm+d/SNOMED codes. All specific drugs of interest to RECOVERY were captured in EPMA data, except for convalescent plasma, with lopinavir-ritonavir records identified in very small numbers (Figure 4.4).

1 Figure 4.3 - Drug capture in EPMA data across BNF chapters

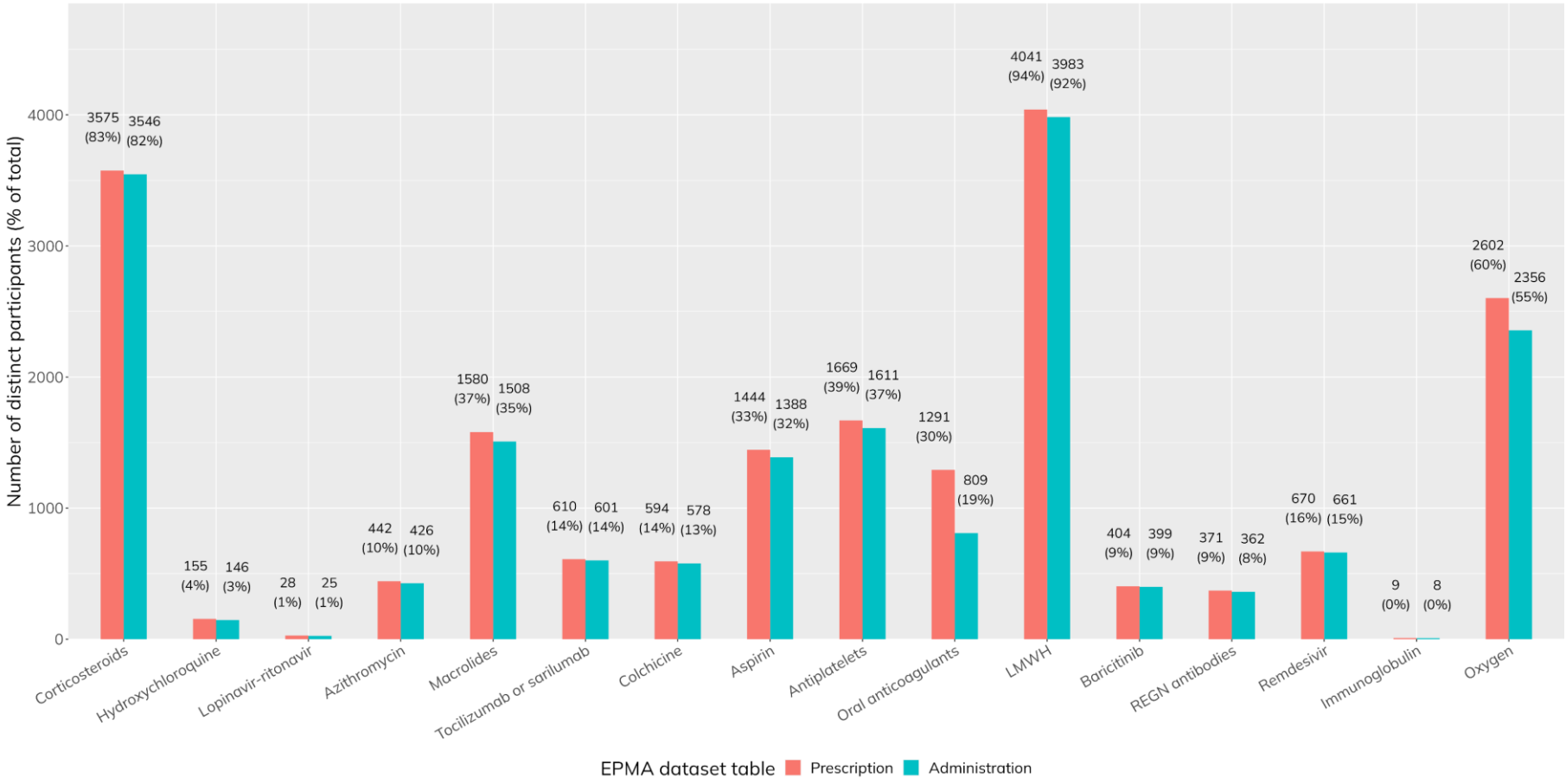
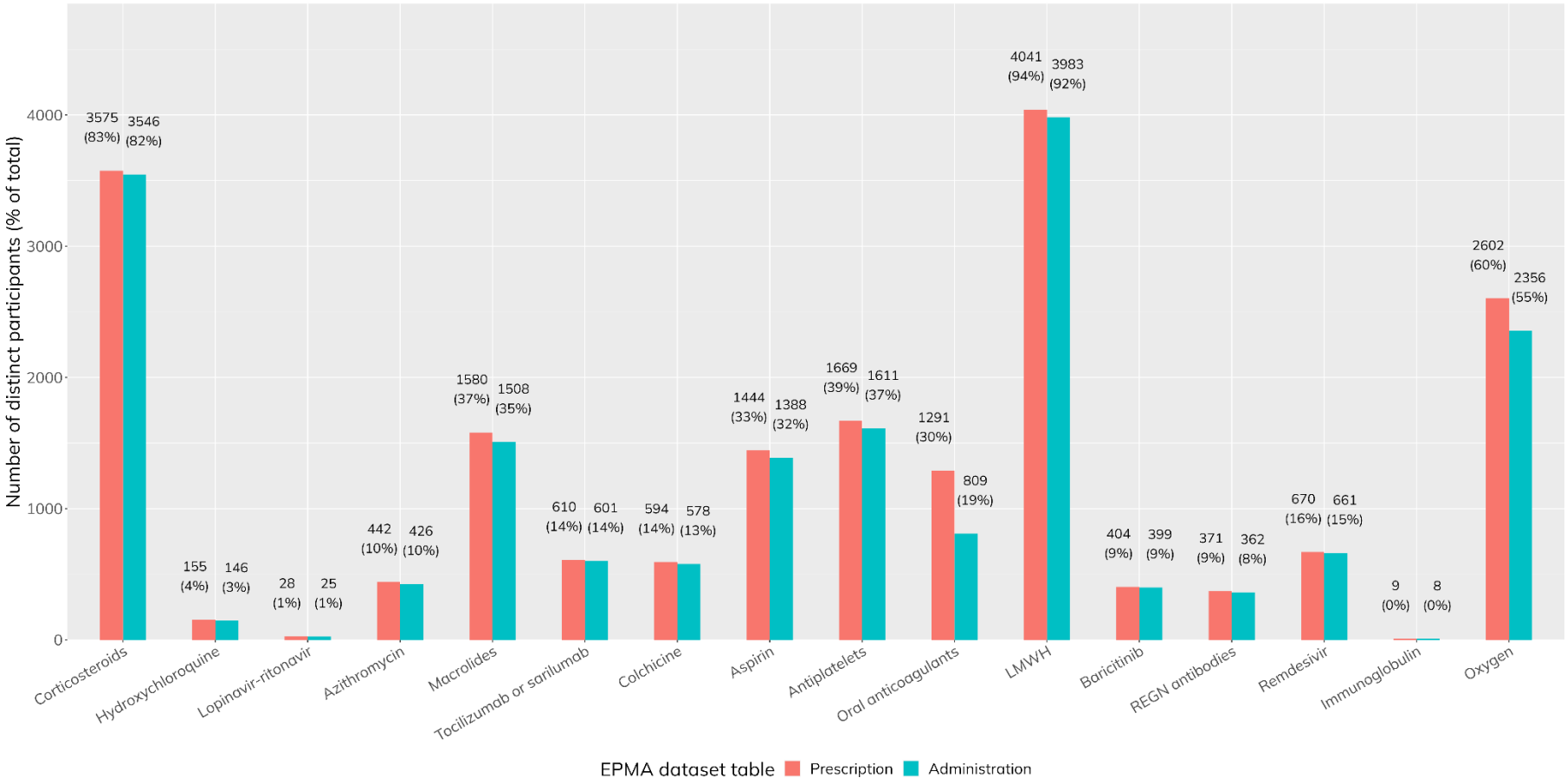


Figure 4.4 - Drug capture in EPMA data for specific drugs



4.4.1.3 Timeliness and stability

The raw EPMA data received included events spanning from early 2019 until after the date of data reception (30th September 2022), with no obvious data recording or submission lags (Figure 4.5). The most recent report date (the date in which the dataset was extracted from the EPMA data reporting system) was the 29th September 2022. Data were mostly stable across two consecutive quarterly extracts, over the entire period considered, in both *prescription* and *administration* tables (Figure 4.6). However, in the two months preceding reception of the earliest extract, using the most recent extract would increase the amount of data available (i.e. the number of distinct participants, records, and drug codes recorded in April 2022 and June 2022 increases from the June 2022 to the September 2022 extract, representing retrospective updates). Therefore a two month lag should be allowed between the end of the analysis period and the point of data reception. Hence, our study population was restricted to participants recruited before June 2022 (but using the September 2022 extract).

Figure 4.5 - Time series of weekly record counts for each date field (record date and reported date), split by Administration and Prescription tables

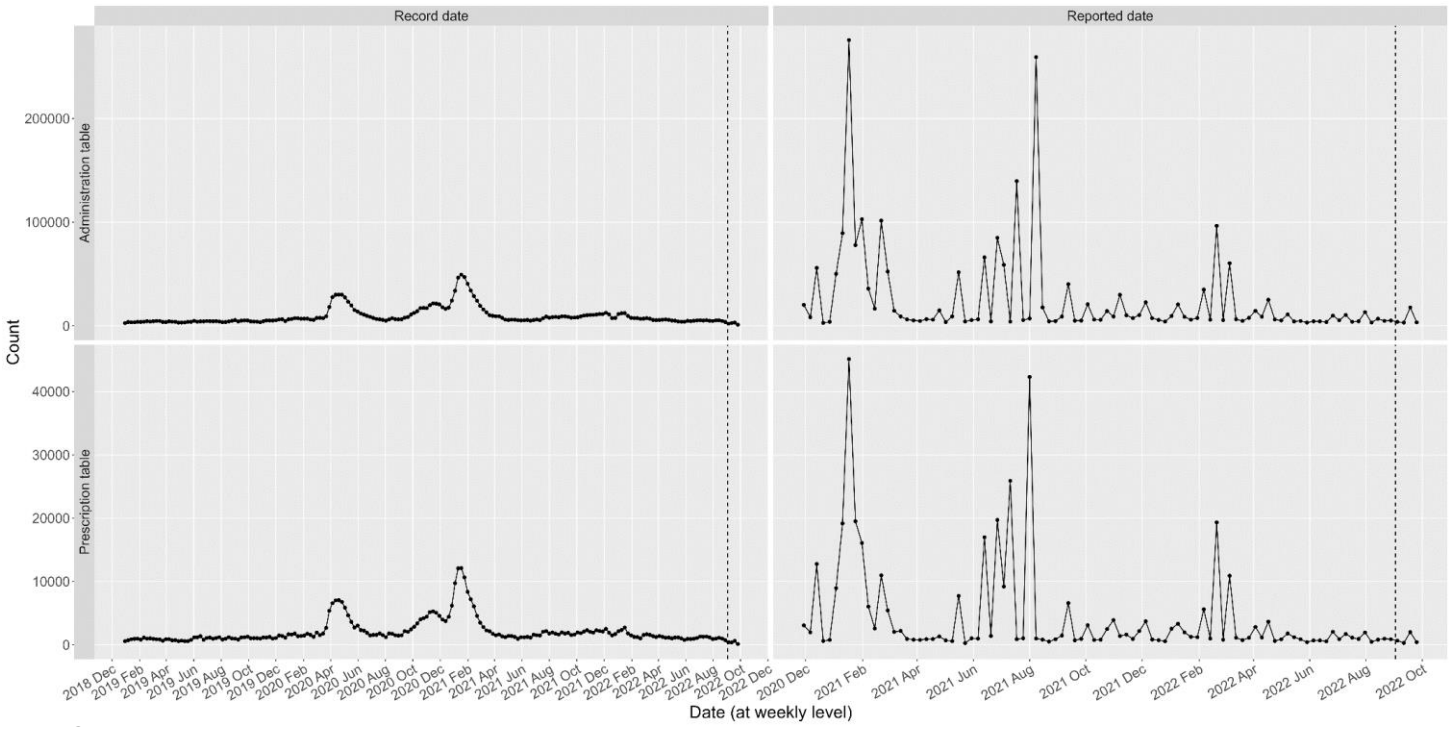


Figure legend: Record date is the date in which an event has occurred. Reported date is the date in which that record was reported to the centralised EPMA data collection system. The vertical axis shows individual record counts, aggregated at a weekly level. The vertical dashed line shows the date of data reception (30th September 2022).

Figure 4.6 - EPMA data stability across consecutive extracts (split by Prescription vs Administration tables)

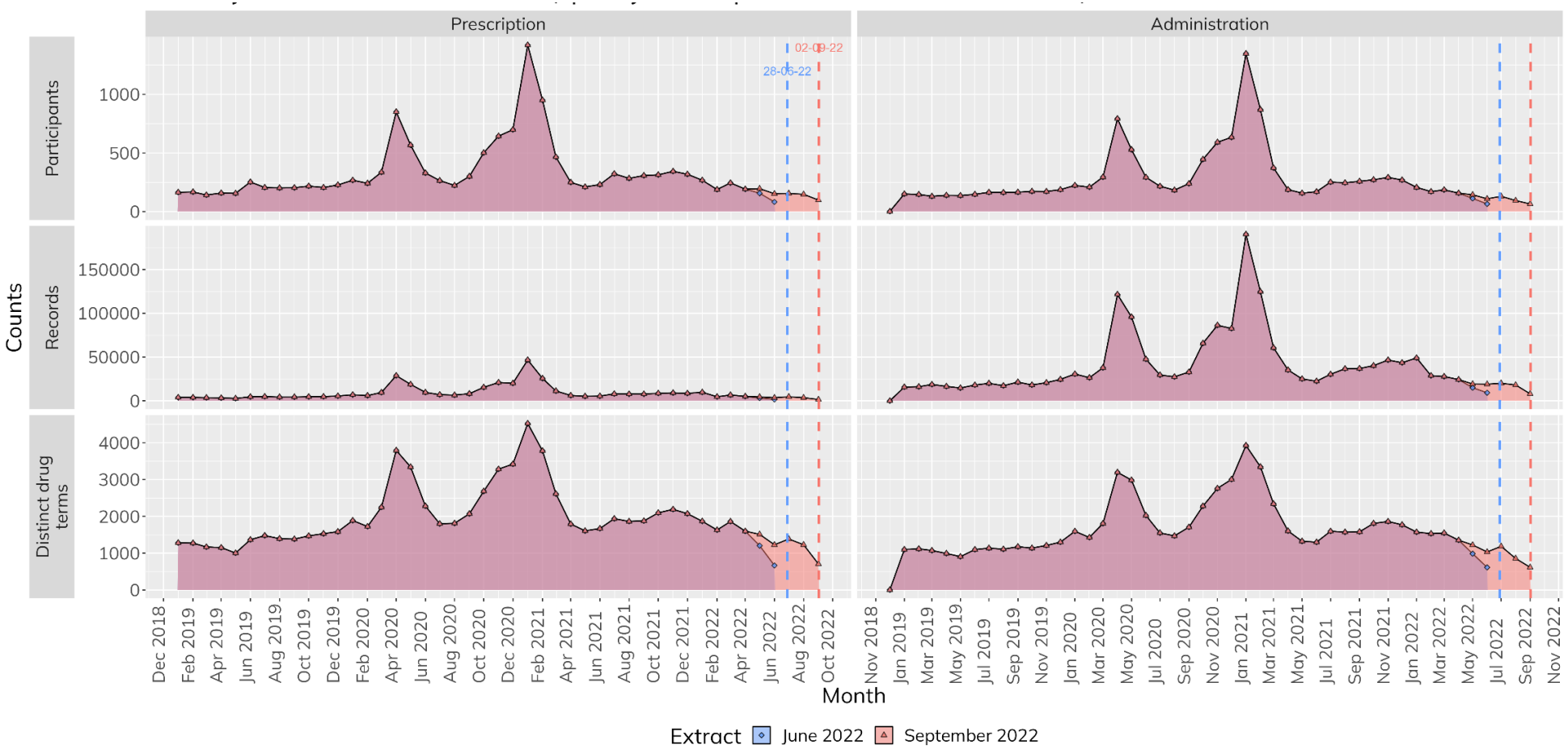


Figure legend: Individual panels are presented for each metric (total participants, total records, and distinct drug terms) and EPMA data table. The number of distinct participants, records, and drug terms in each month is represented by blue squares (June 2022 extract) and red triangles (September 2022 extract). The area below each curve is shaded with the respective colour (with transparency) to illustrate the overlap between both curves. The vertical distance between symbols in a single month represents records added retrospectively in the subsequent extract. Vertical dashed lines show the date in which each extract was received.

4.4.2 Data-driven methodology development

Table 4.7-4.12 depict the sensitivity analyses performed for different data handling rules to extract baseline drug exposure from EPMA data (overall and by drug). For each rule, the option which yielded the maximum agreement (κ and PABAK) was selected (or else the simplest rule if agreement was similar). A final data derivation algorithm was defined using the following rule set:

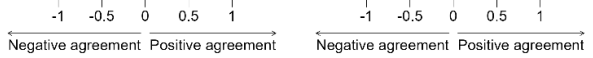
1. *Administration* table date fields: administered date
2. *Prescription* vs *administration* tables: use both
3. Prescription type, status, and formulation: ignore
4. Data recorded in the day of randomisation: consider as baseline exposure

1 Table 4.7 - Sensitivity analysis: administration table date fields

2 *Combined (across all drugs)*

	Counts (participant/drug pairs)					Agreement metrics							
Date field	**Total**	**Both sources**	**CRF only**	**EPMA only**	**Neither source**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Kappa**	**Kappa (95% CI)**	**PABAK**	**PABAK (95% CI)**
Scheduled	24753	8489	3039	1639	11586	73.6	87.6	83.8	79.2		0.62 (0.61 to 0.63)		0.62 (0.61 to 0.63)
Administered	24753	8227	3301	1434	11791	71.4	89.2	85.2	78.1		0.61 (0.6 to 0.62)		0.62 (0.61 to 0.63)
Recorded	24753	8409	3119	1602	11623	72.9	87.9	84	78.8		0.61 (0.6 to 0.62)		0.62 (0.61 to 0.63)
Last updated	24753	8409	3119	1602	11623	72.9	87.9	84	78.8		0.61 (0.6 to 0.62)		0.62 (0.61 to 0.63)
Any*	24753	8464	3064	1629	11596	73.4	87.7	83.9	79.1		0.62 (0.61 to 0.63)		0.62 (0.61 to 0.63)

*Using any non-missing date in a hierarchical sequence: 1) administered 2) scheduled 3) recorded 4) last updated



4 *By drug*

	Counts (participants)					Agreement metrics							
Drug (by administration date field)	**Total**	**Both sources**	**CRF only**	**EPMA only**	**Neither source**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Kappa**	**Kappa (95% CI)**	**PABAK**	**PABAK (95% CI)**
Corticosteroids													
Scheduled	3333	2848	261	95	129	91.6	57.6	96.8	33.1		0.37 (0.3 to 0.43)		0.79 (0.76 to 0.81)
Administered	3333	2783	326	93	131	89.5	58.5	96.8	28.7		0.32 (0.26 to 0.38)		0.75 (0.73 to 0.77)
Recorded	3333	2841	268	94	130	91.4	58	96.8	32.7		0.36 (0.3 to 0.43)		0.78 (0.76 to 0.8)
Last updated	3333	2841	268	94	130	91.4	58	96.8	32.7		0.36 (0.3 to 0.43)		0.78 (0.76 to 0.8)
Any*	3333	2848	261	95	129	91.6	57.6	96.8	33.1		0.37 (0.3 to 0.43)		0.79 (0.76 to 0.81)
Macrolides													
Scheduled	3911	377	42	630	2862	90	82	37.4	98.6		0.44 (0.41 to 0.48)		0.66 (0.63 to 0.68)
Administered	3911	371	48	578	2914	88.5	83.4	39.1	98.4		0.46 (0.42 to 0.5)		0.68 (0.66 to 0.7)
Recorded	3911	375	44	626	2866	89.5	82.1	37.5	98.5		0.44 (0.41 to 0.48)		0.66 (0.63 to 0.68)
Last updated	3911	375	44	626	2866	89.5	82.1	37.5	98.5		0.44 (0.41 to 0.48)		0.66 (0.63 to 0.68)
Any*	3911	378	41	631	2861	90.2	81.9	37.5	98.6		0.45 (0.41 to 0.48)		0.66 (0.63 to 0.68)
Tocilizumab or sarilumab													
Scheduled	1238	198	58	169	813	77.3	82.8	54	93.3		0.52 (0.46 to 0.57)		0.63 (0.59 to 0.68)
Administered	1238	181	75	122	860	70.7	87.6	59.7	92		0.55 (0.49 to 0.6)		0.68 (0.64 to 0.72)
Recorded	1238	197	59	155	827	77	84.2	56	93.3		0.54 (0.48 to 0.59)		0.65 (0.61 to 0.7)
Last updated	1238	197	59	155	827	77	84.2	56	93.3		0.54 (0.48 to 0.59)		0.65 (0.61 to 0.7)
Any*	1238	199	57	165	817	77.7	83.2	54.7	93.5		0.53 (0.47 to 0.58)		0.64 (0.6 to 0.68)
Remdesivir													
Scheduled	3381	430	155	127	2669	73.5	95.5	77.2	94.5		0.7 (0.67 to 0.74)		0.83 (0.81 to 0.85)
Administered	3381	410	175	104	2692	70.1	96.3	79.8	93.9		0.7 (0.66 to 0.73)		0.83 (0.82 to 0.85)
Recorded	3381	426	159	123	2673	72.8	95.6	77.6	94.4		0.7 (0.67 to 0.73)		0.83 (0.81 to 0.85)
Last updated	3381	426	159	123	2673	72.8	95.6	77.6	94.4		0.7 (0.67 to 0.73)		0.83 (0.81 to 0.85)
Any*	3381	427	158	126	2670	73	95.5	77.2	94.4		0.7 (0.67 to 0.73)		0.83 (0.81 to 0.85)
Antiplatelets													
Scheduled	2803	361	88	341	2013	80.4	85.5	51.4	95.8		0.54 (0.5 to 0.58)		0.69 (0.67 to 0.72)
Administered	2803	358	91	323	2031	79.7	86.3	52.6	95.7		0.55 (0.51 to 0.59)		0.7 (0.68 to 0.73)
Recorded	2803	361	88	337	2017	80.4	85.7	51.7	95.8		0.54 (0.5 to 0.58)		0.7 (0.67 to 0.72)
Last updated	2803	361	88	337	2017	80.4	85.7	51.7	95.8		0.54 (0.5 to 0.58)		0.7 (0.67 to 0.72)
Any*	2803	361	88	340	2014	80.4	85.6	51.5	95.8		0.54 (0.5 to 0.58)		0.69 (0.67 to 0.72)
Oral anticoagulants													
Scheduled	2810	202	61	38	2509	76.8	98.5	84.2	97.6		0.78 (0.74 to 0.83)		0.93 (0.91 to 0.94)
Administered	2810	185	78	29	2518	70.3	98.9	86.4	97		0.76 (0.71 to 0.8)		0.92 (0.91 to 0.94)
Recorded	2810	196	67	36	2511	74.5	98.6	84.5	97.4		0.77 (0.73 to 0.82)		0.93 (0.91 to 0.94)
Last updated	2810	196	67	36	2511	74.5	98.6	84.5	97.4		0.77 (0.73 to 0.82)		0.93 (0.91 to 0.94)
Any*	2810	201	62	38	2509	76.4	98.5	84.1	97.6		0.78 (0.74 to 0.82)		0.93 (0.91 to 0.94)
LMWH													
Scheduled	2964	2454	225	73	212	91.6	74.4	97.1	48.5		0.53 (0.48 to 0.58)		0.8 (0.78 to 0.82)
Administered	2964	2398	281	65	220	89.5	77.2	97.4	43.9		0.5 (0.45 to 0.55)		0.77 (0.74 to 0.79)
Recorded	2964	2437	242	71	214	91	75.1	97.2	46.9		0.52 (0.47 to 0.57)		0.79 (0.77 to 0.81)
Last updated	2964	2437	242	71	214	91	75.1	97.2	46.9		0.52 (0.47 to 0.57)		0.79 (0.77 to 0.81)
Any*	2964	2454	225	73	212	91.6	74.4	97.1	48.5		0.53 (0.48 to 0.58)		0.8 (0.78 to 0.82)
Oxygen													
Scheduled	4313	1619	2149	166	379	43	69.5	90.7	15		0.05 (0.02 to 0.08)		-0.07 (-0.1 to -0.04)
Administered	4313	1541	2227	120	425	40.9	78	92.8	16		0.07 (0.05 to 0.1)		-0.09 (-0.12 to -0.06)
Recorded	4313	1576	2192	160	385	41.8	70.6	90.8	14.9		0.05 (0.02 to 0.07)		-0.09 (-0.12 to -0.06)
Last updated	4313	1576	2192	160	385	41.8	70.6	90.8	14.9		0.05 (0.02 to 0.07)		-0.09 (-0.12 to -0.06)
Any*	4313	1596	2172	161	384	42.4	70.5	90.8	15		0.05 (0.02 to 0.08)		-0.08 (-0.11 to -0.05)

*Using any non-missing date in a hierarchical sequence: 1) administered 2) scheduled 3) recorded 4) last updated

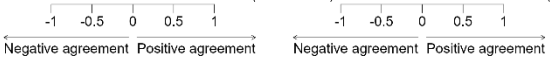


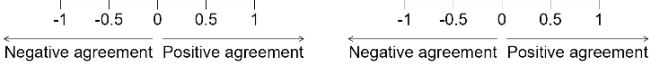
Table 4.8 - Sensitivity analysis: prescription vs administration tables (combined and by drug)

Combined (across all drugs)

Counts (participant/drug pairs)

Data table	Total	Both sources	CRF only	EPMA only	Neither source	Sensitivity	Specificity	PPV	NPV	Kappa	Kappa (95% CI)	PABAK	PABAK (95% CI)
Prescription	24628	8534	2942	1758	11394	74.4	86.6	82.9	79.5		0.61 (0.6 to 0.62)		0.62 (0.61 to 0.63)
Administration	24374	8227	3155	1434	11558	72.3	89	85.2	78.6		0.62 (0.61 to 0.63)		0.62 (0.61 to 0.63)
Any	24753	8681	2847	1781	11444	75.3	86.5	83	80.1		0.62 (0.61 to 0.63)		0.63 (0.62 to 0.64)

Agreement metrics



By drug

Counts (participants)

Drug (by dataset table used)	Total	Both sources	CRF only	EPMA only	Neither source	Sensitivity	Specificity	PPV	NPV	Kappa	Kappa (95% CI)	PABAK	PABAK (95% CI)
Corticosteroids													
Prescription	3315	2827	267	99	122	91.4	55.2	96.6	31.4		0.34 (0.28 to 0.41)		0.78 (0.76 to 0.8)
Administration	3290	2783	291	93	123	90.5	56.9	96.8	29.7		0.33 (0.27 to 0.4)		0.77 (0.74 to 0.79)
Any	3333	2859	250	102	122	92	54.5	96.6	32.8		0.36 (0.29 to 0.42)		0.79 (0.77 to 0.81)
Macrolides													
Prescription	3893	374	45	668	2806	89.3	80.8	35.9	98.4		0.42 (0.39 to 0.46)		0.63 (0.61 to 0.66)
Administration	3833	371	42	578	2842	89.8	83.1	39.1	98.5		0.46 (0.43 to 0.5)		0.68 (0.65 to 0.7)
Any	3911	384	35	676	2816	91.6	80.6	36.2	98.8		0.43 (0.39 to 0.47)		0.64 (0.61 to 0.66)
Tocilizumab or sarilumab													
Prescription	1229	201	54	169	805	78.8	82.6	54.3	93.7		0.53 (0.47 to 0.58)		0.64 (0.59 to 0.68)
Administration	1225	181	71	122	851	71.8	87.5	59.7	92.3		0.55 (0.49 to 0.61)		0.68 (0.64 to 0.72)
Any	1238	202	54	169	813	78.9	82.8	54.4	93.8		0.53 (0.47 to 0.59)		0.64 (0.59 to 0.68)
Remdesivir													
Prescription	3363	431	153	133	2646	73.8	95.2	76.4	94.5		0.7 (0.67 to 0.73)		0.83 (0.81 to 0.85)
Administration	3336	410	174	104	2648	70.2	96.2	79.8	93.8		0.7 (0.66 to 0.73)		0.83 (0.81 to 0.85)
Any	3381	431	154	134	2662	73.7	95.2	76.3	94.5		0.7 (0.66 to 0.73)		0.83 (0.81 to 0.85)
Antiplatelets													
Prescription	2791	361	86	349	1995	80.8	85.1	50.8	95.9		0.53 (0.49 to 0.57)		0.69 (0.66 to 0.71)
Administration	2768	358	84	323	2003	81	86.1	52.6	96		0.55 (0.51 to 0.59)		0.71 (0.68 to 0.73)
Any	2803	368	81	351	2003	82	85.1	51.2	96.1		0.54 (0.5 to 0.58)		0.69 (0.66 to 0.72)
Oral anticoagulants													
Prescription	2798	200	62	60	2476	76.3	97.6	76.9	97.6		0.74 (0.7 to 0.79)		0.91 (0.9 to 0.93)
Administration	2775	185	72	29	2489	72	98.8	86.4	97.2		0.77 (0.72 to 0.81)		0.93 (0.91 to 0.94)
Any	2810	202	61	61	2486	76.8	97.6	76.8	97.6		0.74 (0.7 to 0.79)		0.91 (0.9 to 0.93)
LMWH													
Prescription	2948	2421	245	95	187	90.8	66.3	96.2	43.3		0.46 (0.41 to 0.52)		0.77 (0.75 to 0.79)
Administration	2925	2398	250	65	212	90.6	76.5	97.4	45.9		0.52 (0.47 to 0.57)		0.78 (0.76 to 0.81)
Any	2964	2471	208	97	188	92.2	66	96.2	47.5		0.5 (0.44 to 0.55)		0.79 (0.77 to 0.82)
Oxygen													
Prescription	4291	1719	2030	185	357	45.9	65.9	90.3	15		0.05 (0.02 to 0.08)		-0.03 (-0.06 to 0)
Administration	4222	1541	2171	120	390	41.5	76.5	92.8	15.2		0.07 (0.04 to 0.09)		-0.09 (-0.12 to -0.05)
Any	4313	1764	2004	191	354	46.8	65	90.2	15		0.05 (0.02 to 0.08)		-0.02 (-0.05 to 0.01)

Agreement metrics

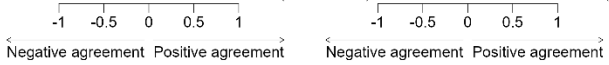
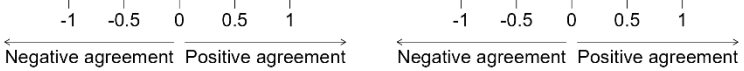


Table 4.9 - Sensitivity analysis: prescription type (combined and by drug)

Combined (across all drugs)

	Counts (participant/drug pairs)					Agreement metrics							
Prescription type	Total	Both sources	CRF only	EPMA only	Neither source	Sensitivity	Specificity	PPV	NPV	Kappa	Kappa (95% CI)	PABAK	PABAK (95% CI)
Inpatient	24753	8645	2883	1767	11458	75	86.6	83	79.9		0.62 (0.61 to 0.63)		0.62 (0.61 to 0.63)
Any	24753	8681	2847	1781	11444	75.3	86.5	83	80.1		0.62 (0.61 to 0.63)		0.63 (0.62 to 0.64)



By drug

	Counts (participants)					Agreement metrics							
Drug (by prescription status)	Total	Both sources	CRF only	EPMA only	Neither source	Sensitivity	Specificity	PPV	NPV	Kappa	Kappa (95% CI)	PABAK	PABAK (95% CI)
Corticosteroids													
Inpatient	3333	2844	265	102	122	91.5	54.5	96.5	31.5		0.34 (0.28 to 0.41)		0.78 (0.76 to 0.8)
Any	3333	2859	250	102	122	92	54.5	96.6	32.8		0.36 (0.29 to 0.42)		0.79 (0.77 to 0.81)
Macrolides													
Inpatient	3911	383	36	673	2819	91.4	80.7	36.3	98.7		0.43 (0.39 to 0.47)		0.64 (0.61 to 0.66)
Any	3911	384	35	676	2816	91.6	80.6	36.2	98.8		0.43 (0.39 to 0.47)		0.64 (0.61 to 0.66)
Tocilizumab or sarilumab													
Inpatient	1238	202	54	169	813	78.9	82.8	54.4	93.8		0.53 (0.47 to 0.59)		0.64 (0.59 to 0.68)
Any	1238	202	54	169	813	78.9	82.8	54.4	93.8		0.53 (0.47 to 0.59)		0.64 (0.59 to 0.68)
Remdesivir													
Inpatient	3381	431	154	133	2663	73.7	95.2	76.4	94.5		0.7 (0.67 to 0.73)		0.83 (0.81 to 0.85)
Any	3381	431	154	134	2662	73.7	95.2	76.3	94.5		0.7 (0.66 to 0.73)		0.83 (0.81 to 0.85)
Antiplatelets													
Inpatient	2803	364	85	350	2004	81.1	85.1	51	95.9		0.53 (0.49 to 0.57)		0.69 (0.66 to 0.72)
Any	2803	368	81	351	2003	82	85.1	51.2	96.1		0.54 (0.5 to 0.58)		0.69 (0.66 to 0.72)
Oral anticoagulants													
Inpatient	2810	202	61	53	2494	76.8	97.9	79.2	97.6		0.76 (0.71 to 0.8)		0.92 (0.9 to 0.93)
Any	2810	202	61	61	2486	76.8	97.6	76.8	97.6		0.74 (0.7 to 0.79)		0.91 (0.9 to 0.93)
LMWH													
Inpatient	2964	2458	221	97	188	91.8	66	96.2	46		0.48 (0.43 to 0.54)		0.79 (0.76 to 0.81)
Any	2964	2471	208	97	188	92.2	66	96.2	47.5		0.5 (0.44 to 0.55)		0.79 (0.77 to 0.82)
Oxygen													
Inpatient	4313	1761	2007	190	355	46.7	65.1	90.3	15		0.05 (0.02 to 0.08)		-0.02 (-0.05 to 0.01)
Any	4313	1764	2004	191	354	46.8	65	90.2	15		0.05 (0.02 to 0.08)		-0.02 (-0.05 to 0.01)

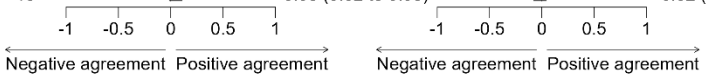
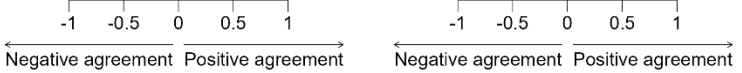


Table 4.10 - Sensitivity analysis: prescription status (combined and by drug)

Combined (across all drugs)

	Counts (participant/drug pairs)					Agreement metrics							
Prescription status	**Total**	**Both sources**	**CRF only**	**EPMA only**	**Neither source**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Kappa**	**Kappa (95% CI)**	**PABAK**	**PABAK (95% CI)**
Active	24753	5646	5882	978	12247	49	92.6	85.2	67.6		0.43 (0.42 to 0.44)		0.45 (0.43 to 0.46)
Any	24753	8681	2847	1781	11444	75.3	86.5	83	80.1		0.62 (0.61 to 0.63)		0.63 (0.62 to 0.64)



By drug

	Counts (participants)					Agreement metrics							
Drug (by prescription status)	**Total**	**Both sources**	**CRF only**	**EPMA only**	**Neither source**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Kappa**	**Kappa (95% CI)**	**PABAK**	**PABAK (95% CI)**
Corticosteroids													
Active	3333	1538	1571	59	165	49.5	73.7	96.3	9.5		0.06 (0.02 to 0.09)		0.02 (-0.01 to 0.06)
Any	3333	2859	250	102	122	92	54.5	96.6	32.8		0.36 (0.29 to 0.42)		0.79 (0.77 to 0.81)
Macrolides													
Active	3911	138	281	232	3260	32.9	93.4	37.3	92.1		0.28 (0.22 to 0.34)		0.74 (0.72 to 0.76)
Any	3911	384	35	676	2816	91.6	80.6	36.2	98.8		0.43 (0.39 to 0.47)		0.64 (0.61 to 0.66)
Tocilizumab or sarilumab													
Active	1238	88	168	115	867	34.4	88.3	43.3	83.8		0.25 (0.17 to 0.32)		0.54 (0.49 to 0.59)
Any	1238	202	54	169	813	78.9	82.8	54.4	93.8		0.53 (0.47 to 0.59)		0.64 (0.59 to 0.68)
Remdesivir													
Active	3381	152	433	38	2758	26	98.6	80	86.4		0.34 (0.28 to 0.39)		0.72 (0.7 to 0.74)
Any	3381	431	154	134	2662	73.7	95.2	76.3	94.5		0.7 (0.66 to 0.73)		0.83 (0.81 to 0.85)
Antiplatelets													
Active	2803	329	120	278	2076	73.3	88.2	54.2	94.5		0.54 (0.5 to 0.58)		0.72 (0.69 to 0.74)
Any	2803	368	81	351	2003	82	85.1	51.2	96.1		0.54 (0.5 to 0.58)		0.69 (0.66 to 0.72)
Oral anticoagulants													
Active	2810	157	106	34	2513	59.7	98.7	82.2	96		0.67 (0.61 to 0.72)		0.9 (0.88 to 0.92)
Any	2810	202	61	61	2486	76.8	97.6	76.8	97.6		0.74 (0.7 to 0.79)		0.91 (0.9 to 0.93)
LMWH													
Active	2964	1679	1000	47	238	62.7	83.5	97.3	19.2		0.19 (0.15 to 0.22)		0.29 (0.26 to 0.33)
Any	2964	2471	208	97	188	92.2	66	96.2	47.5		0.5 (0.44 to 0.55)		0.79 (0.77 to 0.82)
Oxygen													
Active	4313	1565	2203	175	370	41.5	67.9	89.9	14.4		0.04 (0.01 to 0.06)		-0.1 (-0.13 to -0.07)
Any	4313	1764	2004	191	354	46.8	65	90.2	15		0.05 (0.02 to 0.08)		-0.02 (-0.05 to 0.01)

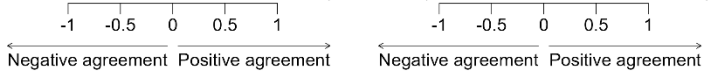
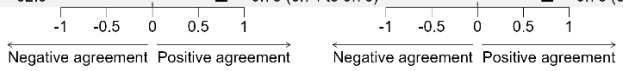


Table 4.11 - Sensitivity analysis: prescription formulation (combined and by drug)

Combined (across all drugs)

Prescription formulation	Total	Both sources	CRF only	EPMA only	Neither source	Sensitivity	Specificity	PPV	NPV	Kappa	Kappa (95% CI)	PABAK	PABAK (95% CI)
	Counts (participant/drug pairs)					Agreement metrics							
Systemic	20440	6789	971	1490	11190	87.5	88.2	82	92		0.75 (0.74 to 0.76)		0.76 (0.75 to 0.77)
Systemic or unknown	20440	6917	843	1589	11091	89.1	87.5	81.3	92.9		0.75 (0.74 to 0.76)		0.76 (0.75 to 0.77)
Any	20440	6917	843	1590	11090	89.1	87.5	81.3	92.9		0.75 (0.74 to 0.76)		0.76 (0.75 to 0.77)



By drug

Drug (by formulation)	Total	Both sources	CRF only	EPMA only	Neither source	Sensitivity	Specificity	PPV	NPV	Kappa	Kappa (95% CI)	PABAK	PABAK (95% CI)
	Counts (participants)					Agreement metrics							
Corticosteroids													
Systemic	3333	2846	263	101	123	91.5	54.9	96.6	31.9		0.35 (0.28 to 0.41)		0.78 (0.76 to 0.8)
Systemic or unknown	3333	2859	250	101	123	92	54.9	96.6	33		0.36 (0.29 to 0.42)		0.79 (0.77 to 0.81)
Any	3333	2859	250	102	122	92	54.5	96.6	32.8		0.36 (0.29 to 0.42)		0.79 (0.77 to 0.81)
Macrolides													
Systemic	3911	380	39	655	2837	90.7	81.2	36.7	98.6		0.44 (0.4 to 0.47)		0.65 (0.62 to 0.67)
Systemic or unknown	3911	384	35	676	2816	91.6	80.6	36.2	98.8		0.43 (0.39 to 0.47)		0.64 (0.61 to 0.66)
Any	3911	384	35	676	2816	91.6	80.6	36.2	98.8		0.43 (0.39 to 0.47)		0.64 (0.61 to 0.66)
Tocilizumab or sarilumab													
Systemic	1238	176	80	123	859	68.8	87.5	58.9	91.5		0.53 (0.47 to 0.59)		0.67 (0.63 to 0.71)
Systemic or unknown	1238	202	54	169	813	78.9	82.8	54.4	93.8		0.53 (0.47 to 0.59)		0.64 (0.59 to 0.68)
Any	1238	202	54	169	813	78.9	82.8	54.4	93.8		0.53 (0.47 to 0.59)		0.64 (0.59 to 0.68)
Remdesivir													
Systemic	3381	363	222	117	2679	62.1	95.8	75.6	92.3		0.62 (0.58 to 0.66)		0.8 (0.78 to 0.82)
Systemic or unknown	3381	431	154	134	2662	73.7	95.2	76.3	94.5		0.7 (0.66 to 0.73)		0.83 (0.81 to 0.85)
Any	3381	431	154	134	2662	73.7	95.2	76.3	94.5		0.7 (0.66 to 0.73)		0.83 (0.81 to 0.85)
Antiplatelets													
Systemic	2803	367	82	336	2018	81.7	85.7	52.2	96.1		0.55 (0.51 to 0.59)		0.7 (0.67 to 0.73)
Systemic or unknown	2803	368	81	351	2003	82	85.1	51.2	96.1		0.54 (0.5 to 0.58)		0.69 (0.66 to 0.72)
Any	2803	368	81	351	2003	82	85.1	51.2	96.1		0.54 (0.5 to 0.58)		0.69 (0.66 to 0.72)
Oral anticoagulants													
Systemic	2810	199	64	61	2486	75.7	97.6	76.5	97.5		0.74 (0.69 to 0.78)		0.91 (0.89 to 0.93)
Systemic or unknown	2810	202	61	61	2486	76.8	97.6	76.8	97.6		0.74 (0.7 to 0.79)		0.91 (0.9 to 0.93)
Any	2810	202	61	61	2486	76.8	97.6	76.8	97.6		0.74 (0.7 to 0.79)		0.91 (0.9 to 0.93)
LMWH													
Systemic	2964	2458	221	97	188	91.8	66	96.2	46		0.48 (0.43 to 0.54)		0.79 (0.76 to 0.81)
Systemic or unknown	2964	2471	208	97	188	92.2	66	96.2	47.5		0.5 (0.44 to 0.55)		0.79 (0.77 to 0.82)
Any	2964	2471	208	97	188	92.2	66	96.2	47.5		0.5 (0.44 to 0.55)		0.79 (0.77 to 0.82)
Oxygen													
Inhaled	4313	497	3271	93	452	13.2	82.9	84.2	12.1		-0.01 (-0.03 to 0)		-0.56 (-0.58 to -0.53)
Inhaled or unknown	4313	1724	2044	191	354	45.8	65	90	14.8		0.04 (0.02 to 0.07)		-0.04 (-0.07 to -0.01)
Any	4313	1764	2004	191	354	46.8	65	90.2	15		0.05 (0.02 to 0.08)		-0.02 (-0.05 to 0.01)

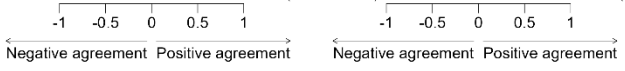
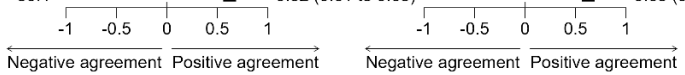


Table 4.12 - Sensitivity analysis: handling of data recorded in the day of randomisation (combined and by drug)

Combined (across all drugs)

	Counts (participant/drug pairs)					Agreement metrics							
Data in the day of randomisation	Total	Both sources	CRF only	EPMA only	Neither source	Sensitivity	Specificity	PPV	NPV	Kappa	Kappa (95% CI)	PABAK	PABAK (95% CI)
Exclude	24753	5983	5545	916	12309	51.9	93.1	86.7	68.9		0.46 (0.45 to 0.47)		0.48 (0.47 to 0.49)
Include	24753	8681	2847	1781	11444	75.3	86.5	83	80.1		0.62 (0.61 to 0.63)		0.63 (0.62 to 0.64)



By drug

	Counts (participants)					Agreement metrics							
Drug (by randomisation date handling rule)	Total	Both sources	CRF only	EPMA only	Neither source	Sensitivity	Specificity	PPV	NPV	Kappa	Kappa (95% CI)	PABAK	PABAK (95% CI)
Corticosteroids													
Exclude	3333	1892	1217	72	152	60.9	67.9	96.3	11.1		0.09 (0.05 to 0.12)		0.23 (0.19 to 0.26)
Include	3333	2859	250	102	122	92	54.5	96.6	32.8		0.36 (0.29 to 0.42)		0.79 (0.77 to 0.81)
Macrolides													
Exclude	3911	303	116	471	3021	72.3	86.5	39.1	96.3		0.43 (0.39 to 0.47)		0.7 (0.68 to 0.72)
Include	3911	384	35	676	2816	91.6	80.6	36.2	98.8		0.43 (0.39 to 0.47)		0.64 (0.61 to 0.66)
Tocilizumab or sarilumab													
Exclude	1238	148	108	50	932	57.8	94.9	74.7	89.6		0.58 (0.51 to 0.64)		0.74 (0.7 to 0.78)
Include	1238	202	54	169	813	78.9	82.8	54.4	93.8		0.53 (0.47 to 0.59)		0.64 (0.59 to 0.68)
Remdesivir													
Exclude	3381	255	330	69	2727	43.6	97.5	78.7	89.2		0.5 (0.45 to 0.55)		0.76 (0.74 to 0.79)
Include	3381	431	154	134	2662	73.7	95.2	76.3	94.5		0.7 (0.66 to 0.73)		0.83 (0.81 to 0.85)
Antiplatelets													
Exclude	2803	236	213	37	2317	52.6	98.4	86.4	91.6		0.61 (0.56 to 0.65)		0.82 (0.8 to 0.84)
Include	2803	368	81	351	2003	82	85.1	51.2	96.1		0.54 (0.5 to 0.58)		0.69 (0.66 to 0.72)
Oral anticoagulants													
Exclude	2810	138	125	41	2506	52.5	98.4	77.1	95.2		0.59 (0.53 to 0.65)		0.88 (0.86 to 0.9)
Include	2810	202	61	61	2486	76.8	97.6	76.8	97.6		0.74 (0.7 to 0.79)		0.91 (0.9 to 0.93)
LMWH													
Exclude	2964	1762	917	64	221	65.8	77.5	96.5	19.4		0.19 (0.14 to 0.23)		0.34 (0.3 to 0.37)
Include	2964	2471	208	97	188	92.2	66	96.2	47.5		0.5 (0.44 to 0.55)		0.79 (0.77 to 0.82)
Oxygen													
Exclude	4313	1249	2519	112	433	33.1	79.4	91.8	14.7		0.04 (0.02 to 0.07)		-0.22 (-0.25 to -0.19)
Include	4313	1764	2004	191	354	46.8	65	90.2	15		0.05 (0.02 to 0.08)		-0.02 (-0.05 to 0.01)

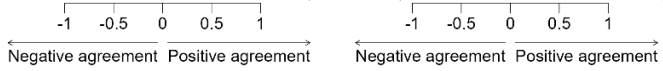


Table legend: Include – data in the day of randomisation included as pre-randomisation exposure; exclude - data in the day of randomisation excluded as pre-randomisation exposure

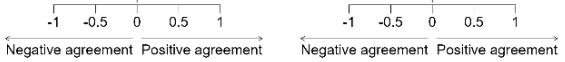
4.4.3 Agreement between EPMA and CRF data

4.4.3.1 Binary baseline and follow-up drug exposure

Overall agreement for baseline exposure was high (κ: 0.62; PABAK: 0.63), and increased if excluding oxygen (κ: 0.75; PABAK: 0.76). Agreement by drug was modest to high (κ: 0.36 for corticosteroids, to 0.74 for oral anticoagulants; PABAK: 0.4 for macrolides, to 0.86 for remdesivir), except for oxygen which exhibited poor agreement (κ: 0.05; PABAK: -0.02; Table 4.13). If excluding oxygen, sensitivity, specificity, PPV and NPV ranged from 73.7%-92.2%, 54.5%-97.6%, 36.2%-96.6%, and 32.8%-98.8% respectively. Besides corticosteroids and LMWH, all drugs had specificity >80%. PPVs were >90% for LMWH and corticosteroids only (both drugs with very high overall prevalence across both EPMA and the CRF). Conversely, NPVs were >93% for most drugs except corticosteroids (32.8%) and LMWH (47.5%).

Table 4.13 - Agreement between EPMA and CRF data for baseline exposure

	Counts (participants or participant/drug pairs)					Agreement metrics							
Drug	**Total**	**Both sources**	**CRF only**	**EPMA only**	**Neither source**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Kappa**	**Kappa (95% CI)**	**PABAK**	**PABAK (95% CI)**
Corticosteroids	3333	2859	250	102	122	92	54.5	96.6	32.8		0.36 (0.29 to 0.42)		0.79 (0.77 to 0.81)
Macrolides	3911	384	35	676	2816	91.6	80.6	36.2	98.8		0.43 (0.39 to 0.47)		0.64 (0.61 to 0.66)
Tocilizumab or sarilumab	1238	202	54	169	813	78.9	82.8	54.4	93.8		0.53 (0.47 to 0.59)		0.64 (0.59 to 0.68)
Remdesivir	3381	431	154	134	2662	73.7	95.2	76.3	94.5		0.7 (0.66 to 0.73)		0.83 (0.81 to 0.85)
Antiplatelets	2803	368	81	351	2003	82	85.1	51.2	96.1		0.54 (0.5 to 0.58)		0.69 (0.66 to 0.72)
Oral anticoagulants	2810	202	61	61	2486	76.8	97.6	76.8	97.6		0.74 (0.7 to 0.79)		0.91 (0.9 to 0.93)
LMWH	2964	2471	208	97	188	92.2	66	96.2	47.5		0.5 (0.44 to 0.55)		0.79 (0.77 to 0.82)
Oxygen	4313	1764	2004	191	354	46.8	65	90.2	15		0.05 (0.02 to 0.08)		-0.02 (-0.05 to 0.01)
Combined	**24753**	**8681**	**2847**	**1781**	**11444**	**75.3**	**86.5**	**83**	**80.1**		**0.62 (0.61 to 0.63)**		**0.63 (0.62 to 0.64)**
Combined (excluding oxygen)	**20440**	**6917**	**843**	**1590**	**11090**	**89.1**	**87.5**	**81.3**	**92.9**		**0.75 (0.74 to 0.76)**		**0.76 (0.75 to 0.77)**

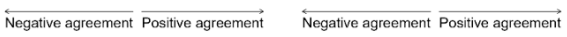


For follow-up drug exposure (Table 4.14), overall agreement was high (κ: 0.77; PABAK: 0.85) and also varied by drug, with corticosteroids exhibiting the lowest agreement (κ: 0.59, PABAK: 0.68) and baricitinib the highest (κ: 0.91; PABAK: 0.94). Sensitivity, specificity, PPV and NPV ranged from 19.3%-91.8%, 67.9%-100%, 68.8%-100%, and

72.6%-99.4% respectively. Four drugs had sensitivity below 70%: lopinavir-ritonavir 19.3%; tocilizumab/salimumab 31.4%; REGN antibodies 49.5%; macrolides 68.7%. Besides corticosteroids, all drugs had specificity >90%. PPVs were >85% for all drugs except aspirin (68.8%), macrolides (81.2%), and tocilizumab/sarilumab (83.8%), while NPVs were >90% for all drugs except corticosteroids (72.6%). Of note, agreement metrics for follow-up exposure were generally higher than those seen for baseline exposure.

Table 4.14 – Agreement metrics between EPMA and case report form (CRF) data for follow-up exposure

	Counts (participants or participant/drug pairs)					Agreement metrics							
Drug	**Total**	**Both sources**	**CRF only**	**EPMA only**	**Neither source**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Kappa**	**Kappa (95% CI)**	**PABAK**	**PABAK (95% CI)**
Corticosteroids	4313	2802	308	386	817	90.1	67.9	87.9	72.6		0.59 (0.56 to 0.62)		0.68 (0.66 to 0.7)
Tocilizumab or sarilumab	4149	165	360	32	3592	31.4	99.1	83.8	90.9		0.42 (0.36 to 0.47)		0.81 (0.79 to 0.83)
Macrolides	4308	620	282	144	3262	68.7	95.8	81.2	92		0.68 (0.65 to 0.71)		0.8 (0.78 to 0.82)
Remdesivir	3510	491	173	51	2795	73.9	98.2	90.6	94.2		0.78 (0.75 to 0.8)		0.87 (0.86 to 0.89)
Hydroxychloroquine	4313	114	24	15	4160	82.6	99.6	88.4	99.4		0.85 (0.8 to 0.9)		0.98 (0.98 to 0.99)
Lopinavir-ritonavir	4313	22	92	0	4199	19.3	100	100	97.9		0.32 (0.18 to 0.46)		0.96 (0.95 to 0.97)
Colchicine	2863	498	90	31	2244	84.7	98.6	94.1	96.1		0.87 (0.84 to 0.89)		0.92 (0.9 to 0.93)
Baricitinib	1990	369	33	22	1566	91.8	98.6	94.4	97.9		0.91 (0.89 to 0.94)		0.94 (0.93 to 0.96)
Aspirin	4313	769	92	348	3104	89.3	89.9	68.8	97.1		0.71 (0.69 to 0.74)		0.8 (0.78 to 0.81)
REGN antibodies	3297	213	217	33	2834	49.5	98.8	86.6	92.9		0.59 (0.54 to 0.64)		0.85 (0.83 to 0.87)
Combined	**37369**	**6063**	**1671**	**1062**	**28573**	**78.4**	**96.4**	**85.1**	**94.5**		**0.77 (0.76 to 0.78)**		**0.85 (0.85 to 0.86)**



After stratification by respiratory status at baseline, overall agreement for both baseline and follow-up exposure was lower among people on IMV/ECMO when compared with those on no oxygen/ventilation, or on oxygen/NIV (Table 4.15-4.16). This was associated with lower overall sensitivity (baseline: no oxygen/NIV 89.4%; oxygen/NIV 88.1%, IMV/ECMO 69.5%; follow-up: 80.8%, 79.9%, 44.4% respectively), while specificity decreased slightly for baseline exposure (90%, 87.8%, 85.7%) and increased to a similar degree for follow-up exposure (95.4%, 95.5%, 97%). Following additional stratification by drug, sensitivity was generally reduced with increasing levels of respiratory support for both baseline and follow-up exposure, across all drugs. On the other hand, specificity at baseline increased with higher respiratory support for antiplatelets, remained similar for

oral anticoagulants and LMWH, and decreased for corticosteroids, macrolides, tocilizumab/sarilumab, and remdesivir. This pattern was most obvious for corticosteroids, with sensitivity decreasing from 73.5% for people on no oxygen/NIV, to 26.8% for those on oxygen/NIV, and 30% for those on IMV/ECMO. For follow-up exposure, specificity was generally unchanged or increased with higher levels of respiratory support across all drugs (including corticosteroids, macrolides, remdesivir, and tocilizumab, and contrary to baseline exposure). Participants with baseline corticosteroid exposure identified only in EPMA data had a high prevalence of “inactive”/”unused” prescription statuses (49/102). Accordingly, specificity would be increased for steroids if only “active” prescription statuses had been used, albeit with lower overall agreement (as shown in the methodology development, Table 4.10). Manual data exploration confirmed the existence of multiple prescription and administration records on distinct dates, both before and after randomisation, representing continued prescriptions (rather than a single record). Follow-up exposure exhibited similar findings.

Table 4.15 - Agreement metrics between EPMA and case report form (CRF) data for baseline exposure, stratified by respiratory status at baseline

	Counts (participants or participant/drug pairs)					Agreement metrics							
Drug (by respiratory status)	**Total**	**Both sources**	**CRF only**	**EPMA only**	**Neither source**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Kappa**	**Kappa (95% CI)**	**PABAK**	**PABAK (95% CI)**
Corticosteroids													
No oxygen or ventilation	335	188	15	35	97	92.6	73.5	84.3	86.6		0.68 (0.6 to 0.76)		0.7 (0.62 to 0.77)
Oxygen requirement or non-invasive ventilatior	2897	2600	215	60	22	92.4	26.8	97.7	9.3		0.1 (0 to 0.2)		0.81 (0.79 to 0.83)
Mechanical ventilation or ECMO	101	71	20	7	3	78	30	91	13		0.05 (-0.26 to 0.36)		0.47 (0.27 to 0.63)
Macrolides													
No oxygen or ventilation	503	29	4	73	397	87.9	84.5	28.4	99		0.37 (0.24 to 0.5)		0.69 (0.62 to 0.75)
Oxygen requirement or non-invasive ventilatior	3243	338	26	571	2308	92.9	80.2	37.2	98.9		0.44 (0.4 to 0.48)		0.63 (0.6 to 0.66)
Mechanical ventilation or ECMO	165	17	5	32	111	77.3	77.6	34.7	95.7		0.36 (0.18 to 0.54)		0.55 (0.41 to 0.67)
Tocilizumab or sarilumab													
No oxygen or ventilation	106	5	0	4	97	100	96	55.6	100		0.7 (0.4 to 0.99)		0.92 (0.81 to 0.98)
Oxygen requirement or non-invasive ventilatior	1104	191	42	162	709	82	81.4	54.1	94.4		0.53 (0.48 to 0.59)		0.63 (0.58 to 0.68)
Mechanical ventilation or ECMO	28	6	12	3	7	33.3	70	66.7	36.8		0.03 (-0.31 to 0.36)		-0.07 (-0.45 to 0.32)
Remdesivir													
No oxygen or ventilation	353	8	3	6	336	72.7	98.2	57.1	99.1		0.63 (0.39 to 0.87)		0.95 (0.9 to 0.98)
Oxygen requirement or non-invasive ventilatior	2925	420	146	114	2245	74.2	95.2	78.7	93.9		0.71 (0.68 to 0.74)		0.82 (0.8 to 0.84)
Mechanical ventilation or ECMO	103	3	5	14	81	37.5	85.3	17.6	94.2		0.15 (-0.19 to 0.5)		0.63 (0.45 to 0.77)
Antiplatelets													
No oxygen or ventilation	280	50	13	29	188	79.4	86.6	63.3	93.5		0.61 (0.5 to 0.72)		0.7 (0.61 to 0.78)
Oxygen requirement or non-invasive ventilatior	2433	310	65	319	1739	82.7	84.5	49.3	96.4		0.53 (0.48 to 0.57)		0.68 (0.65 to 0.71)
Mechanical ventilation or ECMO	90	8	3	3	76	72.7	96.2	72.7	96.2		0.69 (0.45 to 0.93)		0.87 (0.72 to 0.95)
Oral anticoagulants													
No oxygen or ventilation	281	29	10	4	238	74.4	98.3	87.9	96		0.78 (0.66 to 0.89)		0.9 (0.84 to 0.94)
Oxygen requirement or non-invasive ventilatior	2438	173	50	55	2160	77.6	97.5	75.9	97.7		0.74 (0.7 to 0.79)		0.91 (0.9 to 0.93)
Mechanical ventilation or ECMO	91	0	1	2	88	0	97.8	0	98.9		-0.01 (-1.14 to 1.11)		0.93 (0.81 to 0.99)
LMWH													
No oxygen or ventilation	297	231	19	15	32	92.4	68.1	93.9	62.7		0.58 (0.45 to 0.72)		0.77 (0.69 to 0.84)
Oxygen requirement or non-invasive ventilatior	2573	2172	166	81	154	92.9	65.5	96.4	48.1		0.5 (0.44 to 0.56)		0.81 (0.78 to 0.83)
Mechanical ventilation or ECMO	94	68	23	1	2	74.7	66.7	98.6	8		0.09 (-0.22 to 0.4)		0.49 (0.29 to 0.66)
Combined													
No oxygen or ventilation	**2155**	**540**	**64**	**166**	**1385**	**89.4**	**89.3**	**76.5**	**95.6**		**0.75 (0.72 to 0.78)**		**0.79 (0.76 to 0.81)**
Oxygen requirement or non-invasive ventila	**17613**	**6204**	**710**	**1362**	**9337**	**89.7**	**87.3**	**82**	**92.9**		**0.76 (0.75 to 0.77)**		**0.76 (0.76 to 0.77)**
Mechanical ventilation or ECMO	**672**	**173**	**69**	**62**	**368**	**71.5**	**85.6**	**73.6**	**84.2**		**0.57 (0.51 to 0.64)**		**0.61 (0.55 to 0.67)**

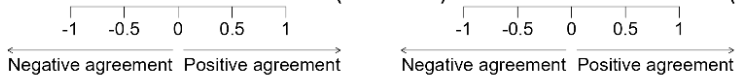
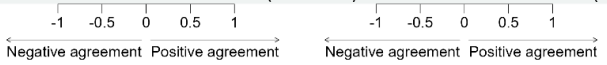


Table 4.16 - Agreement metrics between EPMA and case report form (CRF) for follow-up exposure, stratified by respiratory status at baseline

	Counts (participants or participant/drug pairs)					Agreement metrics							
Drug (by respiratory status)	**Total**	**Both sources**	**CRF only**	**EPMA only**	**Neither source**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Kappa**	**Kappa (95% CI)**	**PABAK**	**PABAK (95% CI)**
Corticosteroids													
No oxygen or ventilation	545	208	23	67	247	90	78.7	75.6	91.5		0.67 (0.61 to 0.73)		0.67 (0.6 to 0.73)
Oxygen requirement or non-invasive ventilatior	3533	2524	227	302	480	91.7	61.4	89.3	67.9		0.55 (0.51 to 0.59)		0.7 (0.68 to 0.72)
Mechanical ventilation or ECMO	235	70	58	17	90	54.7	84.1	80.5	60.8		0.38 (0.26 to 0.49)		0.36 (0.23 to 0.48)
Tocilizumab or sarilumab													
No oxygen or ventilation	519	3	2	1	513	60	99.8	75	99.6		0.66 (0.28 to 1.04)		0.99 (0.97 to 1)
Oxygen requirement or non-invasive ventilatior	3406	159	344	31	2872	31.6	98.9	83.7	89.3		0.41 (0.35 to 0.47)		0.78 (0.76 to 0.8)
Mechanical ventilation or ECMO	224	3	14	0	207	17.6	100	100	93.7		0.28 (-0.08 to 0.65)		0.88 (0.79 to 0.93)
Macrolides													
No oxygen or ventilation	545	61	21	28	435	74.4	94	68.5	95.4		0.66 (0.57 to 0.75)		0.82 (0.77 to 0.87)
Oxygen requirement or non-invasive ventilatior	3529	527	205	108	2689	72	96.1	83	92.9		0.72 (0.69 to 0.75)		0.82 (0.8 to 0.84)
Mechanical ventilation or ECMO	234	32	56	8	138	36.4	94.5	80	71.1		0.35 (0.21 to 0.48)		0.45 (0.33 to 0.57)
Remdesivir													
No oxygen or ventilation	382	15	1	4	362	93.8	98.9	78.9	99.7		0.85 (0.72 to 0.98)		0.97 (0.94 to 0.99)
Oxygen requirement or non-invasive ventilatior	3011	474	165	46	2326	74.2	98.1	91.2	93.4		0.78 (0.75 to 0.8)		0.86 (0.84 to 0.88)
Mechanical ventilation or ECMO	117	2	7	1	107	22.2	99.1	66.7	93.9		0.31 (-0.16 to 0.77)		0.86 (0.74 to 0.94)
Hydroxychloroquine													
No oxygen or ventilation	545	14	9	2	520	60.9	99.6	87.5	98.3		0.71 (0.54 to 0.88)		0.96 (0.93 to 0.98)
Oxygen requirement or non-invasive ventilatior	3533	85	8	13	3427	91.4	99.6	86.7	99.8		0.89 (0.84 to 0.94)		0.99 (0.98 to 0.99)
Mechanical ventilation or ECMO	235	15	7	0	213	68.2	100	100	96.8		0.8 (0.65 to 0.94)		0.94 (0.88 to 0.98)
Lopinavir-ritonavir													
No oxygen or ventilation	545	6	24	0	515	20	100	100	95.5		0.32 (0.06 to 0.59)		0.91 (0.87 to 0.94)
Oxygen requirement or non-invasive ventilatior	3533	16	64	0	3453	20	100	100	98.2		0.33 (0.17 to 0.49)		0.96 (0.95 to 0.97)
Colchicine													
No oxygen or ventilation	282	42	15	8	217	73.7	96.4	84	93.5		0.73 (0.63 to 0.84)		0.84 (0.76 to 0.9)
Oxygen requirement or non-invasive ventilatior	2484	445	65	22	1952	87.3	98.9	95.3	96.8		0.89 (0.87 to 0.91)		0.93 (0.91 to 0.94)
Mechanical ventilation or ECMO	97	11	10	1	75	52.4	98.7	91.7	88.2		0.6 (0.38 to 0.82)		0.77 (0.61 to 0.88)
Baricitinib													
No oxygen or ventilation	195	29	2	3	161	93.5	98.2	90.6	98.8		0.91 (0.82 to 0.99)		0.95 (0.88 to 0.98)
Oxygen requirement or non-invasive ventilatior	1729	339	24	19	1347	93.4	98.6	94.7	98.2		0.92 (0.9 to 0.95)		0.95 (0.93 to 0.96)
Mechanical ventilation or ECMO	66	1	7	0	58	12.5	100	100	89.2		0.2 (-0.36 to 0.76)		0.79 (0.59 to 0.91)
Aspirin													
No oxygen or ventilation	545	67	12	57	409	84.8	87.8	54	97.1		0.59 (0.5 to 0.68)		0.75 (0.69 to 0.8)
Oxygen requirement or non-invasive ventilatior	3533	687	64	275	2507	91.5	90.1	71.4	97.5		0.74 (0.71 to 0.77)		0.81 (0.79 to 0.83)
Mechanical ventilation or ECMO	235	15	16	16	188	48.4	92.2	48.4	92.2		0.41 (0.21 to 0.6)		0.73 (0.63 to 0.81)
REGN antibodies													
No oxygen or ventilation	328	30	4	6	288	88.2	98	83.3	98.6		0.84 (0.74 to 0.94)		0.94 (0.89 to 0.97)
Oxygen requirement or non-invasive ventilatior	2868	180	202	27	2459	47.1	98.9	87	92.4		0.57 (0.52 to 0.62)		0.84 (0.82 to 0.86)
Mechanical ventilation or ECMO	101	3	11	0	87	21.4	100	100	88.8		0.32 (-0.06 to 0.7)		0.78 (0.63 to 0.89)
Combined													
No oxygen or ventilation	**4431**	**475**	**113**	**176**	**3667**	**80.8**	**95.4**	**73**	**97**		**0.73 (0.7 to 0.76)**		**0.87 (0.85 to 0.88)**
Oxygen requirement or non-invasive ventila	**31159**	**5436**	**1368**	**843**	**23512**	**79.9**	**96.5**	**86.6**	**94.5**		**0.79 (0.78 to 0.79)**		**0.86 (0.85 to 0.86)**
Mechanical ventilation or ECMO	**1779**	**152**	**190**	**43**	**1394**	**44.4**	**97**	**77.9**	**88**		**0.5 (0.44 to 0.56)**		**0.74 (0.7 to 0.77)**



When considered over time, specificity for baseline and follow-up corticosteroid exposure decreased after the results of the RECOVERY trial became available, and similarly but to a lesser degree for exposure to tocilizumab/sarilumab during follow-up (Figure 4.7). Importantly, the RECOVERY comparisons for corticosteroids and tocilizumab both yielded positive results.[72,278] For macrolides, aspirin, and colchicine (drugs with negative results in the RECOVERY trial),[279–281] specificity increased after the results became available for both baseline and follow-up exposure, with LMWH exhibiting a similar pattern. For baricitinib and REGN antibodies (drugs with positive trial results, but a lower uptake into clinical practice than corticosteroids and tocilizumab/sarilumab),[80,282] specificity remained mostly unchanged following publication of the trial results.

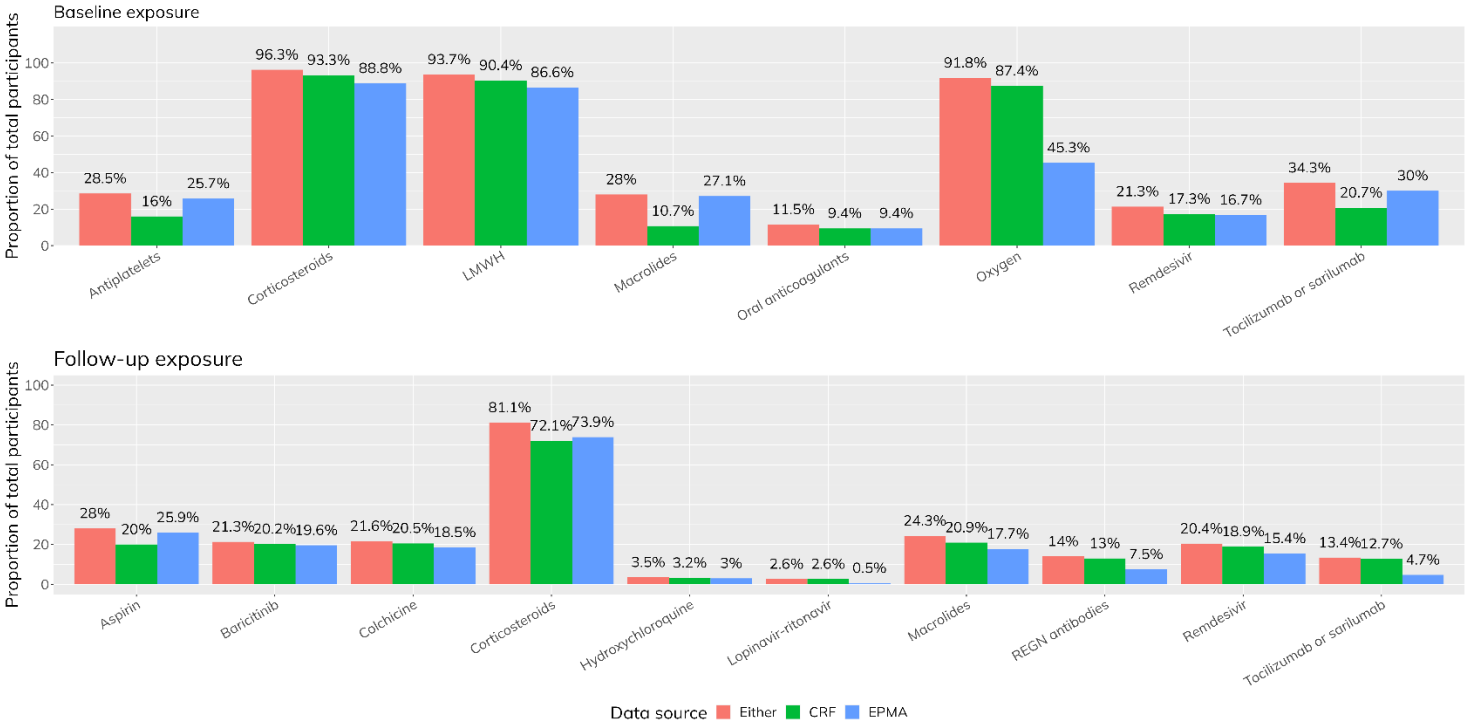
Figure 4.7 - Sensitivity and specificity over time for binary baseline and follow-up drug exposure assessments, stratified by drug group



Figure legend: Green boxes depict the approximate time period during which each drug was being studied in RECOVERY. Error bars depict 95% CIs.

If used as a single source of drug exposure, EPMA data would capture more participants exposed at baseline to antiplatelets (25.7% EPMA vs 15% CRF), macrolides (27.1% vs 10.7%), and tocilizumab/sarilumab (30% vs 20.7%), fewer participants exposed to oxygen, and approximately the same for corticosteroids (88.8% vs 93.3%), oral anticoagulants (9.4% for both), and remdesivir (16.7% vs 17.3%) - Figure 4.8. For follow-up exposure, EPMA data would generally capture fewer or similar numbers of participants than the CRF, except for corticosteroids (73.9% EPMA vs 72.1% CRF) and aspirin (25.9% EPMA vs 20% CRF).

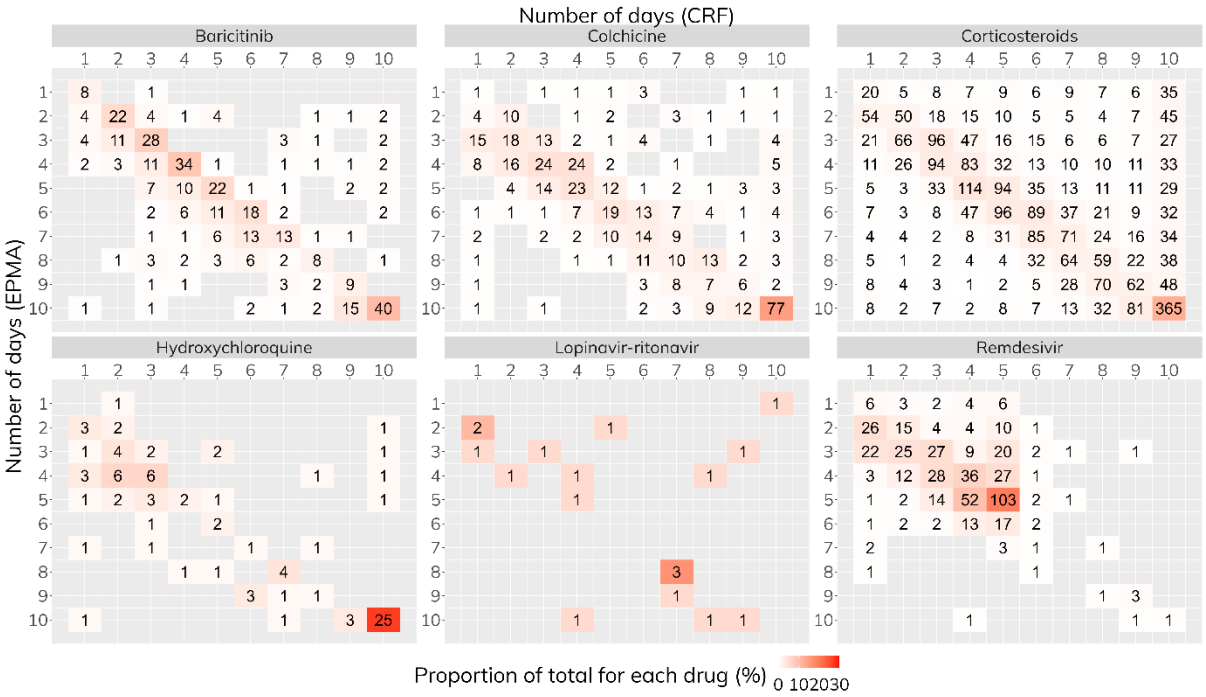
Figure 4.8 - Number of participants identified as baseline or follow-up exposure according to EPMA, case report form (CRF), or either data source



4.4.3.2 Quantitative follow-up exposure

Agreement between EPMA and CRF data for number of days of follow-up drug exposure was moderate (ICC: 0.64, 95% CI: 0.62 to 0.65), with ICCs ranging from 0.40 for lopinavir-ritonavir to 0.76 for baricitinib (Figure 4.9). Average differences between sources (CRF minus EPMA) for exposure during follow-up were small, ranging from -0.8 days [95% LOA: -5.2 to 3.7] for hydroxychloroquine to 0.4 [-4.8 to 5.3] for corticosteroids (overall: 0.1 [-4.6 to 4.8]).

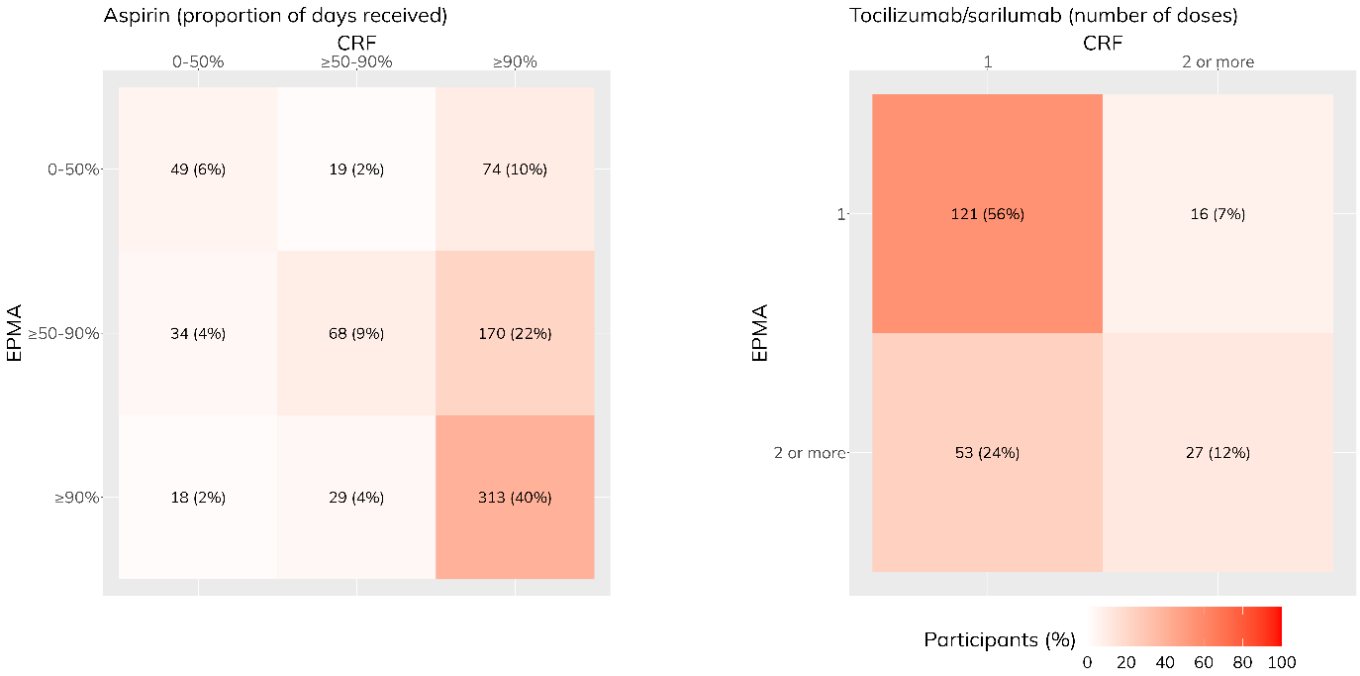
Figure 4.9 - Number of days of follow-up drug exposure according to EPMA vs case report form (CRF) data



Agreement for proportion of days drug received (aspirin) or number of doses received (tocilizumab/sarilumab) was modest (aspirin – κ: 0.25; PABAK: 0.11; tocilizumab/sarilumab - κ: 0.24, PABAK: 0.37; Figure 4.10). Disagreement was mostly due to EPMA underestimating aspirin exposure (33% of participants had lower exposure in EPMA vs CRF, 10% higher, and 55% same), and overestimating tocilizumab/sarilumab exposure (24% of participants had higher exposure in EPMA vs CRF, 7% lower, and 68%

same). Manual data exploration for aspirin records excluded the presence of prescription/administration records not being correctly categorised by the algorithm used.

Figure 4.10 - Proportion of days (for aspirin) or number of doses (for tocilizumab/sarilumab) received during follow-up according to EPMA vs case report form (CRF) data



4.4.4 Use-case scenarios

Figure 4.11 depicts the first use-case scenario (monitoring of CRF drug recording by study site to identify potential data recording issues). κ coefficients by site ranged from 0.37 (95% CI: 0.30-0.44) to 0.74 (0.70-0.78). Investigation of drug recording patterns by site (using proportion of missing records and unconfirmed records) identified distinct potential reasons for disagreement between CRF and EPMA data when compared with aggregate averages. These included: high proportion of missing and unconfirmed records (sites 2, 4, 9, 6); high proportion of unconfirmed records but low proportion of missing records (site 11); high proportion of missing records but low proportion of unconfirmed records (site 1).

Figure 4.11 - Use-case scenario 1: Monitoring of case report form (CRF) drug recording across study sites (using EPMA as reference)

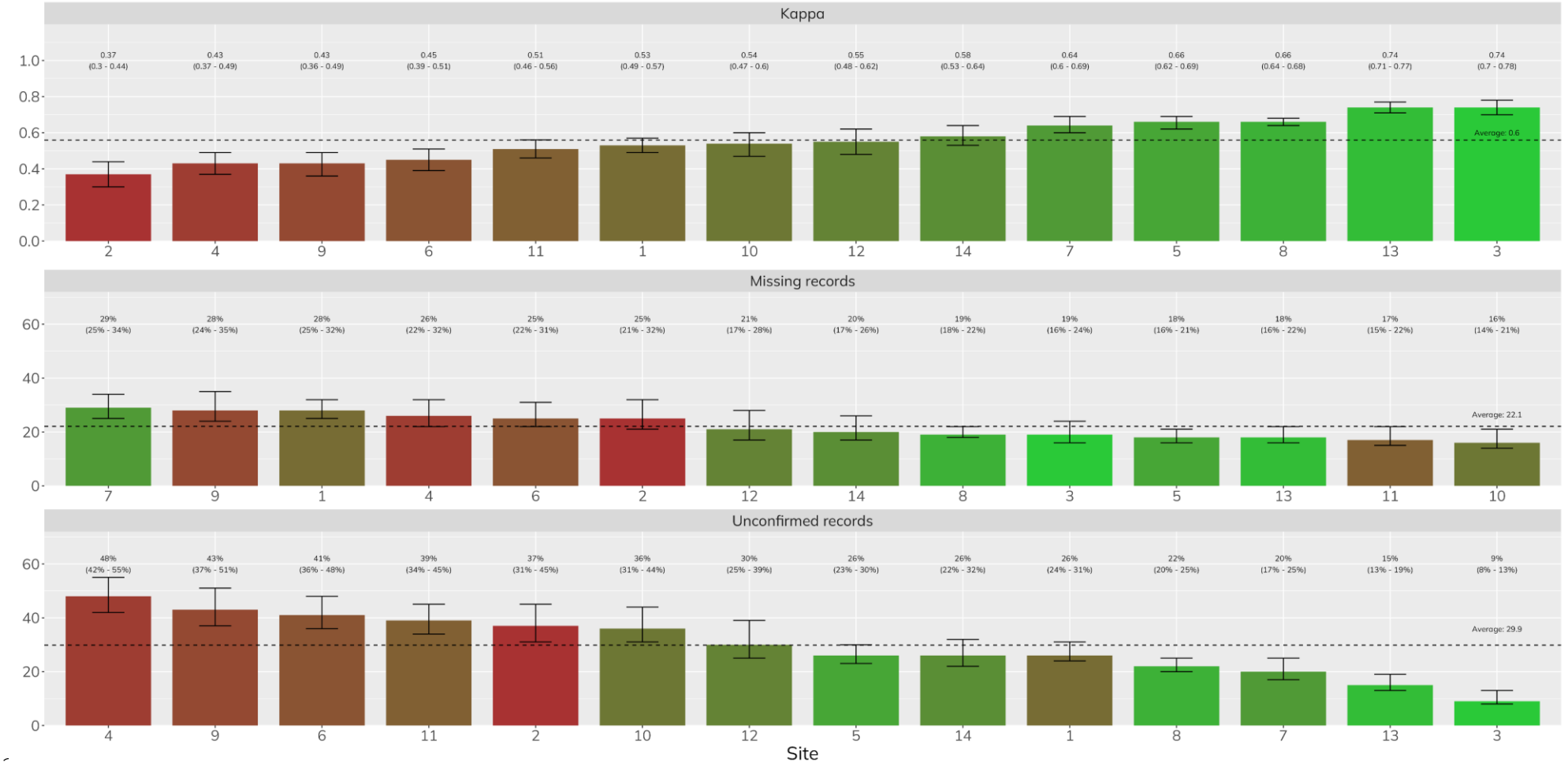


Figure legend: The vertical axes represent the units of the statistic outlined in each panel (Cohen's Kappa, proportion of missing records, or proportion of unconfirmed records). Missing records: proportion of events identified by EPMA data that are not confirmed by CRF data. Unconfirmed records: proportion of events identified by CRF data that are not confirmed by CRF data. The text labels show medians and 95% CIs for each site. Error bars depict 95% CIs. The mean for each metric across all sites is shown at the far right.

The use-case scenario of characterising standard care patterns within the trial over time is shown in Figure 4.12. EPMA data was able to reliably track changes to treatment patterns during and after the trial period for all drugs (e.g. similar use of corticosteroids and hydroxychloroquine while being studied in the trial), and subsequently expanded use of corticosteroids and tocilizumab/sarilumab and decreased use of hydroxychloroquine (following the publication of the RECOVERY trial results). However, for tocilizumab/sarilumab the proportions calculated using EPMA data were consistently lower than the CRF. When stratified by baseline respiratory support, EPMA data consistently captured less participants exposed for people on IMV/ECMO across all drugs, as well as for tocilizumab for participants on oxygen/NIV. Conversely, EPMA data captured more participants exposed to steroids in the no oxygen/NIV group. Of note, adoption of baricitinib into routine practice after positive results appears to have been much lower than that of corticosteroids and tocilizumab/sarilumab.

Figure 4.12 - Use-case scenario 2: Use of select RECOVERY trial medications during the admission of randomisation over time according to EPMA vs case report form (CRF) data, by level of respiratory support at randomisation



Figure legend: Green shades show the period during which each drug was being studied in RECOVERY, with vertical dashed lines showing the date when each result was published in the RECOVERY website (www.recoverytrial.net). Error bars depict 95% CIs.

4.5 Discussion

In this chapter, implementation of RCD on inpatient medications in England was developed using EPMA data within the RECOVERY trial, and validated against CRF data. This is the first reported use of EPMA data for research, and the first validation study of RCD on inpatient medications within a randomised trial. EPMA data covered a representative sample of approximately 11% of RECOVERY participants recruited in England. Agreement between EPMA and CRF was moderate to high, with important variations by drug, respiratory status at baseline, and analysis period (baseline vs follow-up), as well as over time. EPMA data could overestimate numbers of participants exposed to some drugs at baseline vs CRF data, while mostly underestimating exposure during follow-up, but generated similar overall drug exposure patterns over time. These results suggest that EPMA data may be considered a reliable source of overall drug exposure estimates when compared with CRF data. However, limited coverage and poor agreement in specific settings limits its use for assessment on an individual basis.

Contrary to datasets considered in previous chapters, EPMA data only covered approximately 11% of the RECOVERY trial cohort recruited in England. Participants with available EPMA data had similar baseline characteristics to those without, and were recruited in locations spread throughout England, providing a representative sample of the trial population. Although EPMA data may not, at present, provide adequate coverage for use as a single source of drug exposure, this suggests it may be used as an appropriate source of inpatient medications data for research into nationwide care patterns.

EPMA data stability was very high (similarly to the results found for Dispensing data in Chapter 3). Timeliness was also very good, with approximately real-time data reporting and dissemination, in contrast to English RCD resources such as Hospital Episode Statistics (with finalised data released yearly),[148] or US cancer registries (where

dissemination targets are in the order of months).[247] However, a two month period should be allowed before analysis so that data can become stable (due to less frequent data submission schedules for some contributing hospitals).

In the absence of established guidance on how to use these data, a data-driven methodology development process was designed and applied to identify the best procedure to derive drug exposure, using CRF data as reference. While some derivation choices did not have a substantial impact, others could greatly influence the resulting agreement with CRF (namely restricting to active prescriptions). Some of these results could be counterintuitive, highlighting the importance of generating appropriate, in-depth knowledge of a particular RCD source before using it for research purposes, and using adequate validation procedures.[283]

Overall agreement between EPMA and CRF data was substantial for both baseline and follow-up exposures, and similar to levels considered adequate in other validation studies of RCD on medications within trials.[273,284,285] Agreement metrics were generally poorer for comparisons of baseline versus follow-up exposure, possibly as a result of larger impacts of even small discrepancies for a single drug record within the short observation window between hospitalisation and randomisation. For most drugs, sensitivity (i.e. EPMA data not capturing CRF records) was the main limitation. This was likely due to partial uptake of electronic prescribing systems within contributing hospitals (namely suboptimal ICU data coverage), as suggested by generally poorer agreement and sensitivity in participants on IMV/ECMO at randomisation, and acknowledged in the available EPMA documentation[158]. This limitation stresses the importance of considering adequacy of a particular dataset for a given research purpose, and how data recording and collection patterns may differ for particular subgroups.[123] Nonetheless, people on IMV/ECMO represent a minority of RECOVERY trial participants included in this analysis (5.4%), so that aggregate NPVs were >93%, thus suggesting that EPMA data are extremely accurate

at excluding drug exposure. On the other hand, true drug administration without a corresponding record may also occur, even when electronic management systems are used.[286] However, it is unlikely that this behaviour would be captured by study staff completing CRF data.

In contrast to other drugs, specificity (i.e. EPMA records not confirmed in the CRF) was the main limitation for corticosteroids (and similarly, but to a lesser degree, for LMWH exposure at baseline). For LMWH, this finding results from a small number of participants (n=97) wrongly captured as exposed by EPMA data for a drug with very high use prevalence, possibly due to omissions in scheduled medication administration occurring in the short period between hospital admission and randomisation. Dexamethasone was the first drug effective for in-hospital COVID-19 treatment (with benefit only for people with some degree of oxygen or ventilatory support), since becoming the mainstay of in-hospital treatment.[72,287] In this analysis, specificity for corticosteroid exposure dropped substantially with increasing respiratory support levels for baseline exposure (similarly to, but to a much higher degree than other drugs used in COVID-19 treatment, such as macrolides, tocilizumab/sarilumab, and remdesivir), but increased for follow-up exposure (similarly to other drugs). Accordingly, specificity for follow-up exposure decreased after the dexamethasone result was announced (June 2020), and contrary to drugs found to be ineffective in RECOVERY or other trials, for which specificity increased after the trial results were published (e.g. macrolides, aspirin, colchicine, hydroxychloroquine, lopinavir-ritonavir, as well as therapeutic dose LMWH from mid-2021 onwards).[288,289] These findings suggest that recording of corticosteroids in either EPMA or the CRF, as well as other drugs, may have been influenced by changing treatment recommendations over time. One possible explanation is the inclusion of corticosteroids in prescription packages automatically generated for COVID-19 patients upon admission. If these prescriptions were later not administered (e.g. suspected COVID-19 which was subsequently excluded;

clinical improvement after the initial prescription but before CRF completion), this could generate EPMA prescription records which are no longer active, but not necessarily erased from the underlying database. A similar pattern occurred with aspirin and macrolides at baseline, where EPMA greatly overestimated exposure (but not during follow-up). Accordingly, restricting EPMA records to "active" prescriptions increased baseline steroid recording specificity from 55% to 74%, although reducing sensitivity from 92% to 50% (with similar results for remdesivir and macrolides). Alternatively, true medication administration omissions were possible, and are common in UK hospitals[290] and elsewhere.[286,291] In any of these scenarios, one would expect a prescription record without corresponding administration records. These phenomena might contribute to the lower aggregate PPVs (baseline: 81.3%; follow-up: 85.1%) when compared with NPVs (baseline: 92.9%; follow-up: 94.5%), which suggest that RCD sources such as EPMA may be more accurate at excluding drug exposure (i.e. it is unlikely that drug exposure occurs without a corresponding electronic record, except where data coverage is suboptimal such as in ICU patients) than confirming it (a prescribed drug may not be administered for many reasons, but the corresponding record kept). Also, the randomisation CRF question on corticosteroids asked whether these were "currently prescribed", rather than administered, and might have been led to different interpretations by study staff. Nevertheless, restricting to administration records only did not materially impact baseline steroid specificity (56.9% vs 55.2%), neither did the use of more restrictive administration dates (59% administered vs 58% scheduled date), suggesting that there were indeed valid administration records for these drugs.

Taken together, the observations described above put into question whether CRF data can be considered a true gold-standard. CRFs were recorded via manual medical record abstraction by busy research staff, who may not have been directly involved in each participant's care, and thus errors are possible. Albeit not directly comparable, an internal

RECOVERY review showed that CRF data had a sensitivity of 93% for steroids when compared with manual review of local prescription records, while NPV was only 53% (data not shown). This further suggests that absence of CRF recorded steroid exposure may not necessarily mean true non exposure. Accordingly, discrepant records contributing to low corticosteroid specificity (i.e. false-positives, 102/3,333 = 3% for baseline and 386/4313= 8.9% for follow-up exposure) represent only a small proportion of total records, and could potentially be accounted for by a combination of human error and incorrect EPMA data. These findings further confirm the previously-held notion that there is no single gold-standard for drug exposure assessments, and that triangulation between sources is key for precise estimates.[95,116,192,292] Nevertheless, aggregate EPMA data accuracy (i.e. PPV) was 72% for baseline and 85% for follow-up drug exposure. Together with the generally small differences found for total drug exposure estimates at baseline and follow-up, this suggests EPMA data can be used to provide accurate overall estimates. Moreover, it opens the possibility of using EPMA data to investigate CRF recording patterns during the trial period, including whether they are being interpreted and completed as intended. Nonetheless, the importance of the discrepancies found between the two data sources will depend on the specific aim of using these data. For example, some discrepancies might be acceptable if the purpose is to describe overall exposure at randomisation or adherence during follow-up, whereas this threshold might be lower if treatment effects are modified substantially based on the concomitant use of another drug (or if a combination of two drugs is contra-indicated).

Estimates of quantitative exposure showed moderate overall agreement, with small absolute differences. For corticosteroids and aspirin, there were large numbers of people with maximum exposure (10 days) according to the CRF, but much lower exposure according to EPMA data. These discrepancies may either result from the known EPMA data coverage issues, or overestimation by staff completing the CRF (by assuming

continuous administration during admission, or making approximations for proportion of days covered). As discussed previously, it is not possible to confirm which was the case using the available data. On the other hand, EPMA overestimated the number of tocilizumab/sarilumab doses received after randomisation, which might have resulted from planned administrations that did not occur. While these results show less than ideal agreement at an individual level, estimates of quantitative exposure are expected to have a smaller impact on study analyses as binary exposure estimates.

Finally, the use-case scenarios performed demonstrate the feasibility of EPMA data as a reasonable surrogate for aggregate estimates of in-hospital drug-exposure, including changes over time. However, discrepancies between sources become evident when considering respiratory status, and particularly for drugs more likely to be used in ICU settings (e.g. tocilizumab). Thus, it is possible that overall agreement might have been overestimated due to underlying patterns of low sensitivity and low specificity compensating each other (i.e. discrepancies in opposite directions cancelling each other out). The feasibility of using EPMA data for CRF drug recording auditing was also tested with a practical example, showing it could expose marked differences in drug recording patterns between study sites. However, as discussed above, it is impossible to know if these differences represent true errors in CRF recording, or rather variations in EPMA data collection in each hospital. The existing EPMA documentation reports significant variability across contributing hospitals regarding quantity of prescriptions handled via the electronic systems, suggesting the latter may be the case.[158] Nonetheless, the purpose of this assessment would be to highlight *potential* issues with data recording (based on statistical distributions, and in conjunction with other data to determine risk levels), rather than adjudicate *definite* issues. Assessment of overall patterns could therefore be sufficient as a screening tool for this purpose, even if imperfect.

4.5.1 Limitations

A number of limitations should be noted. First, this analysis only included approximately 11% of the RECOVERY cohort recruited in England as a result of EPMA data availability at the present moment. While overall agreement was high, the limited availability hinders implementation of EPMA data as a primary source of drug exposure from the onset. Nonetheless, EPMA data collection is currently expanding to more hospitals across England.[293] If this dataset is found to be of appropriate quality for use in trials, its value as a complementary or primary source of medications data will increase as coverage grows (with the latter potentially incentivised by such assessments and the subsequent increase in demand).

CRF data were used as a reference to guide development and implementation of EPMA data here. While CRF may be considered the *de facto* standard for drug exposure assessment within trials, it is unlikely to be a *gold-standard* for actual medication intake (in which case circulating drug levels could be used).[286] Therefore, discrepancies between CRF and EPMA could in fact be due to wrong CRF records. Moreover, other trials may need to record drug exposure using different metrics or level of detail to that considered in the RECOVERY CRF, further exemplifying that the definition of a reference standard is contextual.

On the other hand, discrepancies could be due to suboptimal EPMA data quality, including resulting from data submission pipelines. To the best of the candidate's knowledge, there is no published information on intrinsic EPMA dataset quality (i.e. comparison of data provided with records in-situ at source, or other methods of internal validation). Such granular level of validation was also outside the scope of this thesis, as it is assumed that publicly-available datasets would have undergone some level of internal quality assurance (as seen for other RCD sources assessed in this thesis). However, the use of computerised data extraction systems is standard practice worldwide, and is expected to

lead to fewer errors than manual abstraction.[294] In Denmark, similar data extraction processes in an inpatient database have yielded 99% accuracy (i.e. PPV) vs local clinician records,[295] but these estimates cannot be directly extrapolated to other settings. Thus, the discrepancies observed may be due to errors in either source considered. Exploratory analyses were developed to further investigate these, but the lack of source medical record data hinders definitive conclusions.

Agreement between sources was assessed with a commonly used statistical approach (κ coefficients), which facilitates comparison with similar studies. Nonetheless, κ coefficients have important limitations (e.g. the possibility of low coefficients despite high agreement in particular circumstances).[263] Moreover, κ was developed for a different purpose than the one used here (i.e. assessing agreement between two observers for subjective ratings, rather than two sources of data for an objective event). To address these limitations, κ was used in combination with PABAK, and sensitivity, specificity, and predictive values. Therefore, although overall agreement between sources in this analysis was considered high, it should be noted that κ coefficients are imperfect estimates of agreement, and that their interpretation is challenging. As noted in previous studies and relevant MHRA and FDA guidance, data quality is context-specific and judgement of whether RCD are appropriate for use in a given setting should be individualised.[24,36,192]

Finally, the applicability of these findings is likely to be limited to studies intending to use EPMA data for similar purposes and in similar care settings. As an example, a trial mainly focused on critical care (or on treatments such as intravenous infusions or blood products) would possibly find that EPMA data do not adequately capture the intended drug exposure in their setting. Nonetheless, other studies interested in collection of similar drug exposure phenotypes and similar drugs should be able to use these results to justify and guide the use of EPMA data, without the need for separate validation or development.

4.6 Conclusion

The analyses presented here demonstrate that EPMA data cover a representative sample of approximately 11% of all RECOVERY trial participants recruited in England, and are able to capture a variety of drug exposures with adequate data quality and timeliness. While overall drug exposure estimates were similar to reference CRF data, more granular analyses exposed potential issues with data coverage and recording for particular patient groups and drug types. In their current form, EPMA data cannot be used to replace CRF recording. However, this may change in the future if coverage is improved, together with a better understanding of intrinsic data quality features and detailed end-to-end data recording audits. These analyses highlight the limitations of conducting research using a resource which was not specifically compiled for this purpose. Nonetheless, the degree of overall agreement found supports the use of these data as an epidemiological research tool.

CHAPTER 5: Routinely-collected nationwide Scottish medications data - development, validation, and application in the LENS trial of fenofibrate in diabetic retinopathy

5.1 Chapter summary

Aims: To develop and validate the use of RCD on medication use in the community in Scotland within the LENS trial of fenofibrate for prevention of diabetic retinopathy progression.

Methods: In LENS, non-study medications are recorded in CRFs at randomisation (all current medications at that point) and during follow-up (initiation of prohibited medications only). Data on medications dispensed in the community in Scotland is also collected from the Prescribing Information Services (PIS) dataset, and available from consent at screening onwards (typically 2 months before randomisation). To validate the use of PIS data within LENS, data quality was first evaluated by verifying data coverage and timeliness. A data-driven methodology development process was then performed by testing different algorithms to define drug exposure at randomisation (using the randomisation CRF as reference). First, different observation windows were tested to define baseline drug exposure at randomisation. Then, approaches for handling missing exact record dates were tested. Accuracy of statin doses extracted from PIS data were also tested against CRF records. Using the best performing algorithm, agreement between PIS and CRF data was then calculated across 18 drugs of interest. Finally, the impact of using PIS data was assessed by calculating drug exposure prevalence (i.e. proportion of participants exposed to each drug) at baseline and during follow-up using PIS vs CRF.

Agreement was assessed using Cohen's κ coefficients, sensitivity, specificity, and predictive values.

Results: PIS data captured >99% (1,147/1,151) of LENS participants, with a 2-3 month lag between the most recent data items included in the dataset and its extraction. Most drugs of interest were dispensed on average every 60 days. Including records in the month of randomisation provided very high agreement with CRF data for baseline drug exposure across all possible lookback windows (κ=0.90-0.92), with a marginal increase if also including records in the month after randomisation (κ=0.92-0.93). Imputing day 1 of the month as exact record date (vs day 15 or last day) generated best agreement between PIS and CRF for baseline exposure (κ=0.92, 0.89, and 0.83 respectively). Statin doses recorded in the randomisation CRF agreed with the strength of the most recent statin record in PIS data for 99% of participants. Using an observation window including 3-months pre-randomisation and the month of randomisation, agreement between PIS and CRF for exposure at randomisation was high (κ>0.80) across most drugs studied (except ezetimibe: κ=0.71). Sensitivity, specificity, and predictive values were also high (approximately >90% for most drugs). PIS and CRF data provide similar counts of participants exposed to each drug at baseline, and for initiation of prohibited drugs during follow-up. However, PIS data complemented CRF data by capturing changes occurring post-randomisation to non-prohibited drugs recorded in the randomisation CRF (mostly initiation of new medications during follow-up).

Conclusions: PIS data are a highly complete and accurate source of information on drug exposure in the community in a randomised trial of fenofibrate in patients with diabetes in Scotland. These results support the use of PIS data across more randomised trials.

5.2 Introduction

In Scotland, RCD have long been used for research purposes in general, and RCTs in particular.[296,297] For example, in 2007 the West of Scotland Coronary Prevention Study of pravastatin for primary cardiovascular prevention in men (WOSCOPS) published its long-term results, using data exclusively from nationwide computerised healthcare records.[165,298] From 2013-2018, ISD (the National Health Service body responsible for publication of official healthcare statistics in Scotland) provided multiple types of linked clinical data to 35 RCTs.[61,299]

RCD on medications prescribed or dispensed in the community and submitted for reimbursement to NHS Scotland have been collated by ISD since 1993 at an aggregate level in the PIS dataset.[162] From 2009, this collection also includes details of individual prescribing and dispensing events. PIS data are available for research purposes, and were used in 40 peer-reviewed publications between 2009-2019.[163] These data were considered high-quality, with publicly-available information on a number of internal data quality assurance processes[300] and a capture rate (i.e. proportion of total prescriptions captured) of approximately 100%.[162,163] In 2017, Hafferty et al published results of a validation study on self-reported medication use versus PIS data among 10,244 participants of the Generation Scotland population-based cohort.[164] Of note, PIS data was used as a reference to validate self-reported drug exposure (rather than self-report validating PIS data). This study was focused on a small number of psychiatric (antidepressants, mood-stabilisers), cardiovascular (anti-hypertensives, lipid-lowering, aspirin, and insulin) and gynaecological drugs (hormone-replacement therapy, oral contraceptives). Agreement was high (i.e. $\kappa>0.70$) for most medications studied, except mood stabilisers ($\kappa=0.45$).

Nonetheless, use of PIS data within RCTs has been limited to two pragmatic trials of feedback interventions to improve prescribing safety. This were the Effective Feedback to

Improve Primary Care Prescribing Safety (EFIPPS) study, assessing cardiovascular, antipsychotic, and anti-inflammatory drugs (published in 2016), and the Feedback of Antibiotic Prescribing in Primary Care (FAPPC) study (published in 2022).[166,167] The lack of widespread adoption of PIS data within trials (in contrast to other Scottish data sources) may be justified by challenges associated with the use of RCD sources, such as uncertainty as to whether PIS data are of adequate quality for use within RCTs, particularly outside pragmatic studies (e.g. in drug intervention trials).[32]

The aims of this chapter are to: 1) assess whether PIS data are of adequate quality to derive outpatient drug exposure use within a randomised trial; 2) develop the methodology needed to implement PIS data for this purpose; and 3) test the impact of using PIS data in possible use-case scenarios within a randomised trial.

5.3 Methods

5.3.1 Study population and time period

LENS is a randomised controlled trial of fenofibrate versus placebo for prevention of retinopathy disease progression in diabetic patients. The trial has recruited 1,151 participants in 16 study sites across all 11 mainland health boards in Scotland, from September 2018 to July 2021.[301] All LENS participants have provided written informed consent, including for linkage to RCD sources. This analysis included all randomised LENS participants.

5.3.2 Medication data sources

5.3.2.1 Case report forms

LENS participants attended a screening visit and a randomisation visit in person with a study nurse. Subsequent follow-up was conducted by post or telephone. Use of prohibited medications (i.e. colchicine, cyclosporine, daptomycin, fibrates, ketoprofen, warfarin or other vitamin K antagonists, and rosuvastatin at a dose of 40mg) was recorded at the screening visit, and individuals taking any of these drugs excluded from the study. Those eligible for the trial at the screening visit entered a run-in phase of approximately 8 weeks, and then attended a randomisation visit at which baseline medical history was recorded. All current medications were also recorded by the study nurse at randomisation, based on participant self-reports and a review of the medication list provided by each participant. Daily dose was also collected for statins. Finally, initiation of prohibited medications was being recorded at each follow-up assessment (with no further data collected on non-prohibited medications).

5.3.2.2 PIS data

Nationwide RCD on outpatient medications data were collected from the PIS dataset.[162] PIS gathers individual prescribing and dispensing data, including unique Community

Health Index (CHI) numbers (equivalent to NHS numbers in Scotland). The data cover all NHS medications prescribed, dispensed, and reimbursed in the community setting in Scotland. Over-the-counter medications and those dispensed in-hospital are not captured. Linkage to PIS data was performed using CHI numbers, with data extracts being provided to LENS approximately every two years. Written informed consent was received for collection of PIS data from the date of screening onwards. Only some cardiovascular medications of specific interest were selected for extraction from PIS data for this study, based on the BNF terminology (Table 5.1).[270]

Table 5.1 - Medications included in the LENS PIS data extraction

BNF chapter	Term
2	Cardiovascular Drugs
2.2	Diuretics
2.4	Beta-adrenoceptor blocking drugs
2.5	Antihypertensive drugs
2.5.5	Drugs affecting the renin-angiotensin system
2.6.2	Calcium-channel blockers
2.8.2	Oral anticoagulants
2.12	Lipid-regulating drugs
6	Endocrine system
6.1.2	Diabetes
6.1.6	Diabetes mellitus, diagnostic and monitoring devices

The first PIS data extract was received in March 2022 and used in this analysis. Only information on dispensing events was requested. Although PIS data includes exact dates of prescribing and dispensing (which are available for dissemination to research teams), these were not requested for the LENS trial, with dates provided as month and year only (NB: data sharing for LENS was agreed prior to the start of this DPhil thesis). Each record provides an approved medication name (usually International Nonproprietary Names) and BNF code (i.e. including formulation and dose). Other variables include dispensing location and number of items dispensed (Table 5.2).

Table 5.2 - PIS data structure

Field name	Description
participant_id	Unique patient ID
dispenser_location	Dispenser location type
disp_ym	Date (month and year) of dispensing
bnf_code	BNF item code
approved_name	Approved medication name
number_items	Number of items

The drug groups selected for this analysis are detailed in Table 5.3. These include 18 non-prohibited drugs of interest to the LENS trial, and 8 prohibited medications (as recorded in the screening and follow-up CRFs).

Table 5.3 - Drugs select for analysis

General drugs of interest to LENS	Prohibited drugs
Antidiabetic drugs	Cyclosporine
Insulin	Colchicine
Metformin	Daptomycin
Sulphonylureas	Fibrates
Thiazolidinediones	Ketoprofen
Dipeptydil peptidase-4 (DPP-4) inhibitors	Rosuvastatin
Glucagon-like peptide 1 (GLP-1) agonists	Vitamin K antagonists
Sodium glucose transporter 2 (SGLT2) inhibitors	Warfarin
Acarbose or similar	Other anti-vitamin K antagonists
Antidyslipidemics	
Statins	
Ezetimibe	
Proprotein convertase subtilisin/kexin 9 (PCSK9) inhibitors	
Antihypertensives	
Angiotensin-converting enzyme (ACE) inhibitors	
Angiotensin receptor blockers (ARB)	
Mineralocorticoid receptor antagonists	
Beta-adrenergic blockers	
Calcium channel blockers (CCB)	
Loop diuretics	
Thiazide diuretics	

For each drug group, a codelist was built using BNF codes via interrogation of the BNF hierarchy within the OpenCodelists platform.[240] Where appropriate, the lists obtained were merged with other publicly-available codelists stored in OpenCodelists.[240] Each

codelist was built by one clinician (the candidate), and manually-reviewed and signed-off by a second clinician. All drug codelists used here are available for inspection and reuse at https://github.com/gpessoaamorim/dphil-lens.

5.3.3 Data analysis

PIS data were assessed to verify their adequacy for the specific aim of identifying outpatient drug exposure to concomitant (i.e. non-study) medications in LENS.[33,244] Data coverage (capture of the intended population and exposure) and timeliness (temporal lag between data recording and dissemination) were evaluated first.[246,247] Then, different algorithms to derive drug exposure from PIS data were developed and tested against CRF data. Using the best performing algorithm, PIS data accuracy was then assessed by evaluating agreement between PIS and CRF data by drug group.[123,192,245–247] Finally, the impact of using PIS data in the trial was tested with possible use-case scenarios. Similarly to previous chapters, data quality was considered in a contextual/external vs internal dimension,[192,242] as it is assumed datasets published by ISD Scotland have undergone rigorous internal data quality checks (the details and results of which are publicly available).[162,163,300]

5.3.3.1 Data quality - coverage and timeliness

Data coverage was initially assessed by calculating the proportion of LENS participants with available PIS data, and computing baseline characteristics for all LENS participants and separately for those with PIS data available vs those without. Coverage of drug groups of interest was evaluated by assessing how many were captured in PIS data. Timeliness was assessed by plotting record counts over time, and assessing the relationship between the most recent data items included in the PIS data extract and the date it was received.

5.3.3.2 Drug exposure methodology development

Previous studies using PIS data have used two different approaches to identify drug exposure, specifically 1) the fixed-time window and 2) the legend-time/medication-on-hand methods. The derivation approaches used by each method have been described in previous chapters.[164,302–304] In contrast to longitudinal drug exposure (i.e. drug adherence) assessments, where the legend-time/medication-on-hand approach is preferred, there is currently no consensus as to the optimal analytical approach for cross-sectional drug exposure (i.e. at a particular time point, such as randomisation), using either PIS or other data sources.[182] The legend-time/medication-on-hand method could not be used here given the lack of exact dates, posologies, or dispensed quantities. Moreover, this method has been shown to be on average inferior to the fixed-window method.[258] The latter was also the method used by Hafferty et al in their study comparing PIS data versus self-reported medication.[164]

In the absence of specific guidance on identifying drug exposure from PIS data, different rules/algorithms were developed for use in the LENS trial. Candidate rules were identified from preliminary data exploration (not shown). Each rule was tested by assessing the impact on resulting agreement observed between PIS and CRF data (used as reference). Specifically, the following options were tested:

1) Which observation window should be used for baseline exposure assessments (when should it start, and when should it end);
2) How should missing exact record dates be handled (for drug initiation events); and
3) Can drug dose be extracted reliably from PIS data (using BNF drug codes).

5.3.3.2.1 Methodology development – lookback windows and handling peri-randomisation records

Two parameters were assessed to define an appropriate observation window for drug exposure at randomisation: the lookback window (i.e. when the observation window starts), and handling of peri-randomisation records (i.e. when the observation window ends).

First, the suitability of different possible fixed lookback windows was explored by calculating average time intervals between consecutive records for the same BNF drug code and each participant, aggregated by drug groups (and, as no exact dates are provided, imputing day 1 as the specific date for each record). This followed a preliminary assessment exploring the number of distinct BNF drug codes and distinct records for each drug group, per participant and per month (data not shown). This exercise showed that most participants had on average only one record per distinct drug code per month, and one record per drug group per month. Then, for each drug group, the proportion of participants who had average intervals between consecutive dispensing records of the same drug group within 1, 2, 3, 6, and 12 months was calculated. Next, the agreement between PIS and CRF data for drug exposure at randomisation was calculated using different lookback windows (i.e. counting records within 1, 2, or 3 months before randomisation, or at any moment before randomisation).

Testing of the different lookback periods specified above was then combined with potential approaches to handling records occurring in the month of randomisation (i.e. include or exclude from baseline exposure), as these could not be reliably allocated as occurring before or after randomisation in the absence of exact dates. Given the availability of a short period of data pre-randomisation (i.e. from screening to randomisation only), the inclusion of records in the month after randomisation as baseline exposure was also tested as a way to increase the amount of data available for this

assessment.[305] In total, 12 different rules for baseline exposure derivation were assessed, evaluating the combined impact of increasing the observation period into the past (rules 1-4: 1 month, 2 months, 3 months, and any time before randomisation) and into the future (rules A-C: exclude month of randomisation; include month of randomisation; and include month of randomisation and month after) - Figure 5.1.

Figure 5.1 - Baseline drug exposure derivation rule logic

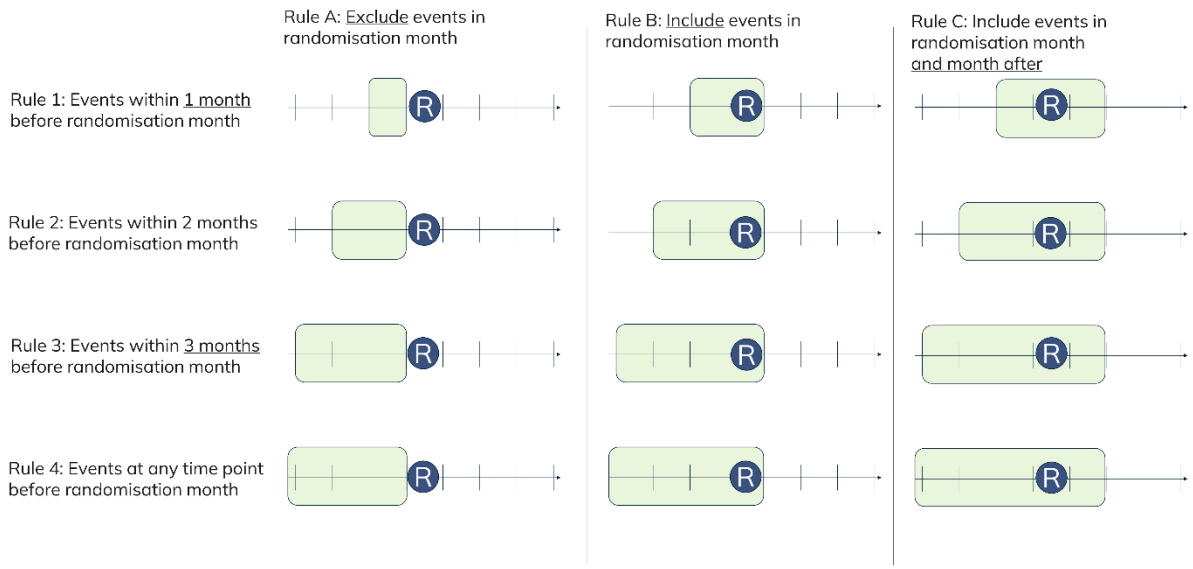


Figure legend: Time periods highlighted in green depict the time period during which PIS records would count towards definition of baseline exposure. The run-in period (from screening to randomisation) typically lasted 8 weeks. R – randomisation

The impact of using each of these rules in LENS was then evaluated by calculating the proportion of participants that would be identified as exposed to a drug at randomisation versus initiation of that drug post-randomisation using each rule (and the corresponding counts from CRF records).

Based on the agreement with CRF data, an optimal observation window was then selected and used for subsequent analyses.

5.3.3.2.2 Methodology development – dispensing dates

Next, given the absence of exact dispensing dates in the data received, different rules for handling dispensing dates were tested, should exact dates be required for time-to-event analyses of drug initiation post-randomisation (e.g. insulin initiation as surrogate for new-onset diabetes). Three different rules were evaluated: 1) imputing day 1; 2) imputing day 15; and 3) imputing last day of the month. As CRF data only recorded a limited number of prohibited drugs after randomisation (8) and occurring in small numbers of participants (n=2), these rules were tested by assessing the resulting agreement observed between PIS data and CRF records for baseline exposure at randomisation – i.e. depending on the imputed date, a record could be included or excluded from baseline assessments based on its relationship with the randomisation date.

5.3.3.2.3 Methodology development – drug doses

Finally, the use of drug doses extracted from PIS data was validated by calculating the proportion of participants for whom the daily statin dose recorded at randomisation matched the strength of the statin medication dispensed in the month of randomisation (or the most recent one previously), split by statin type.

5.3.3.3 Data quality - Agreement between PIS and CRF data

Using the best performing algorithm (identified above), the final resulting agreement between PIS and CRF data for baseline drug exposure was assessed for each drug group, as well as the agreement for initiation of prohibited drugs after randomisation.

5.3.3.4 Use-case scenarios

Finally, the impact of using PIS data in possible use-case scenarios in LENS was evaluated by calculating participant counts identified as exposed to each drug group of interest at baseline in PIS data vs CRF, as well as post-randomisation initiation of prohibited drugs (restricted to follow-up visits occurring up until the 31st August 2021, due to lack of

available PIS data after this date). The additional value of using PIS data to complement assessments of post-randomisation changes for non-prohibited medications (which are not captured in the follow-up CRF data) was also assessed by calculating participant counts identified as exposed or non-exposed at randomisation using CRF data, and quantifying possible changes to drug exposure status during follow-up using PIS data only.

5.3.3.5 Statistical analyses

Agreement for categorical drug exposure at baseline was evaluated using Cohen's κ, an agreement metric widely used in RCD validation studies.[248] Agreement estimates are presented combined across all drug groups (during methodology development), and then split by drug group using the best performing algorithm. For each drug comparison, sensitivity, specificity, and PPV and NPV were also calculated, using the CRF as standard. Sensitivity and specificity were used to assess the performance characteristics of PIS data (i.e. probability of a present or absent CRF record being correctly identified in PIS), and PPV/NPV as measures of PIS data accuracy (i.e. probability that a positive or negative PIS record is true).[219,246,247,275] For post-randomisation initiation of prohibited drugs, only counts in each source are presented due to low event numbers.

Continuous parameters are presented as mean with SD or median with IQR, as appropriate (with distribution normality assessed visually). Frequencies and percentages are presented for categorical parameters. κ coefficients are presented with respective 95% CIs.

5.3.3.6 Data and code

R v4.2.1 was used for all data transformations, statistical analyses, and plotting. The source code used in this work is available at https://github.com/gpessoaamorim/dphil-lens. Further information on the LENS trial is publicly available at

https://www.ctsu.ox.ac.uk/lens. Access to the PIS dataset can be requested for other research purposes from ISD at https://www.publichealthscotland.scot/.

5.3.4 Ethics

LENS is sponsored by the University of Oxford and has been approved by the UK Health Research Authority and the West of Scotland Research Ethics Committee (reference 16/WS/0149). The trial is registered with clinicaltrials.gov (NCT03439345), the ISRCTN registry (ISRCTN15073006), and the European Union registry (EudraCT 2016-002656-24). Permission to use LENS data for this analysis was granted by the trial Steering Committee. Approval for linkage to PIS data were provided by the Scottish Public Benefit and Privacy Panel for Health and Social Care (reference 1516-0338).

5.4 Results

5.4.1 Study population

5.4.1.1 Data coverage and timeliness

This analysis included all 1,151 participants randomised in LENS, of which 1,133 (98.5%) had at least one medication recorded in the randomisation CRF, and 1,147 (99.7%) had available PIS data. Table 5.4 presents baseline characteristics split by linkage status.

Table 5.4 - Baseline characteristics of LENS participants, split by PIS data linkage status

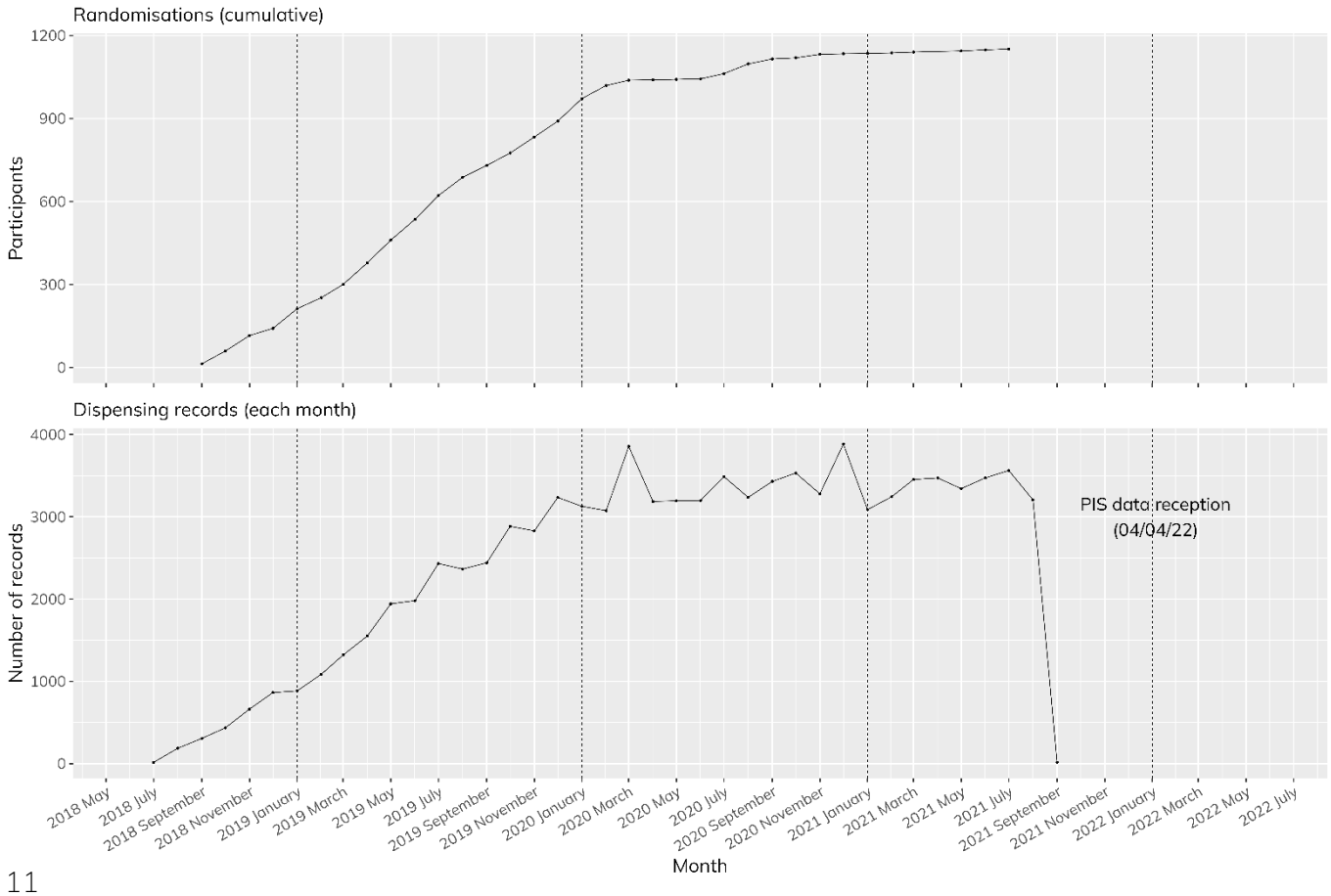
Characteristic	Overall, N = 1,151[1]	Linked, N = 1,147[1]	Non-linked, N = 4[1]
Age, years	60 (12)	60 (12)	50 (12)
Female	312 (27%)	311 (27%)	1 (25%)
Diabetes type			
- Type 1	305 (26%)	303 (26%)	2 (50%)
- Type 2	844 (73%)	842 (73%)	2 (50%)
- Other	2 (0.2%)	2 (0.2%)	0 (0%)
Myocardial infarction	92 (8.0%)	92 (8.0%)	0 (0%)
Stroke/TIA	46 (4.0%)	46 (4.0%)	0 (0%)
Coronary artery bypass graft	39 (3.3%)	39 (3.3%)	0 (0%)
Coronary stent	76 (6.6%)	76 (6.6%)	0 (0%)
Other arterial revascularisation	19 (1.7%)	19 (1.7%)	0 (0%)
Lower limb ulceration	40 (3.5%)	40 (3.5%)	0 (0%)
Lower limb amputation	12 (1.0%)	12 (1.0%)	0 (0%)
Vitrectomy	14 (1.2%)	14 (1.2%)	0 (0%)
Retinal laser treatment	109 (9.5%)	109 (9.5%)	0 (0%)

Table legend: [1]Mean (SD); n (%)

Median time from screening to randomisation was 62 (IQR: 52-70) days, and the median duration of follow-up (until 31st August 2021) was 646 (513-776) days.

PIS data were received in April 2022 and included records starting from July 2018 (first screening visits in the LENS trial) up until September 2021 (Figure 5.2). Data extraction was performed in November 2021, so that there was an underlying 2-3 month lag between the most recent items included in the dataset and the point of data extraction. Most (99%) available records were dispensed in community pharmacies, with the majority including only one item per record (data not shown). Of the pre-specified 18 drug groups of interest, 16 were captured in PIS (exceptions: PCSK9 inhibitors and acarbose or similar).

Figure 5.2 - Cumulative monthly counts of dispensing records and randomisations over time

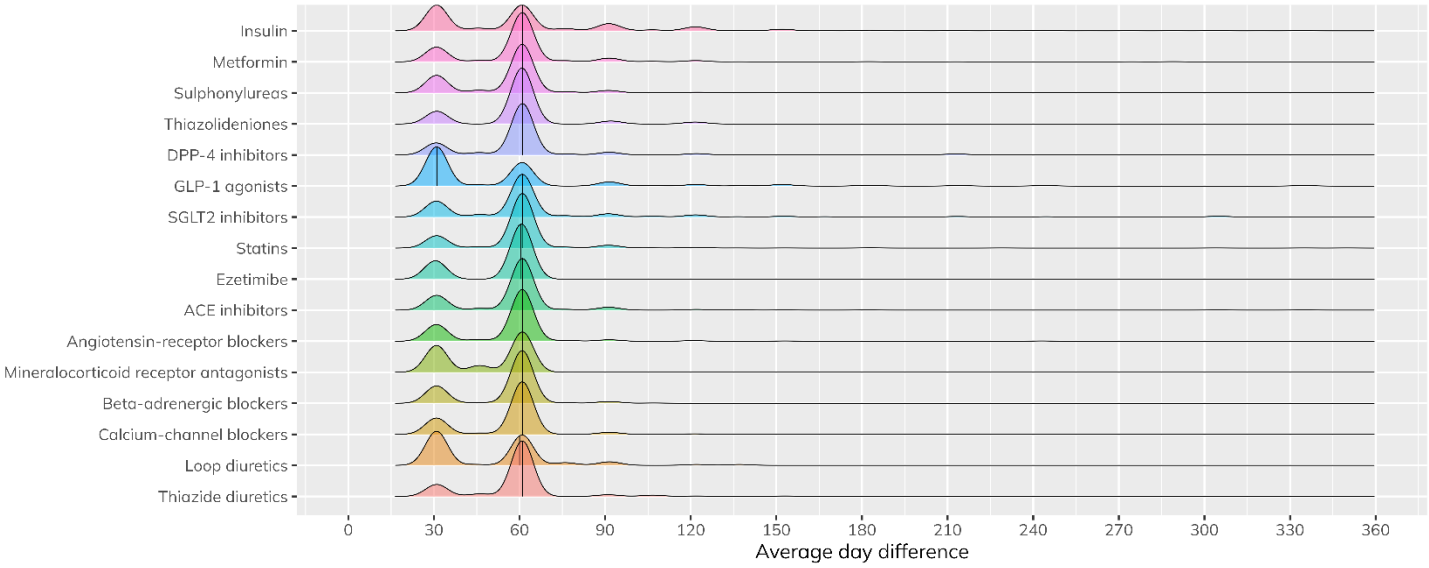


5.4.2 Methodology development

5.4.2.1 Lookback windows and handling peri-randomisation records

The median interval between consecutive dispensing records for the same drug for each participant was 60 days across all drug groups, with the exception of GLP-1 agonists (median 30 days) - Figure 5.3.

Figure 5.3 - Density plots for average time interval in days between consecutive entries of the same BNF code for each participant, per drug group



Approximately 95-100% of individuals had median dispensing intervals shorter than 90 days across all drugs (except insulin: 88%), decreasing to <25-60% of individuals with median dispensing intervals shorter than 2 months (Figure 5.4).

Figure 5.4 – Cumulative proportions of LENS participants with distinct average time intervals in days between consecutive records of the same BNF code in PIS data, by drug group

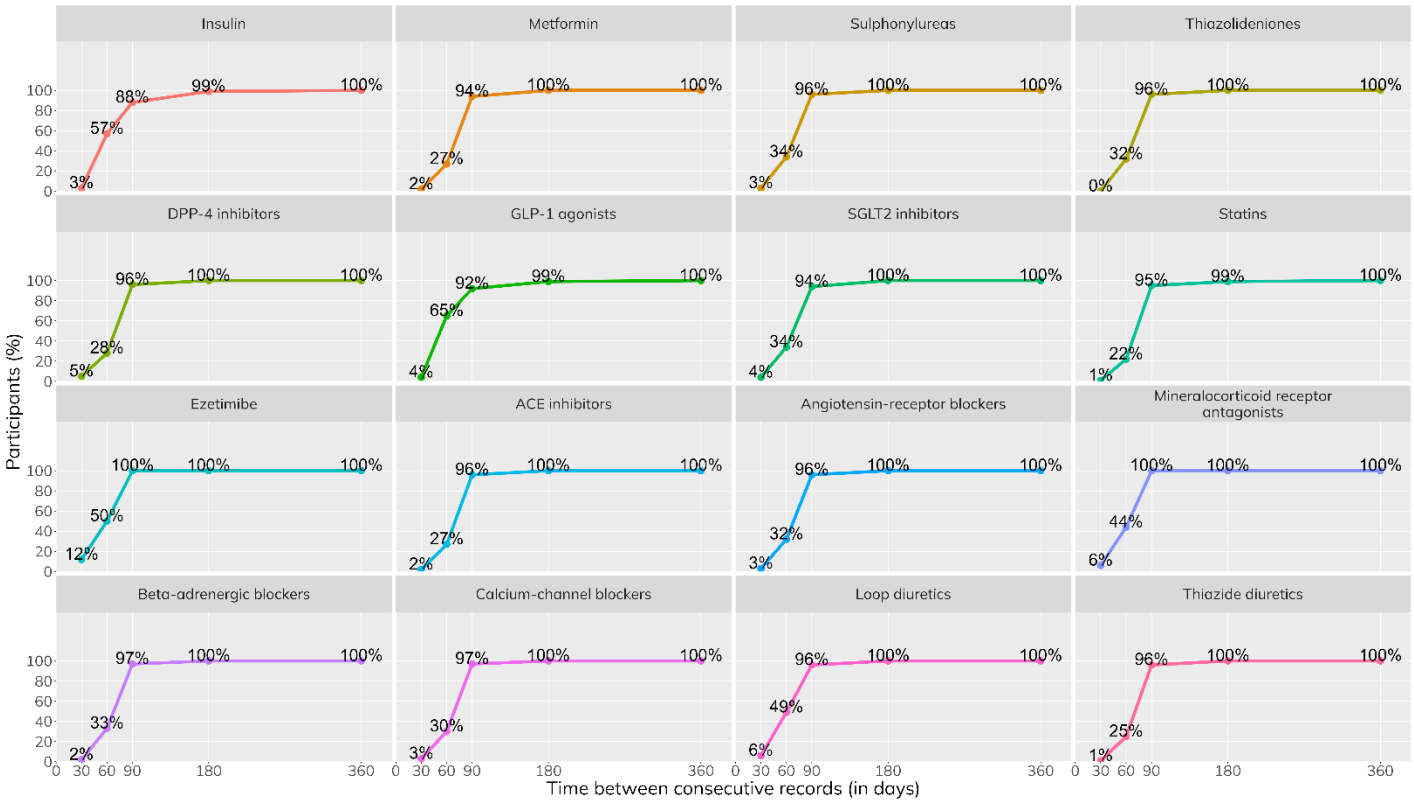


Figure legend: Each point shows the cumulative proportion of participants for whom the average time interval between consecutive drug records for the same drug code is less than 30, 60, 90, 180, or 360 days, split by drug group

Table 5.5 details the impact of using each of the 12 possible rules to derive baseline drug exposure on the agreement between PIS data and the randomisation CRF record, across all drugs of interest combined. Agreement was high (κ>0.80) for 11/12 rules, and increased by both expanding the observation window into the past (rules 1 to 4) and into the future (rules A to C). The largest driver of improved agreement was the inclusion of more recent records, rather than older ones (κ increasing from 0.67 to 0.82, 0.90 to 0.92, and 0.92 to 0.93 by going back one additional month within rules A, B, and C respectively, but with no further improvements with longer lookback windows). Inclusion of data in the month after randomisation as baseline exposure (rule C) only provided marginal benefit when compared with including data up to and including the month of randomisation (κ: 0.93 rule

C vs 0.92 rule B). Sensitivity increased and specificity decreased with increasing fixed window durations.

Table 5.5 - Performance of different rules to derive baseline drug exposure from PIS data vs case report form (CRF) - lookback periods and handling peri-randomisation records

Algorithm rules			Participant-drug pairs				Agreement metrics				
Rule	Lookback period	Handling data in month of randomisation and month after	Both sources	CRF only	PIS data only	Neither source	Sn	Sp	PPV	NPV	κ (95% CI)
1-A	1 month	Exclude month of randomisation	2,473	1,771	166	16,308	58.3	99.0	93.7	90.2	0.67 (0.65 : 0.68)
2-A	2 months	Exclude month of randomisation	3,331	913	253	16,221	78.5	98.5	92.9	94.7	0.82 (0.81 : 0.83)
3-A	3 months	Exclude month of randomisation	3,366	878	255	16,219	79.3	98.5	93.0	94.9	0.82 (0.81 : 0.83)
4-A	Anytime	Exclude month of randomisation	3,374	870	258	16,216	79.5	98.4	92.9	94.9	0.82 (0.81 : 0.83)
1-B	1 month	Include month of randomisation	3,823	421	276	16,198	90.1	98.3	93.3	97.5	0.90 (0.89 : 0.90)
2-B	2 months	Include month of randomisation	3,988	256	303	16,171	94.0	98.2	92.9	98.4	0.92 (0.91 : 0.92)
3-B	3 months	Include month of randomisation	4,004	240	304	16,170	94.3	98.2	92.9	98.5	0.92 (0.91 : 0.93)
4-B	Anytime	Include month of randomisation	4,010	234	307	16,167	94.5	98.1	92.9	98.6	0.92 (0.91 : 0.93)
1-C	1 month	Include month of randomisation and month after randomisation	4,066	178	341	16,133	95.8	97.9	92.3	98.9	0.92 (0.92 : 0.93)
2-C	2 months	Include month of randomisation and month after randomisation	4,109	135	350	16,124	96.8	97.9	92.2	99.2	0.93 (0.92 : 0.94)
3-C	3 months	Include month of randomisation and month after randomisation	4,119	125	351	16,123	97.1	97.9	92.1	99.2	0.93 (0.92 : 0.94)
4-C	Anytime	Include month of randomisation and month after randomisation	4,121	123	354	16,120	97.1	97.9	92.1	99.2	0.93 (0.92 : 0.94)

Table legend: Participant-drug pair counts provided are for all drugs combined.

Figure 5.5 shows the impact of using the different rules tested on proportions of participants identified as exposed at randomisation versus during follow-up (with more restrictive baseline exposure definitions leading to participants being reclassified as initiating that drug after randomisation). Also shown is the proportion identified as baseline exposure using CRF records for that same drug. Rules B and C closely approximate the proportions of participants identified as baseline exposure from PIS data and the CRF.

Figure 5.5 - Impact of different derivation rules on identification of drug exposure as baseline vs post-randomisation initiation, using PIS data vs case report form (CRF)



Figure legend: Horizontal dashed lines show proportions of participants identified as baseline exposure according to the CRF for each drug (i.e. first vertical bar). Each subsequent bar shows proportions classified as exposed to each drug at baseline (i.e. randomisation) vs initiating that drug post-randomisation for each of the RCD derivation rules tested. More restrictive rules for baseline exposure derivation lead to participants being identified as initiating a drug after randomisation instead of exposed at randomisation (if they also had relevant post-randomisation records).

Based on the agreement results obtained, and the average dispensing patterns described earlier, a pragmatic decision was made to select rule 3-B (include records in the 3 months pre-randomisation and the month of randomisation) for subsequent analyses as it is expected that it will better capture dispensing records occurring on average every two-months.

5.4.2.2 Imputing dispensing dates

Table 5.6 shows the impact of different methods to impute exact dates in dispensing records on the agreement between PIS data and the randomisation CRF record, across all drugs of interest combined.

Table 5.6 - Performance of different rules to derive baseline drug exposure from PIS data vs case report form (CRF) – imputing dispensing dates

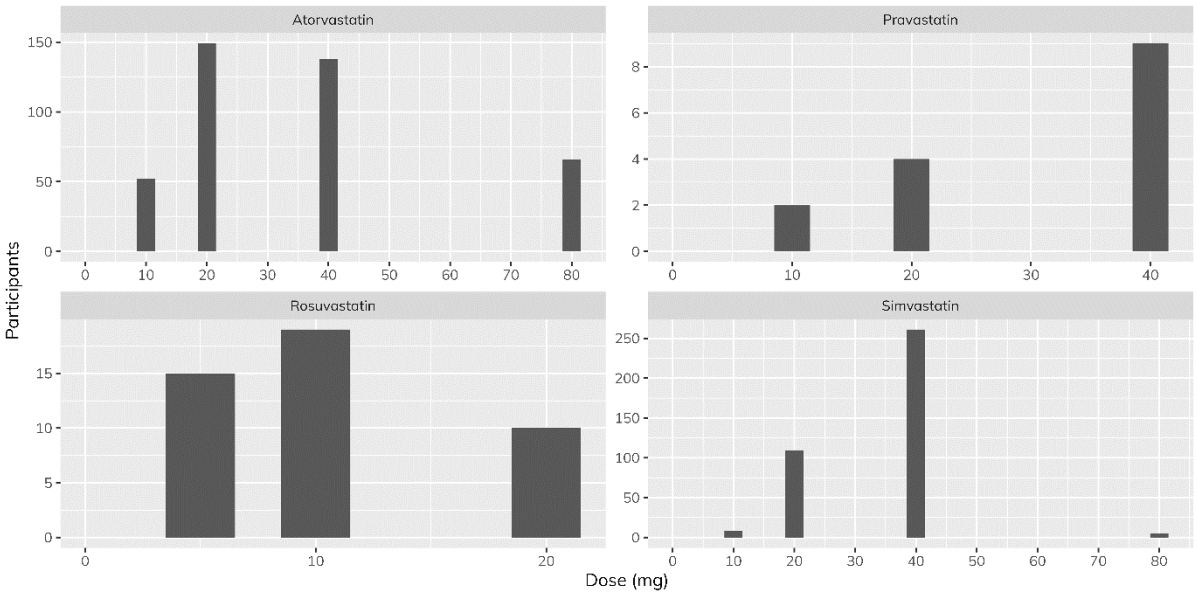
Rule	Participant-drug pairs				Agreement metrics				
	Both sources	CRF only	PIS data only	Neither source	Sn	Sp	PPV	NPV	κ (95% CI)
Use first day of month	3,987	257	301	16,173	93.9	98.2	93.0	98.4	0.92 (0.91 : 0.92)
Use day 15	3,815	429	284	16,190	89.9	98.3	93.1	97.4	0.89 (0.89 : 0.90)
Use last day of month	3,408	836	256	16,218	80.3	98.4	93.0	95.1	0.83 (0.82 : 0.84)

Imputing first day of the month provided best agreement (κ: 0.92, vs 0.89 for day 15 and 0.83 for last day of each month). Based on this assessment, all dispensing records were imputed with day 1 of the corresponding month in subsequent assessments.

5.4.2.3 Validating PIS drug doses

Study participants reported taking 4 different statins in the randomisation CRF, each split across multiple doses (Figure 5.6).

Figure 5.6 - Statin dose distributions (at randomisation)



In 99% of participants with a statin dose recorded in the randomisation CRF (775/784), the strength of the statin PIS record in the month of randomisation (or the most recent one before that) agreed with the daily statin dose recorded in the randomisation CRF (Figure 5.7).

Figure 5.7 - Comparison of statin dose recorded in the randomisation case report form (CRF) vs statin strength in PIS data

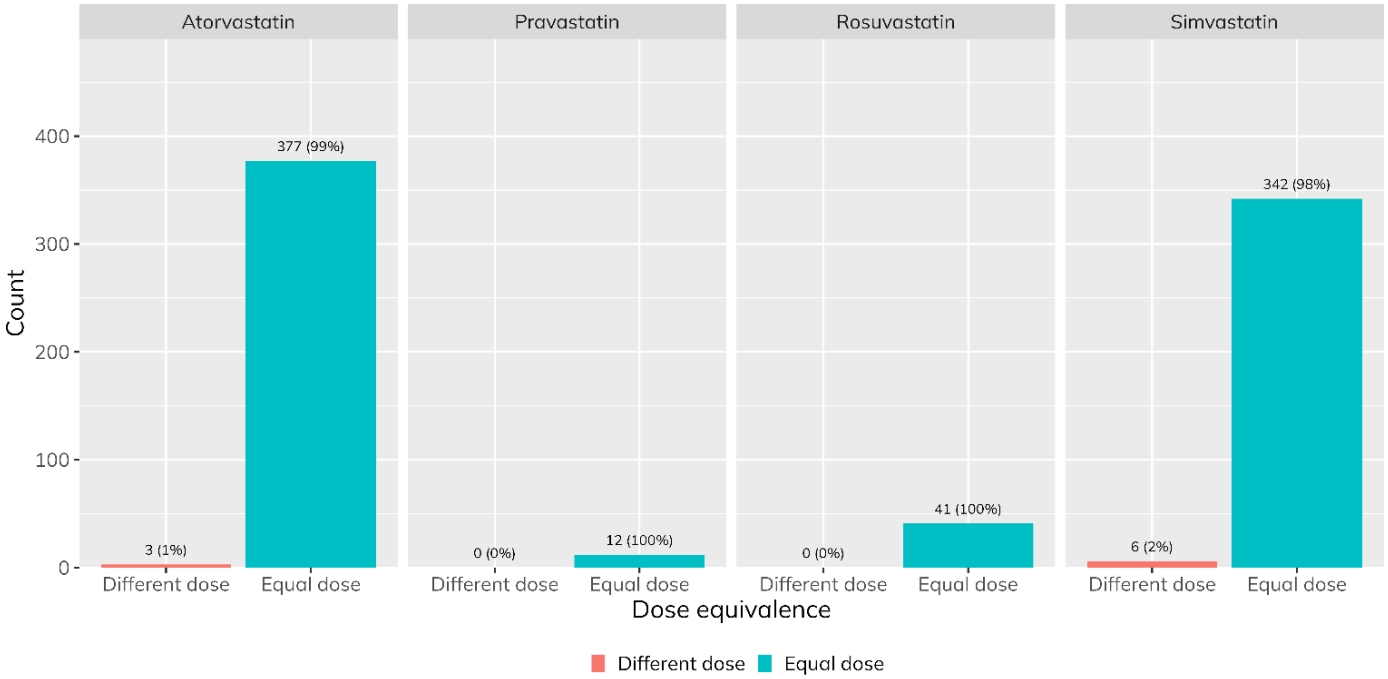


Figure legend: Statin strength from PIS data was extracted from the statin record occurring in the month of randomisation, or the most recent one previously. Only comparisons for drug records where one item was dispensed are shown, as there were very few records with two items dispensed (2 for atorvastatin, 1 for rosuvastatin, all with doses equal to the CRF).

5.4.3 Agreement between PIS and CRF data

Agreement between PIS data and CRF for drug exposure at randomisation was high (κ>0.80) across all drugs investigated, except for ezetimibe (κ=0.71). Sensitivity was >92% for all drugs except ezetimibe (71.4%), a medication with a very small number of participants exposed (9 total identified across both sources). Specificity was >96% for all drugs except statins (90.9%), metformin (90.6%), and insulin (92.4%), which were the three most commonly identified drugs – Table 5.7. PPVs were >85% for all medications except GLP-1 agonists (81.9%), ezetimibe (71.4%), and loop diuretics (78.7%), while NPVs were >97% for all medications except metformin (90.4%), statins (82.3%), and insulin (94%).

1 Table 5.7 - Agreement between PIS data and case report form (CRF) for baseline exposure

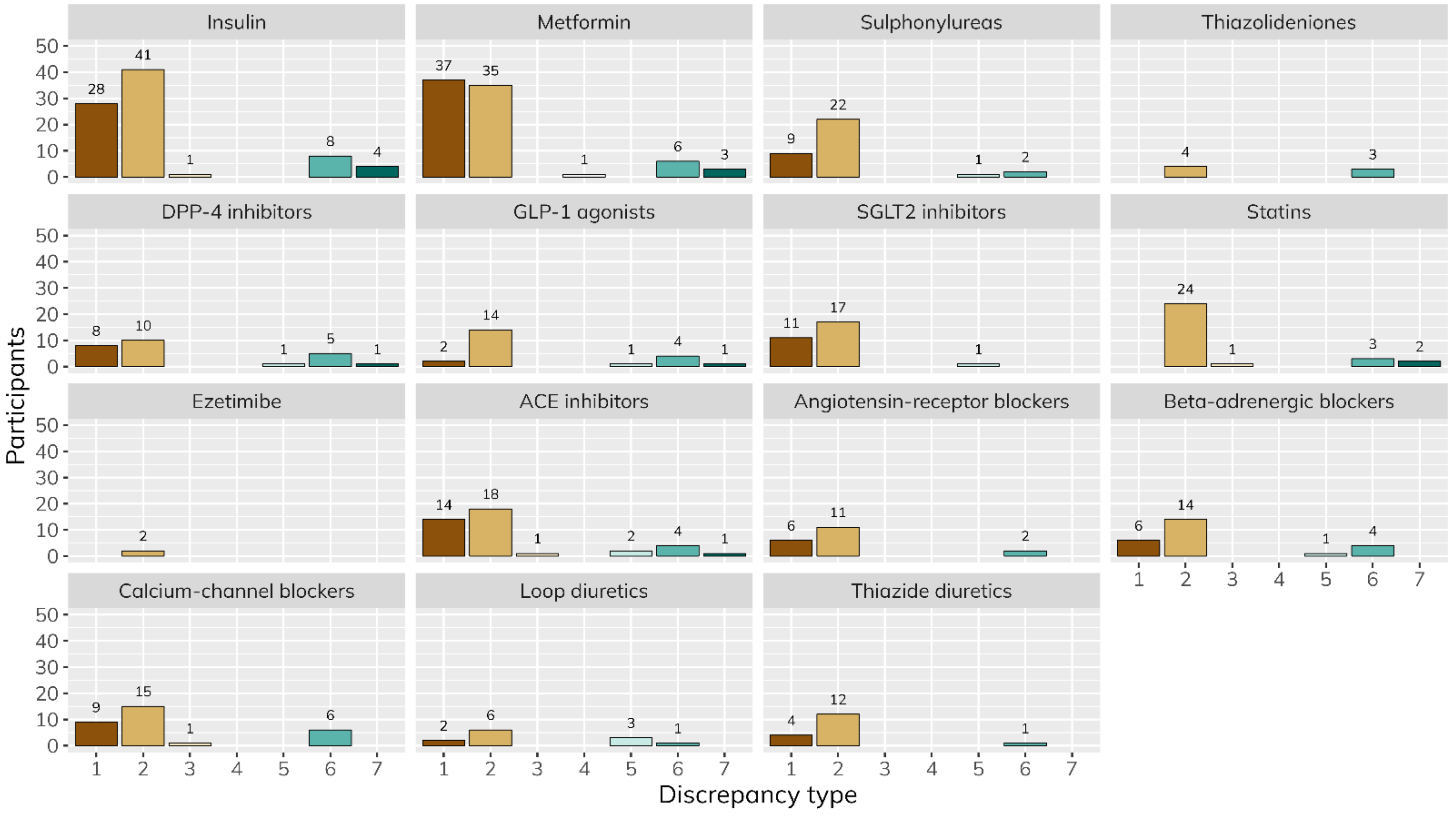
Drug group	Participant counts				Agreement metrics				
	Both sources	CRF only	PIS data only	Neither source	Sn	Sp	PPV	NPV	κ (95% CI)
Insulin	467	38	49	597	92.5	92.4	90.5	94	0.85 (0.82 : 0.88)
Metformin	672	42	41	396	94.1	90.6	94.2	90.4	0.85 (0.82 : 0.88)
Sulphonylureas	286	10	25	830	96.6	97.1	92	98.8	0.92 (0.90 : 0.95)
Thiazolideniones	53	0	7	1,091	100	99.4	88.3	100	0.93 (0.89 : 0.98)
DPP-4 inhibitors	131	10	16	994	92.9	98.4	89.1	99	0.90 (0.86 : 0.94)
GLP-1 agonists	86	7	19	1,039	92.5	98.2	81.9	99.3	0.86 (0.80 : 0.91)
SGLT2 inhibitors	189	11	18	933	94.5	98.1	91.3	98.8	0.91 (0.88 : 0.94)
Statins	796	58	27	270	93.2	90.9	96.7	82.3	0.81 (0.78 : 0.85)
Ezetimibe	5	2	2	1,142	71.4	99.8	71.4	99.8	0.71 (0.43 : 0.99)
ACE inhibitors	476	18	24	633	96.4	96.3	95.2	97.2	0.93 (0.90 : 0.95)
Angiotensin-receptor blockers	182	8	13	948	95.8	98.6	93.3	99.2	0.93 (0.91 : 0.96)
Mineralocorticoid receptor antagonists	13	0	0	1,138	100	100	100	100	1 (1 : 1)
Beta-adrenergic blockers	196	10	19	926	95.1	98	91.2	98.9	0.92 (0.89 : 0.95)
Calcium-channel blockers	277	14	21	839	95.2	97.6	93	98.4	0.92 (0.89 : 0.95)
Loop diuretics	37	3	10	1,101	92.5	99.1	78.7	99.7	0.84 (0.76 : 0.93)
Thiazide diuretics	138	9	13	991	93.9	98.7	91.4	99.1	0.92 (0.88 : 0.95)

Table legend: Baseline exposure calculated using a 3 month lookback period and including data in the month of randomisation.

2

Manual investigation of discrepant records revealed that most discrepancies could be attributed to medications being missed by the CRF (e.g. due to missed recall, or short term medication stop close to randomisation) or PIS data (e.g. medications dispensed in longer intervals and not captured during the run-in period, or no PIS data available), rather than incorrect records being identified from either source (Figure 5.8). This included a number of participants not reporting insulin, metformin, and statin medications at randomisation despite repeat dispensing records for each of these drugs both before and after randomisation (discrepancy type 2). Of note, some of the drug classes with lower sensitivity, specificity, and predictive values (e.g. insulin, metformin, GLP-1 agonists, statins) also had higher numbers of discrepancies with an unclear direction.

Figure 5.8 - Classification of discrepancies between case report form (CRF) and PIS data for baseline drug exposure



Discrepancy type	Description	CRF record	PIS record	Detailed description
1	Missed by PIS (short availability pre-randomisation)	Yes	No	No relevant records during run-in, relevant records within 3 months after randomisation, and average dispensing interval is larger or same as PIS availability pre-randomisation or relevant records before and after randomisation
2	Missed by CRF (established pre-randomisation, continued after)	No	Yes	Distinct relevant PIS records before and after randomisation, several distinct relevant records after randomisation
3	Incorrect recording in CRF (post-randomisation initiation)	Yes	No	First relevant PIS record after month of randomisation, and PIS availability pre-randomisation is longer than the larger dispensing interval for that drug
4	Incorrect recording in CRF (medication stopped before randomisation)	Yes	No	Relevant PIS records before randomisation, but none in the month of randomisation or after
5	Incorrect recording in PIS (medication stopped)	No	Yes	Relevant records only before or in the month of randomisation
6	Unclear (possible initiation in month of randomisation)	Yes	No	First PIS record in month of randomisation, several records after
7	Unclear (missed by PIS or initiation in month after randomisation)	Yes	No	First relevant PIS record in the month after randomisation, several relevant PIS records after randomisation, average dispensing interval shorter than PIS availability before randomisation

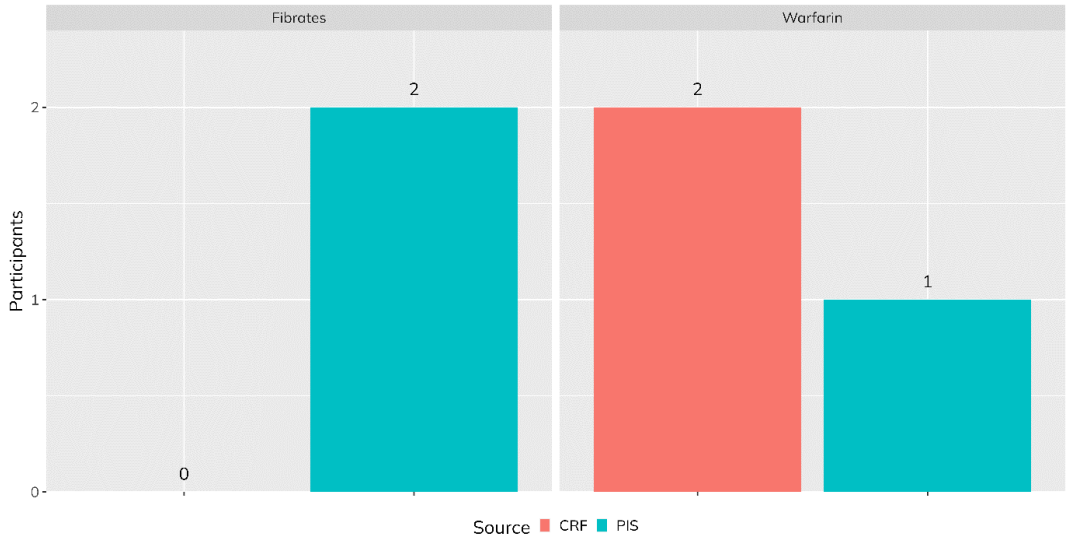
5.4.4 Use-case scenarios

Figures 5.9-5.10 show the possible impact of using PIS data vs CRF to calculate proportions of participants exposed to each drug group at baseline and after randomisation, respectively. Proportions identified from PIS and CRF for baseline exposure were similar across all drug groups considered (maximum absolute difference: 2.7% for statins; PIS: 823 vs CRF 854). During follow-up, initiation of prohibited drugs was recorded in the CRF for 2 participants (warfarin in both cases), one of which was also captured by PIS data. Conversely, PIS data captured two participants with post-randomisation dispensing of fibrates which were not captured by the CRF (although this was due to only one and two distinct dispensing events occurring shortly after randomisation in each case). Of note, PIS data only captured a limited number of prohibited drugs (fibrates, warfarin, and rosuvastatin 40mg), so others could not be assessed.

Figure 5.9 - Baseline exposure to specific drug groups in PIS data vs case report form (CRF)



Figure 5.10 - Post-randomisation initiation of prohibited drugs in PIS data vs case report form (CRF)



When compared with drug exposure at randomisation recorded in the CRF, implementation of PIS data would generally increase proportions of study participants identified as exposed to non-prohibited concomitant drugs during follow-up, with possible impacts on safety and efficacy assessments (e.g. insulin: 43.8% at randomisation to 50.8% during follow-up; metformin: 62% to 67.9%) – Figure 5.11.

Figure 5.11 - Changes to drug exposure after randomisation captured by PIS data if using case report form (CRF) as baseline source

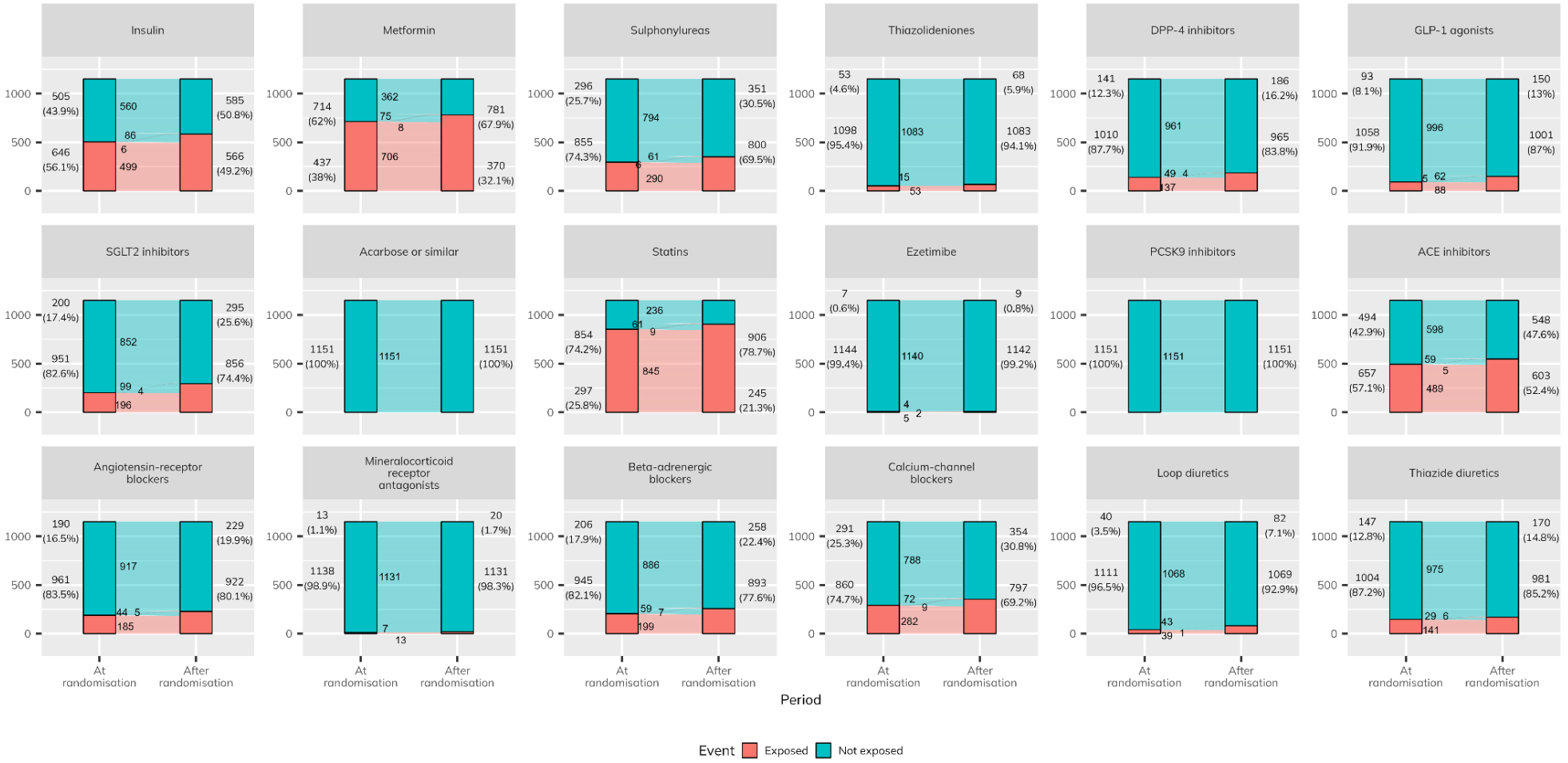


Figure legend: Numbers presented between the "At randomisation" and "After randomisation" columns depict the flow of participants from one period to the other, in the following order (top to bottom): 1) no drug exposure (no exposure at randomisation or after randomisation); 2) drug initiation after randomisation (no exposure at randomisation but confirmed exposure after randomisation); 3) drug stopping after randomisation (exposure at randomisation but no confirmed exposure after randomisation); 4) continued drug exposure (exposure identified at randomisation and after randomisation)

5.5 Discussion

Collection of current medications from participant self-reports or medication reviews, and subsequent recording in CRFs, is considered the standard approach for concomitant medication data collection in trials.[99] However, this method has a number of known limitations (e.g. failed recall, social desirability bias, entry/transcription errors) and is time-consuming.[306–308] RCD sources such as PIS may help tackle these limitations by providing higher quality data at a lower cost. Despite being infrequently used in trials, PIS data have been used in the past to validate self-reported medication exposure for epidemiological research given their high quality (and similarly to studies in other countries).[124,227,260,272,274,295,305–307,309–321] Thus, it would seem reasonable to employ the same approach within trials. This chapter presents the first reported use of PIS data within a drug intervention trial, and the second study directly comparing PIS with a collateral data source (greatly expanding the breadth of medications previously assessed).[164] PIS data captured >99% of LENS participants and had a high level of agreement with CRF data across all medications studied, yielding similar numbers of individuals exposed to each medication at randomisation. These results suggest that PIS data are of adequate quality for use within trials in Scotland (including of drug interventions), with potential for complementing or replacing standard methods of concomitant medication data collection.

PIS data captured almost all (99.7%) LENS participants. This confirms previous reports of very high capture rates in this dataset (close to 100% across the entire dataset, and 96% in Hafferty et al).[164] There was a 2-3 month lag between the most recent data item included and the time of data extraction. This hinders using PIS data for real-time medication use monitoring (as advised by NHS Scotland),[162] but is similar to that seen with other RCD sources in the UK and US.[247]

Agreement between PIS and CRF data for drug exposure at randomisation was high (κ>0.80) across all but one possible observation window. K coefficients were influenced by the duration of the observation window pre-randomisation, increasing with inclusion of older records but also (and to a larger degree) more recent records, including those occurring after randomisation. While unusual, the use of data occurring after an event to define a pre-existing exposure is not unprecedented, and could be justified in this situation given the short span of data available before randomisation.[305] However, the added value of including post-randomisation events was residual if already including records in the month of randomisation (rule C vs rule B). Moreover, if including the month of randomisation in the observation window, agreement remained mostly unchanged despite inclusion of older records (rules 1B/2B/3B/4B). The final rule selected (3B – inclusion of records in the 3 months preceding randomisation and in the month of randomisation) yielded very high agreement with the CRF, without requiring inclusion of post-randomisation events. This rule also replicates the underlying average dispensing intervals of 2 months observed here. Hafferty et al found that a 6-month window outperformed a 3-month window for agreement between self-report and PIS data, albeit by small differences (κ increased by 0.00-0.06, and by larger amounts for gynaecological medications such as hormone-replacement therapy and oral contraceptives, which are likely to be dispensed at longer intervals).[164] In the present study, addition of older records (rules 4-A, 4-B, 4-C) had a residual impact on agreement versus shorter windows, and tended to over-estimate exposure (due to loss of specificity) when compared with CRF records. Therefore, the short temporal availability of PIS data before randomisation is unlikely to have meaningfully harmed the observed agreement with CRF data. Moreover, this limitation could be overcome by requesting consent for transfer of earlier records in future studies.

The methodological assessment of different rules to impute dispensing dates was restricted by low numbers of post-randomisation medication initiation events (for which an exact study visit date could potentially have been used as reference). In the absence of such data, PIS records were imputed with different possible exact dates, and the resulting agreement calculated versus CRF recorded drug exposure at randomisation. Imputing day 1 of each month generated the best agreement, but with small differences versus other options. These results might reflect the artefactual inclusion of more recent events within baseline exposure assessments if an earlier date is used, thus compensating for the short period of data available before randomisation, in keeping with the results for the observation window assessment described above (particularly the large impact of including records closer to randomisation). While this approach might not be needed for other studies if exact dates are requested, this derivation procedure provides empirical evidence for imputing day 1 in each record if exact dates need to be used in future analyses within LENS. Altogether, the methodological development exercises performed support the use of the best performing algorithm found here in other trials looking to extract similar drug exposure phenotypes from PIS data. At the same time, the general data-driven methodological framework used can also be repurposed in circumstances where other investigators may wish to perform validation and development of their own, or where one might expect the algorithm chosen here to be impractical or not perform as well.

Using the selected baseline drug exposure derivation rule, comparisons between PIS data and CRF across various drug groups yielded high agreement, with $\kappa>0.80$ for all drugs except ezetimibe (the latter with small numbers). Likewise, sensitivity was >92% for all drugs except ezetimibe, and specificity >96% for all drugs except statins, metformin, and insulin. Positive and negative predictive values were also generally high. Most of these findings may be explained by the impact of small numbers of records missed by CRF reports

when the underlying prevalence is low, thus leading to lower sensitivity and PPVs (e.g. ezetimibe, GLP-1 agonists), and records missed by PIS data when the underlying prevalence is high, leading to lower NPVs (metformin, statins, insulin). Manual investigation of discrepant records suggests that these are likely the result of failed participant recall at randomisation (impacting specificity and PPV), lack of enough PIS data before randomisation to adequately capture dispensing of chronic medications (impacting sensitivity and NPV), and discrepancies with unclear direction, thus suggesting that PIS data are highly accurate in this setting. Moreover, the discrepancy patterns observed suggest that either data source is more likely to miss a correct drug exposure, than to capture an incorrect one. These findings are similar to those reported by Hafferty et al, using the same data source and similar drug groups and derivation procedures.[164] The agreement observed is also generally similar or superior to previous studies in other geographical locations, in which distinct RCD sources were considered adequate for drug exposure assessments.[124,227,260,272,274,295,305–307,309–321] Finally, as shown by the use-case scenarios tested, using PIS data within LENS would provide virtually identical results in quantifying baseline drug exposure. PIS data would also improve capture of post-randomisation exposure (not collected in LENS follow-up CRFs), such as the initiation over time of medications used to control diabetes (as expected in this population). Importantly, there were very few discrepancies with regards to use of prohibited medications after randomisation (for the two such medications that were included in the PIS data extraction). The two discrepant cases for fibrates (captured in PIS but not in the CRF) could possibly represent errors made by the participants GP by misinterpreting participation in the trial as requiring a prescription for fenofibrate.

Overall, the long lag before data dissemination and infrequent data extraction schedule in LENS hinder the use of PIS data for real-time safety monitoring assessments, namely

verifying eligibility for study entry and ongoing eligibility for study treatment based on the exclusion of prohibited medications. However, the high agreement observed between CRF records and PIS data suggests that a hybrid approach that employs PIS data for most drug exposure derivation assessments, complemented by verification of self-reported prohibited drugs, is a valid and feasible approach for concomitant medication recording within randomised trials in Scotland.

5.5.1 Limitations

Several limitations should be acknowledged. First, these results apply to randomised trials in Scotland and to the PIS dataset only, and should not be used to guide or validate the use of RCD sources for trials in other settings. Second, this cohort comprised clinical trial participants with a diagnosis of diabetes, who usually receive one or more prescription medications. Therefore, these results may not be directly applicable to other disease cohorts, and to broader cohorts of diabetic patients outside trials. Nevertheless, these findings are broadly similar to those in Hafferty et al within a population-based cohort study.[164] Third, only some drugs of specific interest to the LENS trial were selected for analysis here. A focus on other drugs with different dispensing patterns or not adequately covered by PIS (e.g. over-the-counter, hospital administration) might have yielded inferior results. Fourth, PIS data were only available for a short period of time before randomisation. Although a limitation in this study, this is likely to only have negatively biased the agreement results obtained (particularly sensitivity and negative predictive values). Fifth, while CRF data (from participant reports) was used here as a reference to guide validation and implementation of PIS data, it may not be a perfect source for identifying current medication exposure. However it is accepted that no single source can be considered a gold-standard for this purpose, and that this choice depends on the specific study design, purpose, and availability of relevant sources.[110] Hence, in this setting the use of self-reports (i.e. the standard

procedure) can be considered not only a reasonable comparator, but also the most informative one in terms of assessing the value of using PIS data in the context of established practice. Finally, agreement between sources was assessed using Cohen's κ, a widely used agreement metric which facilitates comparison with other studies. However, it is acknowledged that κ coefficients are influenced by properties other than observed agreement alone, were not developed for the specific purpose of comparing two sources of data, and their interpretation may be challenging.[262,263] Thus, interpretation of these results should take into account not only κ coefficients in isolation, but also complementary metrics (sensitivity, specificity, predictive values). Moreover, appropriateness of PIS data for use in a particular setting will depend on the specific drug, population, and intended use of the data, so that a given κ coefficient may be adequate in one circumstance but not another.[24,36,192]

5.6 Conclusion

Nationwide, RCD on medications dispensed in the community setting in Scotland from the PIS dataset provides extensive coverage and accurate identification of exposure to a range of cardiovascular medications for participants within a randomised trial of fenofibrate in diabetes. Overall, these results support expanded use of PIS data in future trials in Scotland. The data-driven methodology approach used here may also serve as a framework to guide validation and implementation of PIS data for other medications or trial settings.

CHAPTER 6: Validation, development, and implementation of routinely-collected data on medications in England in a phase-III, regulatory submission trial of inclisiran for secondary cardiovascular disease prevention (ORION-4)

6.1 Chapter summary

Aims: To develop and validate the use of RCD on medications dispensed in the community setting in England within the ORION-4 trial of inclisiran for secondary prevention of cardiovascular events in patients with established cardiovascular disease.

Methods: In ORION-4, non-study medications are recorded in CRFs at each study visit. These include screening (focused on statins and ezetimibe) and randomisation (all current medications at that point), and all follow-up visits (3 months after randomisation and then every 6 months, with all current medications being recorded). RCD on medication use in the community in England from April 2018 onwards are available from the Medicines Dispensed in the Community (Dispensing) dataset. To validate the use of Dispensing data in ORION-4, data quality was first assessed by evaluating capture of the study population and 27 cardiovascular drug groups of interest, as well as data dissemination timelines. Then, different possible algorithms were tested to identify drug exposure from Dispensing data using the study randomisation CRF as reference. As Dispensing data does not include exact record dates (provided as month only), the impact of using different imputation rules (day 1, day 15, or the last day of each month) on agreement for time to drug initiation after randomisation was compared with CRF records (for drugs not being used at randomisation). Extraction of daily statin dose from statin strength in Dispensing data was also tested. Using

the best performing algorithm, agreement between Dispensing and CRF data was then calculated for each visit type (screening, randomisation, follow-up), aggregate and by drug. Finally, the impact of using Dispensing data vs CRF records on particular analyses of interest to ORION-4 was estimated by calculating overall drug exposure prevalence at each visit, describing historic statin Dispensing records for participants not taking a statin at screening, and comparing estimates of statin treatment intensity at randomisation according to each source. Agreement for binary drug exposure (yes/no) was assessed using Cohen's κ coefficients, together with sensitivity, specificity, and predictive values. Agreement for time to drug initiation after randomisation (in days) was assessed using ICCs and Bland-Altman plots.

Results: Dispensing data captured >99% (10,020/10,038) of ORION-4 participants recruited in England up until the 9th March 2023 for whom linkage data was requested, and 26/27 drug groups of interest (exception: PCSK9 inhibitors). Most participants had at least 12 months of historic Dispensing data available before randomisation, with a 3-month lag between the most recent data items included in the dataset and its dissemination. Using Dispensing records in the previous 3 calendar months to define baseline drug exposure at randomisation provided very high agreement with CRF records (κ=0.93 across all drug groups combined), with slight improvement if also including records in the month of randomisation (κ=0.94). All Dispensing date imputation rules yielded similarly highly agreement for estimates of time to drug initiation after randomisation when compared with the date a drug was first recorded in the follow-up CRF (ICC=0.89-0.90). However, agreement was higher if the CRF date was instead imputed as the mid-point between the first follow-up visit with a new drug record after randomisation and the preceding visit (ICC=0.92). Daily statin dose recorded in the CRF agreed with the most recent previous Dispensing statin strength record for approximately 85-90% of participants, across all

statin types and study visits. Using the best performing derivation algorithm, overall agreement between CRF and Dispensing data across all drugs groups combined was high across all visits (κ=0.94-0.96). Sensitivity, specificity, and predictive values were generally above 95%. However, there were important discrepancies for some lipid-lowering medications which were prohibited in conjunction with the study drug (PCSK9 inhibitors and non-study inclisiran), mostly due to these drugs not being adequately captured in Dispensing data. Dates of drug initiation in the Dispensing data were on average 42 days earlier than those from the CRF. Dispensing data yielded generally similar or slightly higher estimates for medication use prevalence at randomisation and during follow-up when compared with the CRF (up to 2% absolute overestimation), except for statins and aspirin (up to 3% absolute underestimation), while agreeing with the CRF for statin treatment intensity at randomisation for 85% of participants.

Conclusions: Dispensing data are a highly complete and accurate source of information on exposure to cardiovascular drugs in the community in England when compared with the study CRF for participants in a trial of secondary cardiovascular prevention. However, Dispensing data dissemination may be slow or miss information on important drug exposures, and should therefore be used in conjunction with standard data collection methods in a hybrid model.

6.2 Introduction

RCD are increasingly being used to support planning, conduct, and analysis of RCTs, globally and in the UK.[49,61] However, repurposing of RCD within RCTs has mostly been limited to trials testing non-pharmaceutical interventions.[49,59,60] This is likely due to concerns about data quality, availability, and the need to develop specific and complex algorithms to incorporate RCD in RCTs.[32] At the same time, most trials assessing new drugs typical employ "explanatory" (i.e. testing if an intervention works) as opposed to "pragmatic" designs (testing which intervention works best, or how to best implement an existing intervention), generally entailing comprehensive and/or intensive data collection schedules.[7,322]

RCD are substantially similar to that collected for purpose in trials, including information on concomitant (non-study) medications.[33,187] The systematic processes underpinning RCD collection may also contribute to improved data quality when compared to that collected by trial staff, while helping reduce burden for participants and costs.[48,49] Thus, use of RCD to streamline randomised trials is not necessarily an impediment to trial quality but may actually help improve it[51,60] (as noted by senior representatives of the US Food and Drug Agency[323] and in a recent report commissioned by the UK Government),[324] and has been endorsed by regulatory agencies worldwide.[24,36]

Before widespread implementation, harnessing RCD on medications within trials of new pharmaceutical interventions requires appropriate validation.[32,33] Chapters 3-4 have investigated the potential value of using two novel RCD sources (on medication dispensing in the community and medication use in-hospital in England) within the RECOVERY trial of inpatient COVID-19 treatments. Those analyses were focused on a small number of medications, and performed in a highly-pragmatic trial undertaken in extraordinary circumstances amidst a global public health crisis.[51,323] Moreover, RECOVERY trial data were

collected using medical records review, while most trials in the community setting typically rely on self-reported medication lists to identify background medication use.

In this chapter, data from a large trial of a new cholesterol-lowering medication for secondary cardiovascular prevention were used to validate, develop, and test the value of using a novel RCD source on medication use in the community in England for RCTs of new pharmaceutical interventions in the outpatient setting.

6.3 Methods

6.3.1 Study population and time period

ORION-4 is an ongoing randomised, placebo-controlled trial of inclisiran (a novel injectable cholesterol lowering-medication) for secondary prevention among men aged 40+ years or women aged 55+ years with established cardiovascular disease.[142] The trial is taking place in the UK and the US. At the beginning of the study, inclisiran was not yet licensed in either country. The trial was therefore planned as a phase-III, cardiovascular outcomes, regulatory submission trial. Following results from three phase-III RCTs (whose primary outcomes were changes in LDL cholesterol),[325] inclisiran has since become approved for the treatment of primary hypercholesterolemia or mixed dyslipidaemia in patients with cardiovascular disease in the UK in 2021. In the US, it was approved in 2021 for lowering LDL cholesterol in adults with familial hypercholesterolemia or clinical atherosclerotic cardiovascular disease, and in 2023 for treatment of adults with high LDL-cholesterol and increased risk of cardiovascular disease. However, ORION-4 is the first trial designed and powered to assess the impact of inclisiran on cardiovascular events.

From March 2019 to September 2023, ORION-4 recruited 16,124 participants, 13,414 of which are in the UK (and 12,398 in England). All participants have provided written informed consent, including for linkage to RCD sources in the UK. The study population of interest for this analysis is composed of ORION-4 participants randomised in England up to the 9th March 2023.

6.3.2 Medication data sources

6.3.2.1 Case report forms

Potentially-eligible ORION-4 participants in the UK were invited by post, after centralised pre-screening using NHS secondary care records. Those willing to take part were asked to telephone the central study office for a pre-screening call. If considered eligible, they were

then invited to attend an initial screening visit in-person at one of the study sites (and asked to bring a list of current medications). At the screening visit, a focused medical history was taken and eligibility to the study confirmed. A randomisation visit then followed approximately 8 weeks later, in which ongoing study eligibility was confirmed, and any updates to medical history were recorded. The first follow-up visit was then scheduled approximately 3 months later, with subsequent visits approximately every 6 months until the end of the study.

Current concomitant medications were recorded by research nurses in a medication list using an electronic CRF at each visit. At screening, study staff were trained to focus on accurate recording of lipid-lowering medications (namely statins and ezetimibe), but were instructed to record all current concomitant medications at each following visit. At each visit, participants were reminded to bring an updated list of their current medications to the next visit. Daily drug doses were only recorded for statins, and used to define intensity of cholesterol-lowering treatment (high intensity: atorvastatin ≥40mg or rosuvastatin ≥20mg; moderate intensity: atorvastatin 10-39 mg, rosuvastatin 5-19 mg, simvastatin ≥20mg, pravastatin ≥40mg, lovastatin ≥40mg, fluvastatin 80mg, pitavastatin ≥1mg; low-intensity: any other lower dose statin regimen). Participants not taking a statin at screening were also asked to provide a reason (intolerance to one statin, intolerance to more than one statin, clinical recommendation, patient preference, or unsure).

Table 6.1 depicts the medication groups of interest to ORION-4 selected for exploration in this analysis (in collaboration with the trial Principal Investigators). These included concomitant medications of interest, as well as prohibited medications (non-study inclisiran, and PCSK9 inhibitors).

Table 6.1 - Drug groups of interest to ORION-4 included for analysis

General medication class	Medication group
Lipid-lowering	Statins (combined and individual statins)
	Ezetimibe
	PCSK9 inhibitors
	Inclisiran
	Bempedoic acid
	Icosapent ethyl
	Fibrates
	Resins
Diabetes	Insulin
	Metformin
	Sulphonylureas
	Thiazolidinediones
	DDP-4 inhibitors
	GLP-1 agonists
	SGLT2 inhibitors
	Acarbose or similar
Hypertension/ heart failure	ACE inhibitors
	ARBs
	Mineralocorticoid receptor antagonists
	Beta-adrenergic blockers
	Calcium channel blockers
	Loop diuretics
	Thiazide diuretics
Antithrombotic	Aspirin
	Other antiplatelets (clopidogrel, ticagrelor, prasugrel, dipyridamole)
	Vitamin K antagonists (warfarin, acenocoumarol, coumadine)
	Non-vitamin K oral anticoagulants (NOACs; apixaban, edoxaban, rivaroxaban, dabigatran)

6.3.2.2 Medicines Dispensed in the Community dataset

Nationwide, RCD on medication use in the community are being extracted from the Medicines Dispensed in the Community (Dispensing) dataset.[230] The Dispensing dataset has been previously described in Chapter 3. Briefly, this resource includes information on medications prescribed in the community setting in England, and submitted for reimbursement in England/Scotland/Wales/Isle of Man/Jersey/Guernsey/Alderney to the NHS Business Services Authority since April 2018. Amongst other information, the data include medication name, SNOMED and BNF codes, and monthly period in which a drug

was submitted for reimbursement by the dispenser.[270] More details on data collection pathways and data structure are publicly available from NHSE.[230]

Linkage to Dispensing data was performed by NHSE using a combination of personal identifiers including name, biological sex, date-of-birth, and NHS number. Written informed consent was received from ORION-4 participants for linkage to all available Dispensing data. The dataset included here was received by the study team on the 18th April 2023, following submission of a cohort of ORION-4 participants randomised in England up to the 09th March 2023. Only one data extract was available for analysis.

Following advice provided from NHSE (personal communication), drug exposure was extracted using the *prescribed_bnf_code*, *date_of_processing, and prescribed_strength* fields. For each drug group, a codelist of BNF codes was built via interrogation of the BNF hierarchy within the OpenCodelists platform.[240] All codelists were built by one clinician (the candidate) and manually reviewed and approved by a second one, and are available for inspection and reuse at https://github.com/gpessoaamorim/dphil-orion-4.

6.3.3 Data analysis

The purpose of this analysis was to assess whether Dispensing data can be used as a potential source of information on exposure to non-study medications for ORION-4 participants recruited in England.[33,244] As described in previous chapters, the general framework used followed recommendations issued by the FDA and the MHRA for the use of RCD in clinical research.[191,192,242] These documents recommend considering data quality in a contextual/extrinsic dimension, rather than intrinsic/internal data quality. Details on internal data quality assessments for the Dispensing dataset are publicly available.[239] The contextual data quality dimensions considered here were completeness/coverage, timeliness, and accuracy/agreement with a relevant reference standard (in this case, collateral medication data captured in the study CRF). Similarly to previous chapters, given

the novelty of this dataset and in the absence of specific guidance on how to extract drug exposure from this resource, different possible drug exposure derivation algorithms were tested (using agreement with the study randomisation CRF as reference). Using the best performing algorithm, agreement between RCD and CRF was then assessed for different possible drug exposures of interest. Finally, the impact of using Dispensing data on specific use-case scenarios relevant to ORION-4 was tested by comparing the estimates obtained from RCD vs the CRF.

6.3.3.1 Data quality – coverage and timeliness

First, data coverage was assessed by calculating counts of individuals in the study population with available Dispensing data, and computing baseline characteristics (as recorded in the randomisation CRF) for those with vs without available RCD (i.e. linkage status). To check for temporal or geographical gaps, date and place of randomisation were also plotted by linkage status. Capture of drug groups of interest was then assessed by calculating counts of distinct records and participants included in the Dispensing dataset for each group. Finally, timeliness was investigated by plotting monthly counts of Dispensing records over time, and assessing historical data length (i.e. how far the oldest records included spanned) and recency (i.e. the interval between data reception and the more recent records included).

6.3.3.2 Drug exposure methodology development

6.3.3.2.1 Lookback window and handling of peri-randomisation records

As described in Chapter 2, multiple methods can be used to identify drug exposure at a particular time point using RCD.[326] A number of candidate rules to define cross-sectional drug exposure from RCD at randomisation were identified and tested against the corresponding CRF records. The randomisation visit was chosen as reference for this development exercise as this was felt to be the visit in which more emphasis was made

during investigator training on correct recording of all medications, and therefore was considered likely to be the most accurate source (vs screening or follow-up).

The following derivation options were tested:

1) How long should the observation window span into the past (1, 2, 3, 6, or 12 months)
2) Should records in the month of randomisation be included or ignored for derivation of drug exposure at randomisation (as they cannot reliably be ascertained as pre- or post-randomisation due to the lack of exact record dates)

For this purpose, underlying drug dispensing patterns were first evaluated in the Dispensing dataset by calculating average intervals between consecutive records of the same BNF code for each participant and for each drug group (for participants who had at least two relevant records for each drug group). This assessment followed preliminary data exploration which showed that, on average, most study participants had one record for each distinct BNF code each month, one distinct BNF code per drug group per participant, and one distinct record per drug group per month (data not shown). Then, for each drug group, cumulative proportions of participants with average dispensing intervals within 1, 2, 3, 6, and 12 months were calculated to test the potential impact of using each of these lookback windows. Next, the agreement between Dispensing data and CRF recorded medications at randomisation was tested using each possible lookback window (i.e. defining drug exposure from RCD as the presence of a relevant drug record in the 1, 2, 3, 6, or 12 months preceding the month of randomisation), and at the same time either including or ignoring records in the month of randomisation. In total, 10 (5x2) different rules were tested. The impact of using each of these rules in the ORION-4 trial was then evaluated by calculating the proportion of participants that would be classified as exposed to a drug at randomisation versus initiating that drug post-randomisation. The rule which yielded the highest level of agreement with the randomisation CRF was then selected for subsequent analyses.

6.3.3.2.2 Handling exact dates for drug initiation during follow-up

As previously described, Dispensing data does not provide exact record dates (recorded as monthly period only). While this can be circumvented in some cases by using calendar months for drug exposure derivation (as outlined above for exposure at randomisation), it may be of interest to impute a specific date for some potential uses of medications data (e.g. safety assessments using time-to-event analyses, such as insulin initiation as surrogate for new-onset diabetes). At the same time, the CRF does not record the exact date in which a drug was initiated, and only the date of the visit in which that drug was first recorded is available.

To explore the possibility of using exact medication initiation dates, CRF and Dispensing records were extracted for participants identified as initiating a new drug post-randomisation according to the CRF (i.e. drug group not recorded at randomisation, but recorded later), and who had at least one matching Dispensing record after the month of randomisation. Exact dates were then imputed using two sets of rules:

1. Dispensing date handling: imputing day 1, day 15, or last day of the month;
2. CRF date handling: taking the date of the first visit in which a new drug was recorded, or imputing the mid-point between that visit and the preceding visit (on or after the date of randomisation)

Using the exact dates obtained with each rule combination, the agreement between both sources for time to drug initiation (calculated in days from randomisation) was then assessed.

6.3.3.2.3 Statin dose

Information on medication strength is provided in Dispensing data, but not prescribed daily dose. To test extraction of daily dose from medication strength, the daily statin dose recorded in the CRF at each study visit was compared with the strength of the statin

medication dispensed in that month (or the most recent one before that), separately by statin type and by study visit type (screening, randomisation, or follow-up).

6.3.3.3 Data quality – agreement between RCD and CRF

The final resulting agreement between RCD and CRF records was then calculated for each visit type (screening, randomisation, follow-up), aggregate and by drug group. Agreement during follow-up was calculated both by comparing drugs recorded at each individual visit, and drugs recorded at any time across the entire duration of follow-up. Agreement for time to drug initiation after randomisation was then assessed for participants with a relevant drug record in both the CRF and RCD, and restricted to RCD records occurring after the month of randomisation.

6.3.3.4 Use-case scenarios

Finally, the impact of using Dispensing data in specific analyses of particular interest to ORION-4 was tested in comparison with the CRF. These included:

1) calculating overall proportions of participants exposed to each drug of interest at screening, randomisation, and during follow-up;

2) characterising historic statin exposure for participants who self-reported not taking a statin at screening; and

3) defining cholesterol-lowering treatment intensity at randomisation.

6.3.3.5 Statistical analyses

Agreement for binary drug exposure at each visit was evaluated using Cohen's κ, an agreement metric widely used in RCD validation studies.[248] Agreement estimates are presented combined across all drug groups during methodology development, and then both aggregate and split by drug group in the final agreement comparisons (performed using the best performing derivation algorithm). Sensitivity, specificity, and PPV/NPV, were also calculated alongside κ coefficients, using CRF as standard.[219,246,247,275] Agreement

between CRF and RCD estimates of time to drug initiation since randomisation was assessed using ICCs and Bland-Altman plots. ICCs were computed using two-way mixed effects models, agreement between sources (vs measurement consistency), and single measurements as unit of analysis (vs average of several measurements), following previously published methodology.[135,214,249,250]

Continuous parameters are presented as mean with SD or median with IQR, as appropriate (with distribution normality assessed visually). Frequencies and percentage are presented for categorical parameters. κ coefficients and ICCs are presented with respective 95% CIs. Bland-Altman plots are shown with estimates of mean differences and 95% LOA. Error bars in all plots show standard errors, unless specified otherwise.

6.3.3.6 Data and code availability

R v4.3.1 was used for all data transformations, statistical analyses, and plotting. The source code used in this work is available at https://github.com/gpessoaamorim/dphil-orion-4. Access to the Dispensing dataset can be requested for other research purposes from NHSE at https://digital.nhs.uk/services/data-access-request-service-dars.

6.3.4 Ethics

ORION-4 is jointly sponsored by the University of Oxford and Novartis (previously The Medicines Company), and has been approved by the UK Health Research Authority and the Oxfordshire C Research Ethics Committee (Ref. No 18/SC/0243). The trial is registered with clinicaltrials.gov (NCT03705234), the ISRCTN registry (ISRCTN88876914), and the European Union registry (EudraCT 2017-005066-22). Permission to use ORION-4 data for this analysis was granted by the trial Steering Committee. Further information on the ORION-4 trial is publicly available at https://www.orion4trial.org/.

Results

6.3.5 Study population

The overall study population considered in this analysis consisted of 10,038 ORION-4 participants randomised in England up to the 09th March 2023, of whom 10,020 (99.8%) had available linked Dispensing data. Table 6.2 shows baseline characteristics of the study population, split by linkage status. Manual review of CRF recorded medications for the 18 participants with no Dispensing data showed that all had records for multiple medications at the randomisation visit, including several cardiovascular medications selected for this analysis. These participants were recruited throughout the duration of the study (Figure 6.1), and in several sites across England (Figure 6.2).

Table 6.2 - Baseline characteristics of ORION-4 participants recruited in England and submitted in data linkage request, split by Dispensing dataset linkage status

Characteristic	Overall, N = 10,038	Linkage data available, N = 10,020	Linkage data unavailable, N = 18
Age in years, mean (SD)	70 (8)	70 (8)	64 (9)
Female	2,813 (28%)	2,809 (28%)	4 (22%)
Race			
White	9,826 (98%)	9,808 (98%)	18 (100%)
Asian	109 (1.1%)	109 (1.1%)	0 (0%)
Black	39 (0.4%)	39 (0.4%)	0 (0%)
Mixed	34 (0.3%)	34 (0.3%)	0 (0%)
Other	30 (0.3%)	30 (0.3%)	0 (0%)
Diabetes	1,906 (19%)	1,904 (19%)	2 (11%)
Myocardial infarction	7,366 (73%)	7,353 (73%)	13 (72%)
CABG	1,668 (17%)	1,667 (17%)	1 (5.6%)
PCI	4,798 (48%)	4,791 (48%)	7 (39%)
Stroke	2,275 (23%)	2,271 (23%)	4 (22%)
Lower limb arterial revascularisation	1,129 (11%)	1,128 (11%)	1 (5.6%)
Aortic aneurysm repair (surgery or stent)	498 (5.0%)	498 (5.0%)	0 (0%)

Figure 6.1 – Number of monthly randomisations over time, split by linkage status

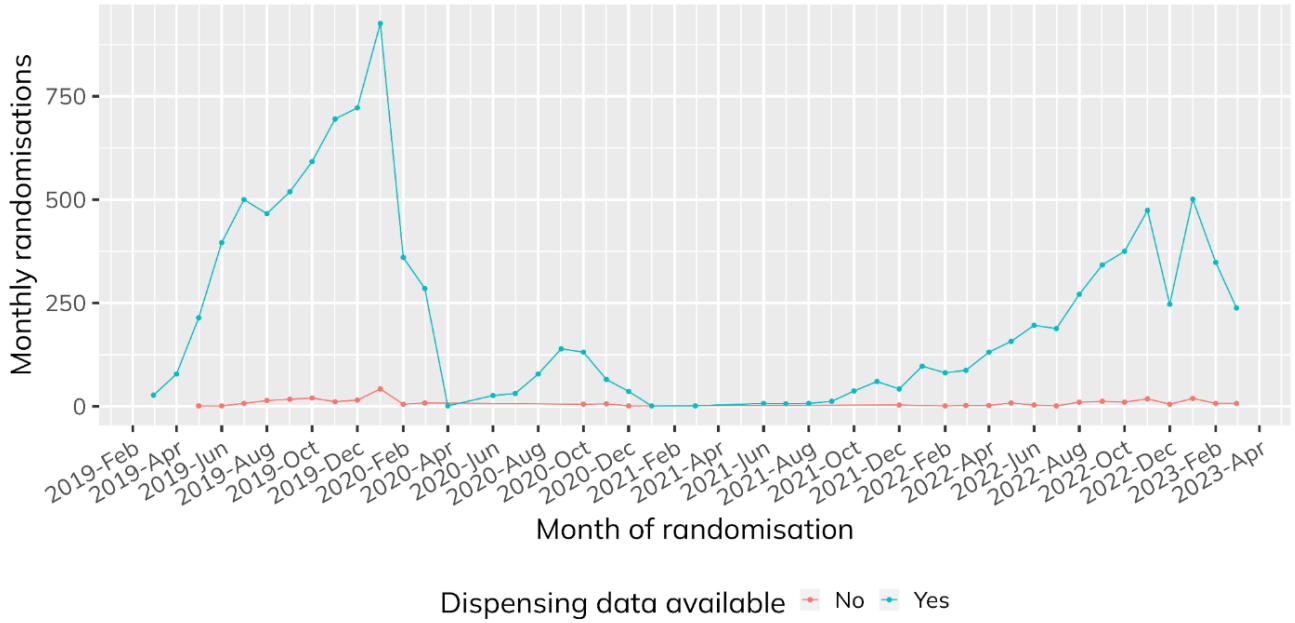
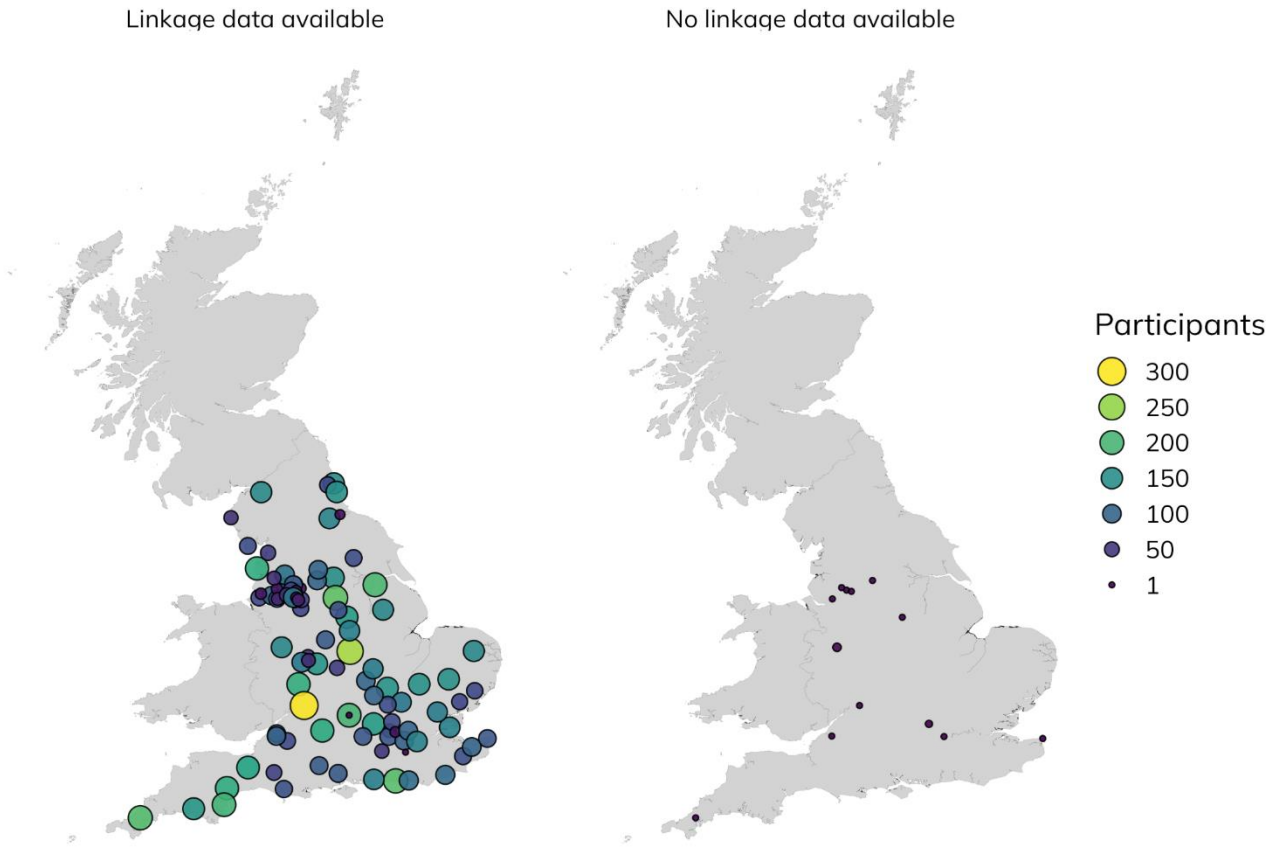


Figure 6.2 - Recruitment location of the study population, split by linkage status



6.3.6 Drug coverage and timeliness

All medication groups of interest were captured in the Dispensing dataset, with the exception of PCSK9 inhibitors (Figure 6.3). Of note, PCSK9 inhibitors are only prescribed within secondary care in England, while inclisiran may be prescribed in either primary or secondary care.

Figure 6.3 - Capture of drug groups of interest in the Dispensing dataset

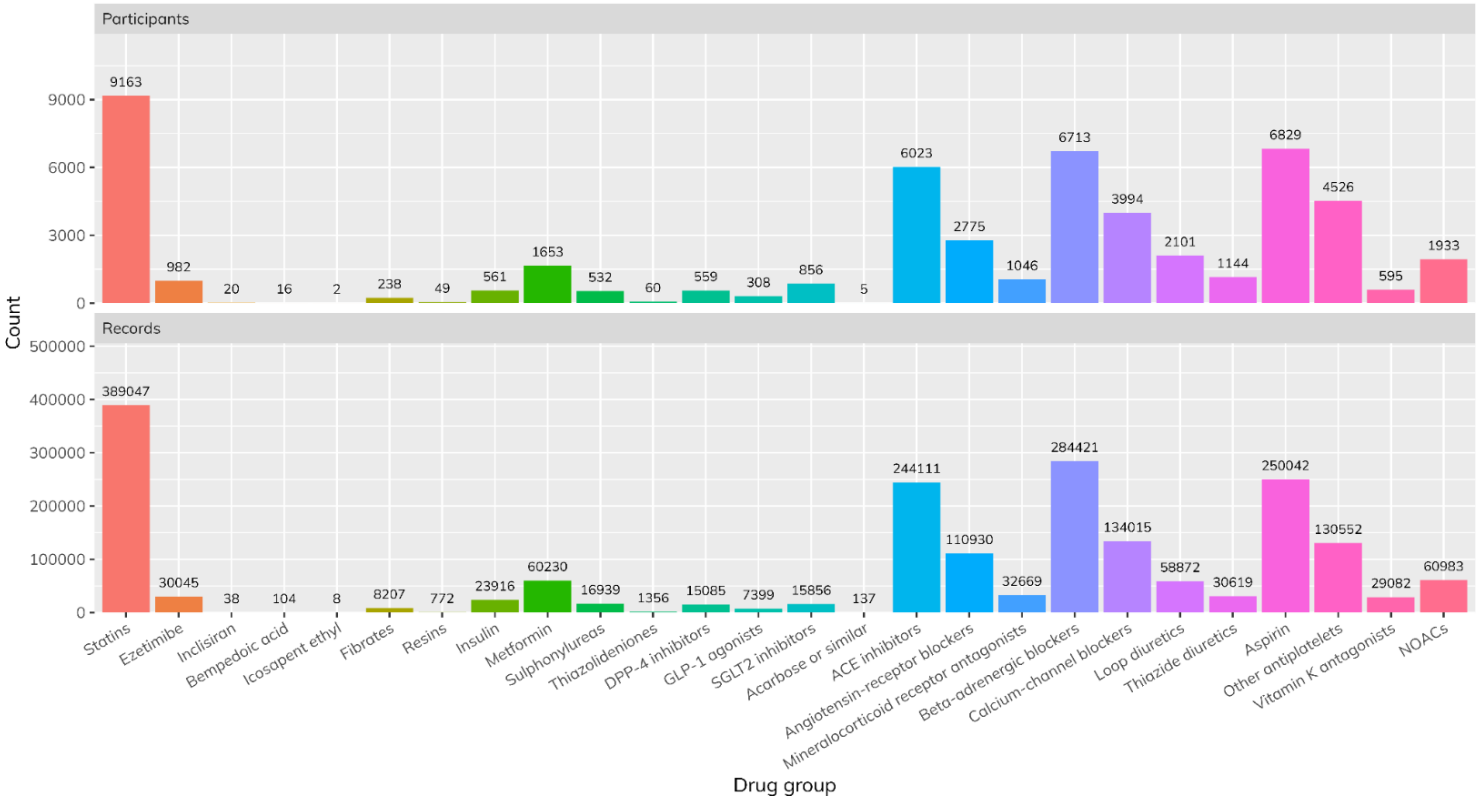


Figure legend: Each panel shows counts for total distinct participants and total distinct records (i.e. distinct participants and dates) for each drug group. All drug groups of interest were captured, except PCSK9 inhibitors (0 records).

The Dispensing data extract used here was received in April 2023, and included records spanning from April 2018 up to January 2023 – i.e. there was a 3 month gap between the last data items included in the dataset and its dissemination to the study team (data not shown). Most participants had at least 12 months of historic Dispensing data before randomisation (Figure 6.4). Following this assessment, subsequent analyses comparing

RCD and CRF records were restricted to data recorded in study visits up to the 31st January 2023 inclusive.

Figure 6.4 - Histogram of historic Dispensing data available before randomisation for each participant

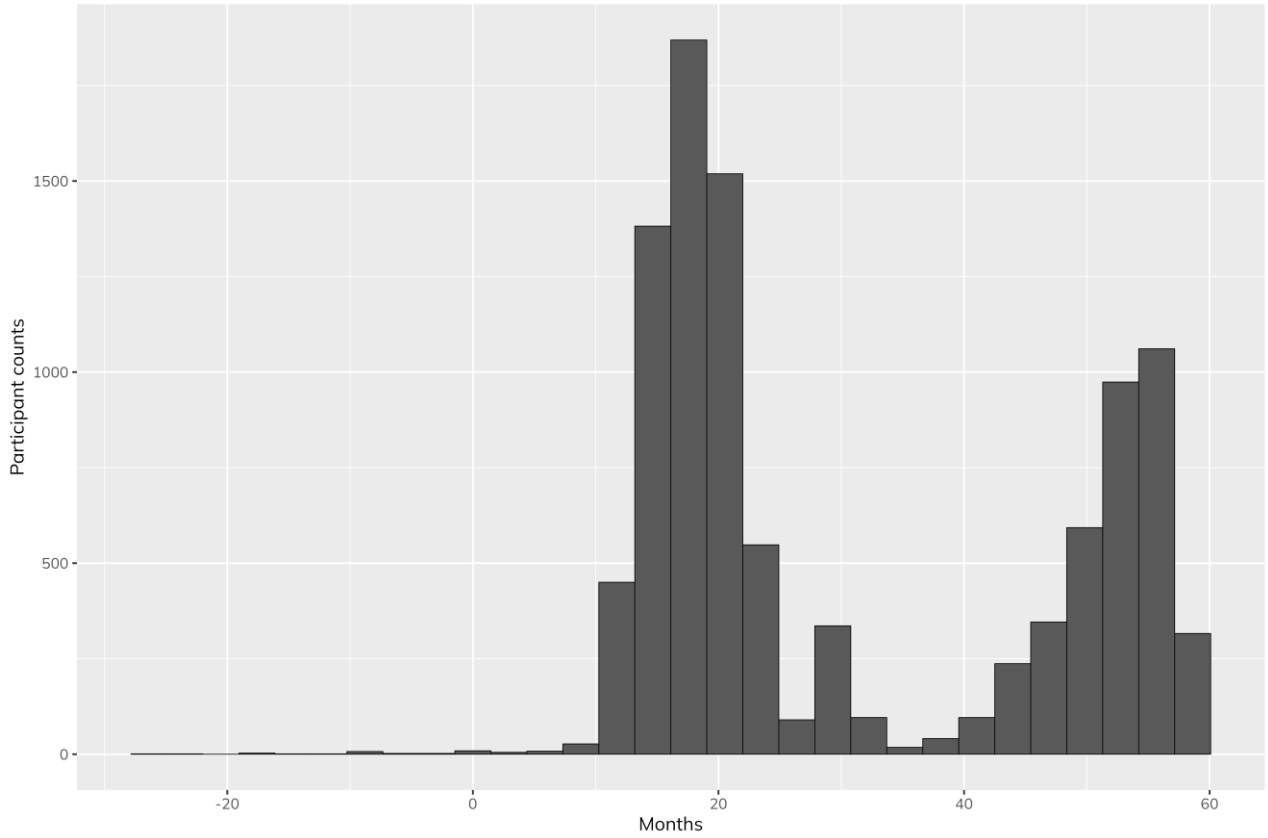


Figure legend: The horizontal axis depicts time from the first Dispensing record to randomisation (with positive values representing participants for whom the first Dispensing record is before randomisation). Months defined as 30-day periods

6.3.7 Methodology development

6.3.7.1 Lookback windows and handling of peri-randomisation records

The median interval between consecutive dispensing records for the same drug for each participant was 30 days across all drug groups, with the exception of icosapent ethyl, resins, and insulin (45 days), vitamin K antagonists (which includes warfarin; 60 days) and inclisiran (90 days) - Figure 6.5. Approximately >90% participants had median dispensing intervals shorter than 90 days across all drugs (except inclisiran: 38%) - Figure 6.6.

Figure 6.5 - Histograms of average intervals between consecutive records of the same BNF code for each participant, by drug group

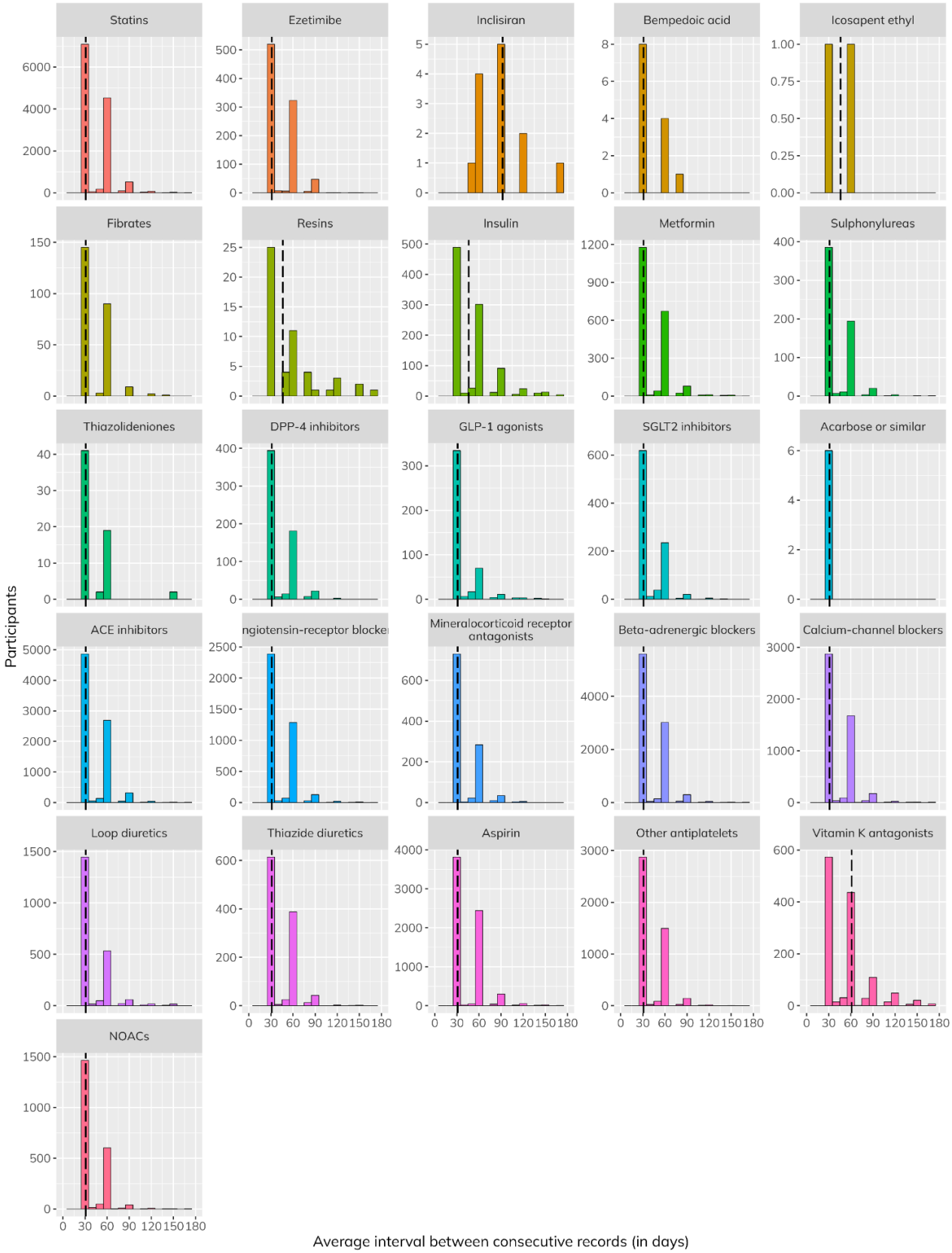


Figure legend: Averages calculated as medians. Vertical black lines depict medians in each drug group. Horizontal axes capped at 180 days

Figure 6.6 - Cumulative proportions of ORION-4 participants with different average time intervals between consecutive records of the same BNF code, by drug group

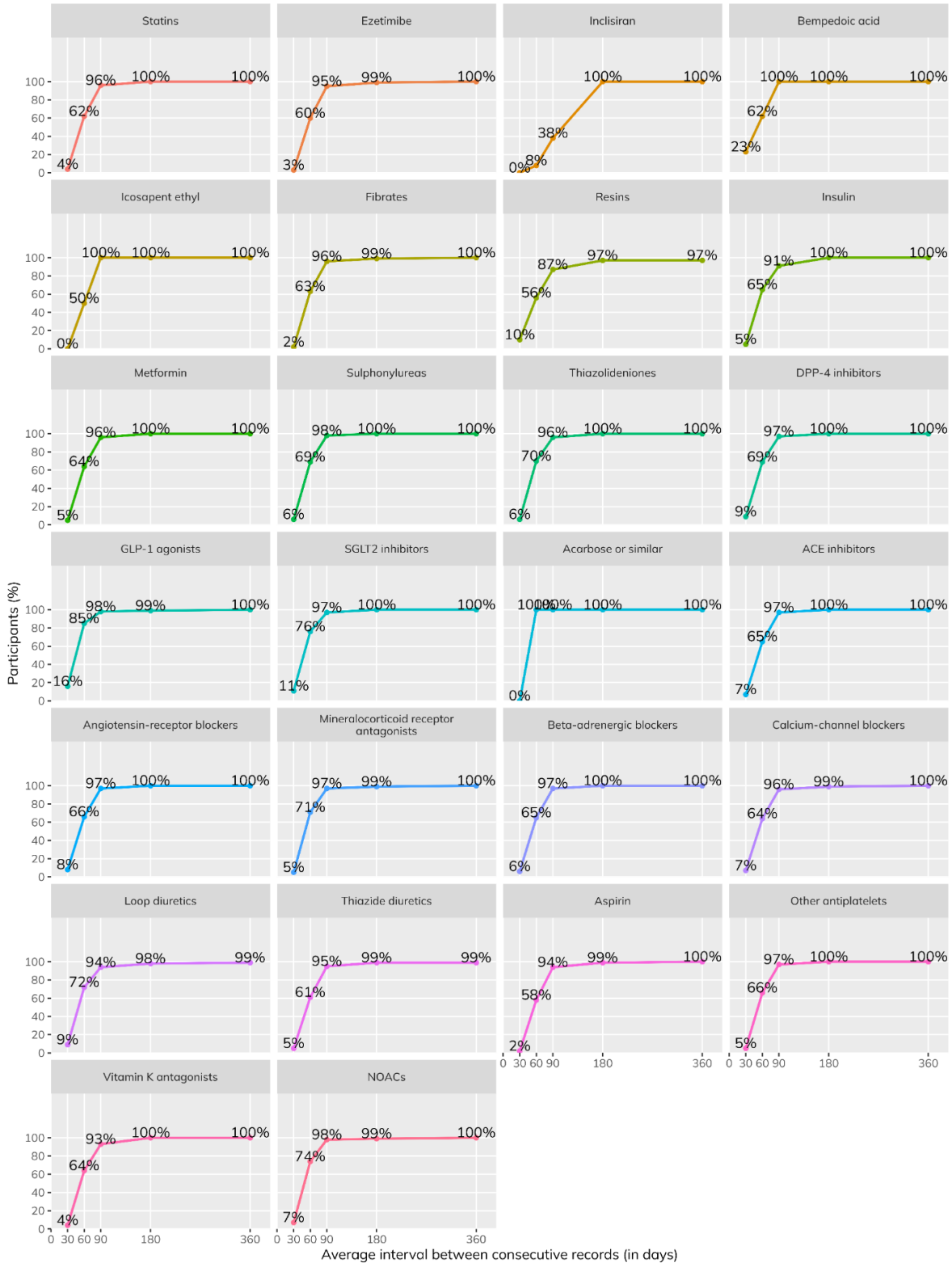


Figure legend: Each point shows the cumulative proportion of participants for whom the average time interval between consecutive drug records for the same drug code is less than 30, 60, 90, 180, or 360 days, split by drug group. Averages calculated as medians.

Table 6.3 shows the impact of using each of the 10 possible rules used to derive baseline drug exposure on the agreement between Dispensing data and the randomisation CRF, across all drugs groups combined. Agreement was high ($\kappa \geq 0.90$) for all rules tested, except rule 1-A (including only records in the month before randomisation, driven by low sensitivity of 67.3%). Agreement was highest with lookback periods of 3-6 months if excluding records in the month of randomisation ($\kappa = 0.93$), or 3 months if including records in the month of randomisation ($\kappa = 0.94$).

Table 6.3 - Performance of different rules to derive baseline drug exposure from Dispensing data when compared with self-reported exposure at the randomisation visit

Rules			Participant-drug pair counts				Agreement parameters				
Rule	Lookback period	Handling data in month of randomisation	Both sources	CRF only	RCD only	Neither source	Sn	Sp	PPV	NPV	K (95% CI)
1-A	1 month	Exclude month of randomisation	26,391	12,823	1,717	219,589	67.3	99.2	93.9	94.5	0.75 (0.75 : 0.76)
2-A	2 months	Exclude month of randomisation	35,328	3,886	2,385	218,921	90.1	98.9	93.7	98.3	0.90 (0.90 : 0.91)
3-A	3 months	Exclude month of randomisation	37,055	2,159	2,698	218,608	94.5	98.8	93.2	99.0	0.93 (0.93 : 0.93)
4-A	6 months	Exclude month of randomisation	38,024	1,190	3,369	217,937	97.0	98.5	91.9	99.5	0.93 (0.93 : 0.94)
5-A	12 months	Exclude month of randomisation	38,323	891	4,812	216,494	97.7	97.8	88.8	99.6	0.92 (0.92 : 0.92)
1-B	1 month	Include month of randomisation	34,819	4,395	2,367	218,939	88.8	98.9	93.6	98.0	0.90 (0.89 : 0.90)
2-B	2 months	Include month of randomisation	37,082	2,132	2,646	218,660	94.6	98.8	93.3	99.0	0.93 (0.93 : 0.93)
3-B	3 months	Include month of randomisation	37,735	1,479	2,877	218,429	96.2	98.7	92.9	99.3	0.94 (0.93 : 0.94)
4-B	6 months	Include month of randomisation	38,225	989	3,493	217,813	97.5	98.4	91.6	99.5	0.93 (0.93 : 0.94)
5-B	12 months	Include month of randomisation	38,444	770	4,920	216,386	98.0	97.8	88.7	99.6	0.92 (0.92 : 0.92)

Figure 6.7 shows the proportion of participants identified as being exposed to a drug at randomisation using CRF records (first vertical bar), and the impact of using the different rules tested on proportions of participants identified as exposed at randomisation versus initiating each drug after randomisation (as more restrictive definitions of pre-randomisation exposure may lead to reclassification as post-randomisation initiation for participants with multiple records over time). Rules 3-A, 4-A, 3-B, and 4-C closely approximate estimates of drug exposure at randomisation from Dispensing data and the CRF, but with small differences across rules. For example, the prevalence of statin use at randomisation according to the CRF is 85.4%, with rules 4A (including records in the 6 months pre-randomisation), 3B (records in the 3 months pre-randomisation or in the month of randomisation), and 4B (records in the 6 months pre-randomisation or in the month of randomisation) providing the closest estimates (85.4%, 84.6%, and 86.1% respectively).

Based on these assessments, exposure at randomisation was defined in Dispensing data as the presence of a relevant drug record within the three calendar months before or within the month of randomisation.

Figure 6.7 - Impact of different derivation rules on identification of drug exposure as baseline vs post-randomisation initiation from Dispensing data vs case report form (CRF)

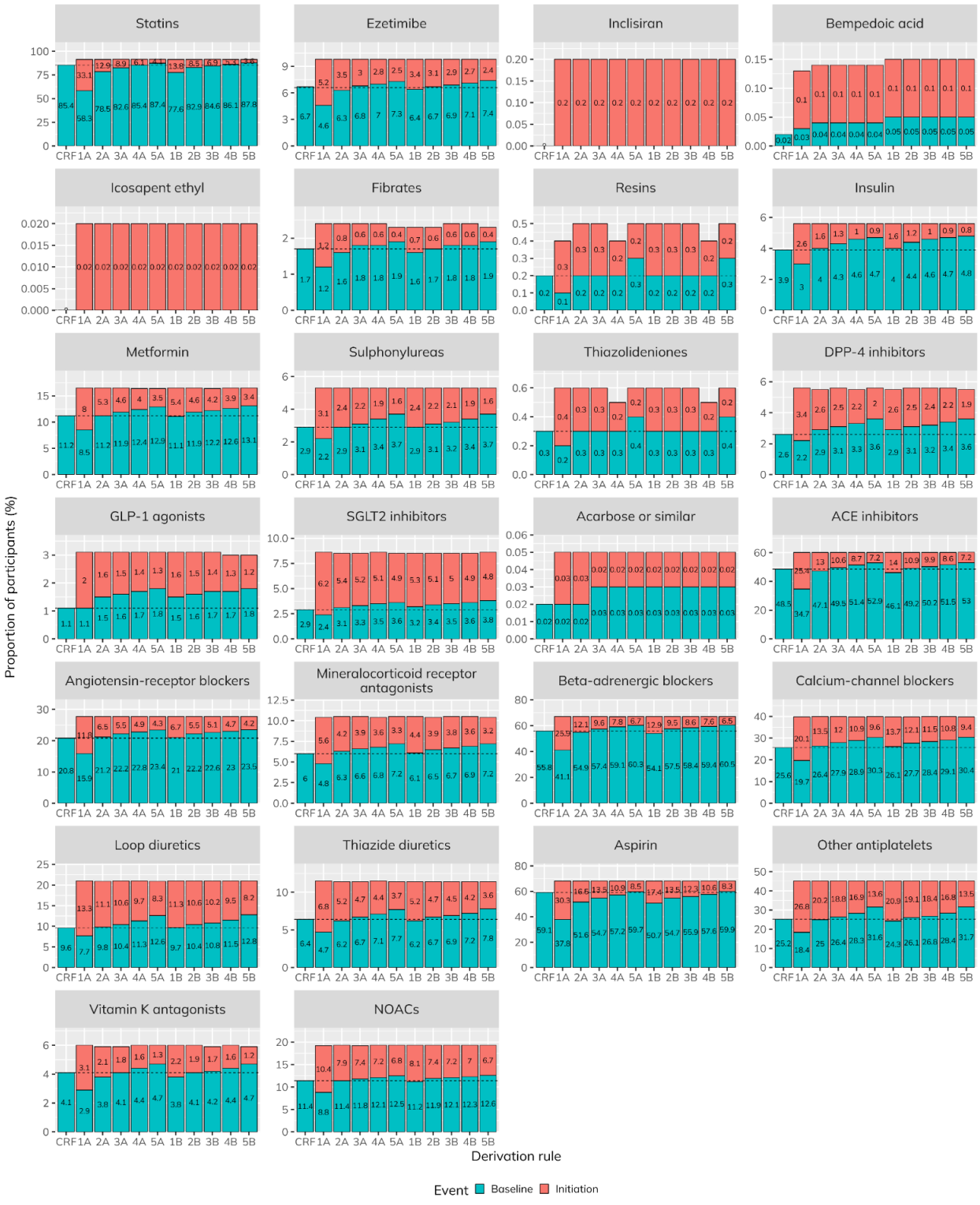


Figure legend: Horizontal dashed lines show proportions of participants identified as baseline exposure according to the CRF for each drug (i.e. first vertical bar). Each subsequent bar shows proportions classified as exposed to each drug at baseline (i.e. randomisation) vs initiating that drug post-randomisation for each of the RCD derivation rules tested. More restrictive rules for baseline exposure derivation lead to participants being identified as initiating a drug after randomisation instead of exposed at randomisation (if they also had relevant post-randomisation records).

6.3.7.2 Handling exact dates for drug initiation during follow-up

Exploration of different possible approaches to impute exact dispensing dates was initiated by comparing estimates of time to first drug record after randomisation between CRF and Dispensing data (calculated in days), for participants identified as initiating a drug after randomisation according to the CRF (and restricting to Dispensing records occurring after the month of randomisation). Imputing the last day of each month as the exact date in all Dispensing records yielded the best agreement with CRF records (using the visit date in which a new drug was first recorded; ICC=0.90), but agreement was high across all Dispensing date imputation rules - Figure 6.8. However, agreement between Dispensing data and the CRF was improved if the CRF initiation date was instead imputed as the mid-point between the first follow-up visit with a new drug recorded and the previous visit, and then similar regardless of the Dispensing date imputed (ICC=0.92). Therefore, subsequent assessments of time to drug initiation were performed by imputing the mid-point date in CRF records (as it was assumed that a choice that improves agreement between two collateral sources expected to be of reasonably high quality would maximise estimate accuracy). Dispensing records were imputed with day 15 of each month as a pragmatic choice.

Figure 6.8 - Comparison of estimates for time to drug initiation since randomisation from case report form (CRF) vs Dispensing data, across combinations of possible date handling rules

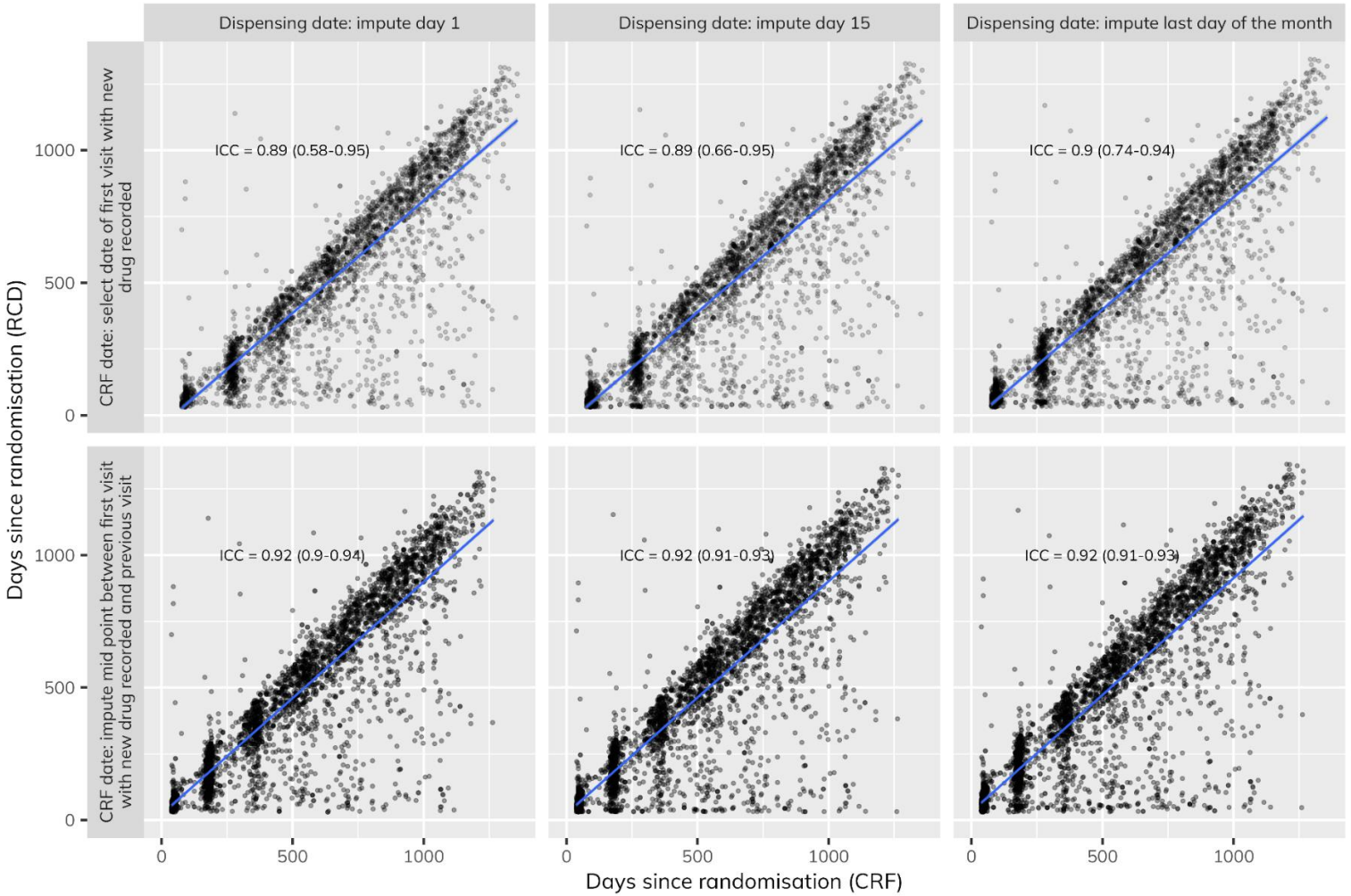


Figure legend: Each panel shows one rule combination (3 possible rules to impute exact dates in Dispensing records, and 2 to handle uncertain drug initiation dates from CRF records). Each dot shows one individual participant-drug pair. ICCs are calculated across all participant-drug pairs for each rule combination, and shown along 95% Cis. Linear regression lines are also shown, along with respective 95% CI margins.

6.3.7.3 Validating statin drug doses from RCD records

Figure 6.9 shows distributions of statin doses recorded at each study visit. For most participants, the daily statin dose recorded in the study CRF was the same as the strength of the most recent statin record in Dispensing data for the same statin, with similar results across all visits and statins (approximately 85-90% for screening, 90% for randomisation, and 90-95% for follow-up; Figure 6.10). Discrepancies were due to both different statin strengths in Dispensing data, and absence of Dispensing records for that particular statin.

Figure 6.9 - Statin dose distributions from case report form (CRF) data, by statin and study visit

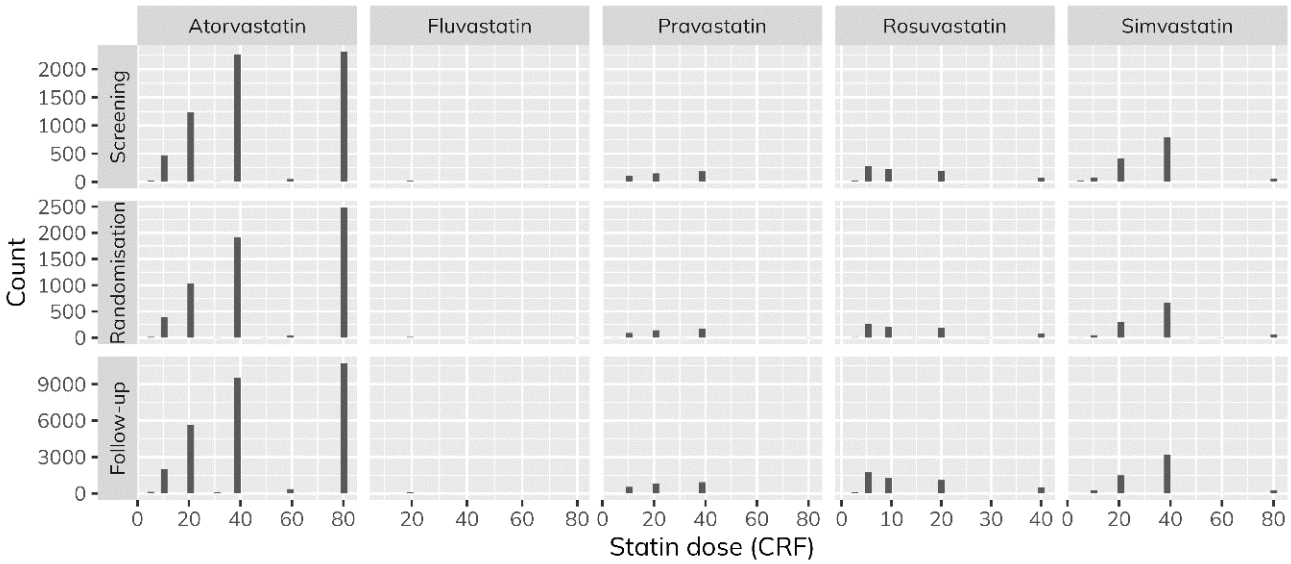


Figure legend: The vertical axis shows participant counts for screening and randomisation visits, and participant-visit pairs for follow-up visits.

Figure 6.10 - Comparison of statin dose recording in case report form (CRF) vs Dispensing data, by statin and study visit

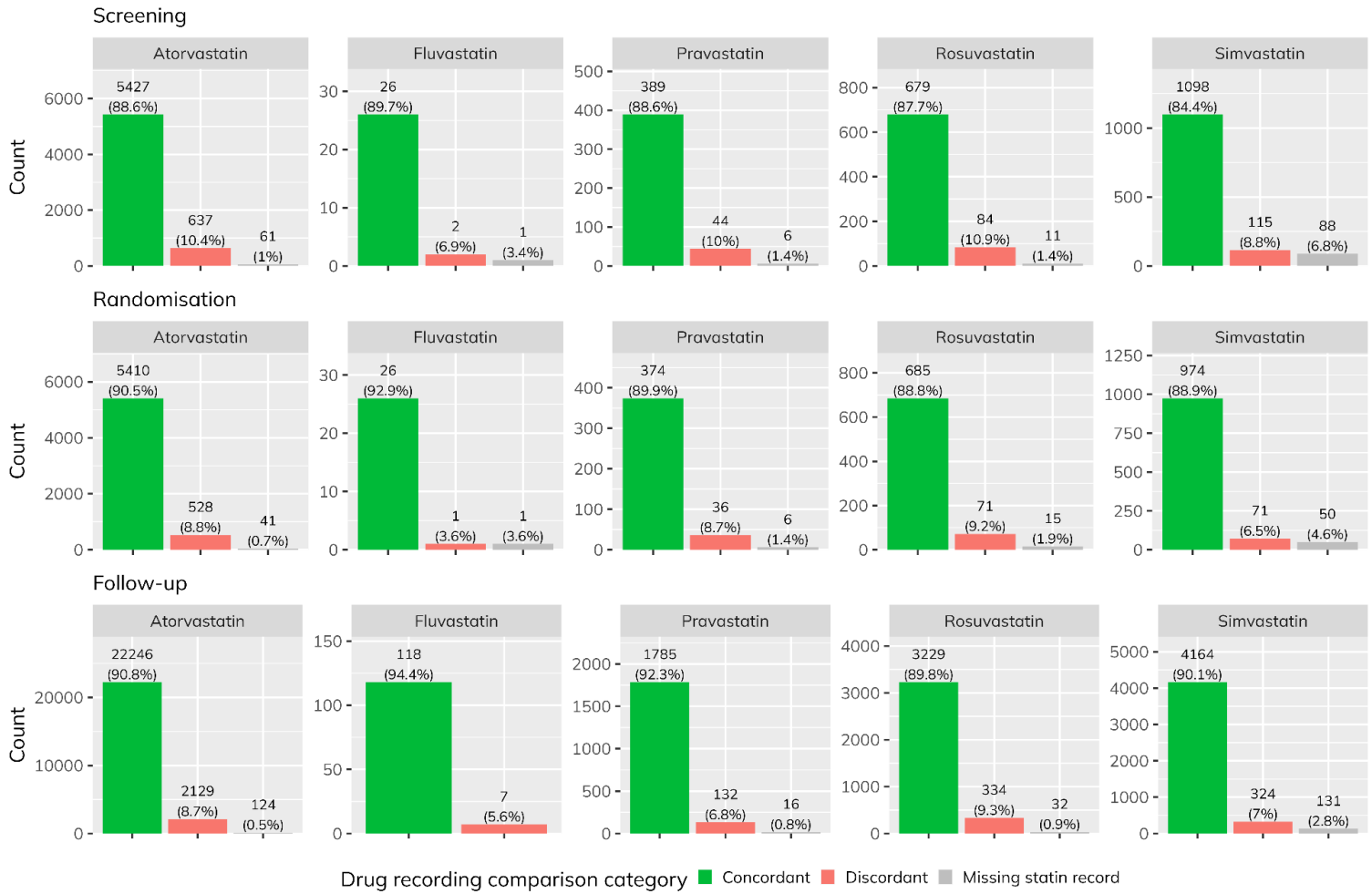


Figure legend: The vertical axis shows participant counts for screening and randomisation visits, and participant-visit pairs for follow-up visits. Statin dose from RCD data extracted as the strength of the statin dispensed in the month of the corresponding visit, or the most recent record before that.

6.3.8 Agreement for drug exposure between Dispensing data and case report form (CRF)

6.3.8.1 Binary drug exposure at each visit

Table 6.4 shows aggregate agreement between RCD and CRF for drug exposure captured at each study visit type, across all drug groups of interest combined. Combined agreement between sources was high across all study visits (κ=0.94 to 0.96).

Table 6.4 - Agreement between Dispensing data and case report form (CRF) for drug exposure at each visit type, across all drug groups

Visit	Participant-drug pairs				Agreement metrics				
	Both sources	CRF only	RCD only	Neither source	Sn	Sp	PPV	NPV	κ (95% CI)
Screening (statins and ezetimibe only)	8,831	217	186	8,848	97.6	97.9	97.9	97.6	0.96 (0.95 : 0.96)
Randomisation	36,017	1,397	2,779	209,121	96.3	98.7	92.8	99.3	0.94 (0.93 : 0.94)
Follow-up (records at any point during follow-up)	35,737	922	2,392	183,294	97.5	98.7	93.7	99.5	0.95 (0.94 : 0.95)
Follow-up (records at each individual visit)	160,061*	8,204*	8,904*	938,984*	95.1	99.1	94.7	99.1	0.94 (0.94 : 0.94)

Table legend: Counts shown in this row are for participant-drug-visit pairs

Tables 6.5-6.8 show agreement by individual drug group for each visit type. At screening, agreement was high for statins and ezetimibe, albeit with a moderate NPV of 73.2% for statins and PPV of 79% for ezetimibe. At randomisation, agreement was generally high for all medications (κ≥0.80; sensitivity, specificity, and predictive values >85%), except for resins and bempedoic acid (with small numbers each), and some diabetes medications and diuretics (mostly associated with lower PPVs of approximately 65-85%). Of note, agreement between RCD and CRF improved from screening to randomisation for both statins and ezetimibe (κ: 0.80 to 0.89 and 0.86 to 0.95, respectively), with increases in specificity (91.6% to 93.2%) and NPV (73.2% to

88.2%) for statins, and sensitivity (95.9% to 97.2%), specificity (98.4% to 99.5%), and PPV (79% to 93.3%) for ezetimibe. Agreement across the entire follow-up period was also lower for diabetes medications and diuretics (due to lower PPVs of approximately 50-85%), as well as inclisiran, PCSK9 inhibitors, and icosapent ethyl (with both low sensitivities and low PPBs, but small numbers). Agreement during follow-up was slightly lower when comparisons were made for each visit separately (versus entire follow-up period) due to lower sensitivity across most drugs, but with similar specificity, PPVs, and NPVs.

Table 6.5 - Agreement for drug exposure at screening between Dispensing data and case report form (CRF), by drug group

Drug group	Participant counts				Agreement metrics				
	Both sources	CRF only	RCD only	Neither source	Sn	Sp	PPV	NPV	κ (95% CI)
Statins	8,317	195	49	532	97.7	91.6	99.4	73.2	0.80 (0.77 : 0.82)
Ezetimibe	514	22	137	8,316	95.9	98.4	79	99.7	0.86 (0.83 : 0.88)

1 Table 6.6 - Agreement for drug exposure at randomisation between Dispensing data and case report form (CRF), by drug group

Drug group	Participant counts				Agreement metrics				
	Both sources	CRF only	RCD only	Neither source	Sn	Sp	PPV	NPV	κ (95% CI)
Statins	7,977	178	98	1,336	97.8	93.2	98.8	88.2	0.89 (0.88 : 0.90)
Ezetimibe	627	18	45	8,899	97.2	99.5	93.3	99.8	0.95 (0.94 : 0.96)
PCSK9 inhibitors	0	0	0	9,589	N/A	100	N/A	100	N/A
Inclisiran	0	0	0	9,589	N/A	100	N/A	100	N/A
Bempedoic acid	2	0	2	9,585	100	100	50	100	0.67 (0.2 : 1.00)
Icosapent ethyl	0	0	0	9,589	N/A	100	N/A	100	N/A
Fibrates	165	6	12	9,406	96.5	99.9	93.2	99.9	0.95 (0.92 : 0.97)
Resins	15	3	6	9,565	83.3	99.9	71.4	100	0.77 (0.62 : 0.92)
Insulin	357	22	86	9,124	94.2	99.1	80.6	99.8	0.86 (0.84 : 0.89)
Metformin	1,038	35	140	8,376	96.7	98.4	88.1	99.6	0.91 (0.90 : 0.92)
Sulphonylureas	273	10	37	9,269	96.5	99.6	88.1	99.9	0.92 (0.89 : 0.94)
Thiazolideniones	26	2	7	9,554	92.9	99.9	78.8	100	0.85 (0.76 : 0.95)
DPP-4 inhibitors	234	10	72	9,273	95.9	99.2	76.5	99.9	0.85 (0.81 : 0.88)
GLP-1 agonists	103	2	59	9,425	98.1	99.4	63.6	100	0.77 (0.71 : 0.83)
SGLT2 inhibitors	265	5	60	9,259	98.1	99.4	81.5	99.9	0.89 (0.86 : 0.91)
Acarbose or similar	2	0	1	9,586	100	100	66.7	100	0.80 (0.41 : 1.00)
ACE inhibitors	4,521	109	281	4,678	97.6	94.3	94.1	97.7	0.92 (0.91 : 0.93)
Angiotensin-receptor blockers	1,956	35	208	7,390	98.2	97.3	90.4	99.5	0.93 (0.92 : 0.93)
Mineralocorticoid receptor antagonists	550	16	81	8,942	97.2	99.1	87.2	99.8	0.91 (0.90 : 0.93)
Beta-adrenergic blockers	5,224	100	352	3,913	98.1	91.7	93.7	97.5	0.90 (0.90 : 0.91)
Calcium-channel blockers	2,405	53	330	6,801	97.8	95.4	87.9	99.2	0.90 (0.89 : 0.91)
Loop diuretics	881	42	164	8,502	95.4	98.1	84.3	99.5	0.88 (0.87 : 0.90)
Thiazide diuretics	557	56	110	8,866	90.9	98.8	83.5	99.4	0.86 (0.84 : 0.88)
Aspirin	5,064	579	275	3,671	89.7	93	94.8	86.4	0.82 (0.81 : 0.83)
Other antiplatelets	2,332	83	238	6,936	96.6	96.7	90.7	98.8	0.91 (0.90 : 0.92)
Vitamin K antagonists	384	14	31	9,160	96.5	99.7	92.5	99.8	0.94 (0.93 : 0.96)
NOACs	1,059	19	84	8,427	98.2	99	92.7	99.8	0.95 (0.94 : 0.96)

2

1 Table 6.7 - Agreement for drug exposure at any point during follow-up between Dispensing data and case report form (CRF), by drug
2 group

Drug group	Participant-visit counts				Agreement metrics				
	Both sources	CRF only	RCD only	Neither source	Sn	Sp	PPV	NPV	κ (95% CI)
Statins	7,062	61	76	1,036	99.1	93.2	98.9	94.4	0.93 (0.92 : 0.94)
Ezetimibe	671	13	63	7,488	98.1	99.2	91.4	99.8	0.94 (0.93 : 0.95)
PCSK9 inhibitors	0	11	0	8,224	0	100	N/A	99.9	N/A
Inclisiran	9	4	3	8,219	69.2	100	75	100	0.72 (0.51 : 0.93)
Bempedoic acid	5	0	2	8,228	100	100	71.4	100	0.83 (0.60 : 1.00)
Icosapent ethyl	2	0	0	8,233	100	100	100	100	1.00 (1.00 : 1.00)
Fibrates	169	5	10	8,051	97.1	99.9	94.4	99.9	0.96 (0.93 : 0.98)
Resins	17	3	5	8,210	85	99.9	77.3	100	0.81 (0.68 : 0.94)
Insulin	405	9	51	7,770	97.8	99.3	88.8	99.9	0.93 (0.91 : 0.95)
Metformin	1,126	9	126	6,974	99.2	98.2	89.9	99.9	0.93 (0.92 : 0.94)
Sulphonylureas	298	16	44	7,877	94.9	99.4	87.1	99.8	0.90 (0.88 : 0.93)
Thiazolideniones	25	2	11	8,197	92.6	99.9	69.4	100	0.79 (0.68 : 0.91)
DPP-4 inhibitors	304	4	69	7,858	98.7	99.1	81.5	99.9	0.89 (0.86 : 0.91)
GLP-1 agonists	158	6	67	8,004	96.3	99.2	70.2	99.9	0.81 (0.76 : 0.85)
SGLT2 inhibitors	444	9	156	7,626	98	98	74	99.9	0.83 (0.81 : 0.86)
Acarbose or similar	2	0	2	8,231	100	100	50	100	0.67 (0.20 : 1.00)
ACE inhibitors	4,231	53	152	3,799	98.8	96.2	96.5	98.6	0.95 (0.94 : 0.96)
Angiotensin-receptor blockers	2,006	12	137	6,080	99.4	97.8	93.6	99.8	0.95 (0.94 : 0.96)
Mineralocorticoid receptor antagonists	638	13	111	7,473	98	98.5	85.2	99.8	0.90 (0.89 : 0.92)
Beta-adrenergic blockers	4,952	48	188	3,047	99	94.2	96.3	98.4	0.94 (0.93 : 0.95)
Calcium-channel blockers	2,565	44	274	5,352	98.3	95.1	90.3	99.2	0.91 (0.90 : 0.92)
Loop diuretics	1,130	39	287	6,779	96.7	95.9	79.7	99.4	0.85 (0.83 : 0.87)
Thiazide diuretics	611	58	112	7,454	91.3	98.5	84.5	99.2	0.87 (0.85 : 0.89)
Aspirin	4,768	399	149	2,919	92.3	95.1	97	88	0.86 (0.85 : 0.87)
Other antiplatelets	2,420	75	157	5,583	97	97.3	93.9	98.7	0.93 (0.93 : 0.94)
Vitamin K antagonists	393	9	17	7,816	97.8	99.8	95.9	99.9	0.97 (0.95 : 0.98)
NOACs	1,326	20	123	6,766	98.5	98.2	91.5	99.7	0.94 (0.93 : 0.95)

3

1 Table 6.8 - Agreement for drug exposure at each follow-up visit between Dispensing data and case report form (CRF), by drug group

Drug group	Participant-visit counts				Agreement metrics				
	Both sources	CRF only	RCD only	Neither source	Sn	Sp	PPV	NPV	κ (95% CI)
Statins	33,420	904	380	6,635	97.4	94.6	98.9	88	0.89 (0.89 : 0.90)
Ezetimibe	2,998	136	151	38,054	95.7	99.6	95.2	99.6	0.95 (0.94 : 0.96)
PCSK9 inhibitors	0	34	0	41,305	0	100	N/A	99.9	N/A
Inclisiran	12	4	4	41,319	75	100	75	100	0.75 (0.58 : 0.92)
Bempedoic acid	6	4	4	41,325	60	100	60	100	0.60 (0.32 : 0.88)
Icosapent ethyl	2	0	0	41,337	100	100	100	100	1.00 (1.00 : 1.00)
Fibrates	796	54	50	40,439	93.6	99.9	94.1	99.9	0.94 (0.93 : 0.95)
Resins	66	16	21	41,236	80.5	99.9	75.9	100	0.78 (0.71 : 0.85)
Insulin	1,776	90	251	39,222	95.2	99.4	87.6	99.8	0.91 (0.90 : 0.92)
Metformin	4,844	173	438	35,884	96.6	98.8	91.7	99.5	0.93 (0.93 : 0.94)
Sulphonylureas	1,175	128	184	39,852	90.2	99.5	86.5	99.7	0.88 (0.87 : 0.89)
Thiazolideniones	101	29	43	41,166	77.7	99.9	70.1	99.9	0.74 (0.68 : 0.80)
DPP-4 inhibitors	1,114	116	270	39,839	90.6	99.3	80.5	99.7	0.85 (0.83 : 0.86)
GLP-1 agonists	543	54	253	40,489	91	99.4	68.2	99.9	0.78 (0.75 : 0.80)
SGLT2 inhibitors	1,232	48	402	39,657	96.2	99	75.4	99.9	0.84 (0.83 : 0.85)
Acarbose or similar	11	2	3	41,323	84.6	100	78.6	100	0.81 (0.65 : 0.98)
ACE inhibitors	19,493	586	685	20,575	97.1	96.8	96.6	97.2	0.94 (0.94 : 0.94)
Angiotensin-receptor blockers	9,061	230	605	31,443	97.5	98.1	93.7	99.3	0.94 (0.94 : 0.95)
Mineralocorticoid receptor antagonists	2,571	110	327	38,331	95.9	99.2	88.7	99.7	0.92 (0.91 : 0.92)
Beta-adrenergic blockers	23,115	538	885	16,801	97.7	95	96.3	96.9	0.93 (0.93 : 0.93)
Calcium-channel blockers	11,321	554	1,038	28,426	95.3	96.5	91.6	98.1	0.91 (0.90 : 0.91)
Loop diuretics	4,494	405	778	35,662	91.7	97.9	85.2	98.9	0.87 (0.86 : 0.87)
Thiazide diuretics	2,647	384	371	37,937	87.3	99	87.7	99	0.87 (0.86 : 0.87)
Aspirin	21,784	2,787	656	16,112	88.7	96.1	97.1	85.3	0.83 (0.83 : 0.84)
Other antiplatelets	10,398	582	630	29,729	94.7	97.9	94.3	98.1	0.92 (0.92 : 0.93)
Vitamin K antagonists	1,682	88	88	39,481	95	99.8	95	99.8	0.95 (0.94 : 0.96)
NOACs	5,399	148	387	35,405	97.3	98.9	93.3	99.6	0.95 (0.94 : 0.95)

6.3.8.2 Time to drug initiation after randomisation

Estimates for time to drug initiation after randomisation using Dispensing data were strongly associated with those from the CRF (ICC: 0.92, 95% CI: 0.91 - 0.93). Dispensing data would generally anticipate drug recording in the CRF across all drugs, by an average of 41.6 days - Figure 6.11 and Table 6.9 (or 116.9 days if using the date of the first visit with a new drug recorded; data not shown).

Figure 6.11 – Individual estimates of time to drug initiation after randomisation according to Dispensing data and case report form (CRF), by drug

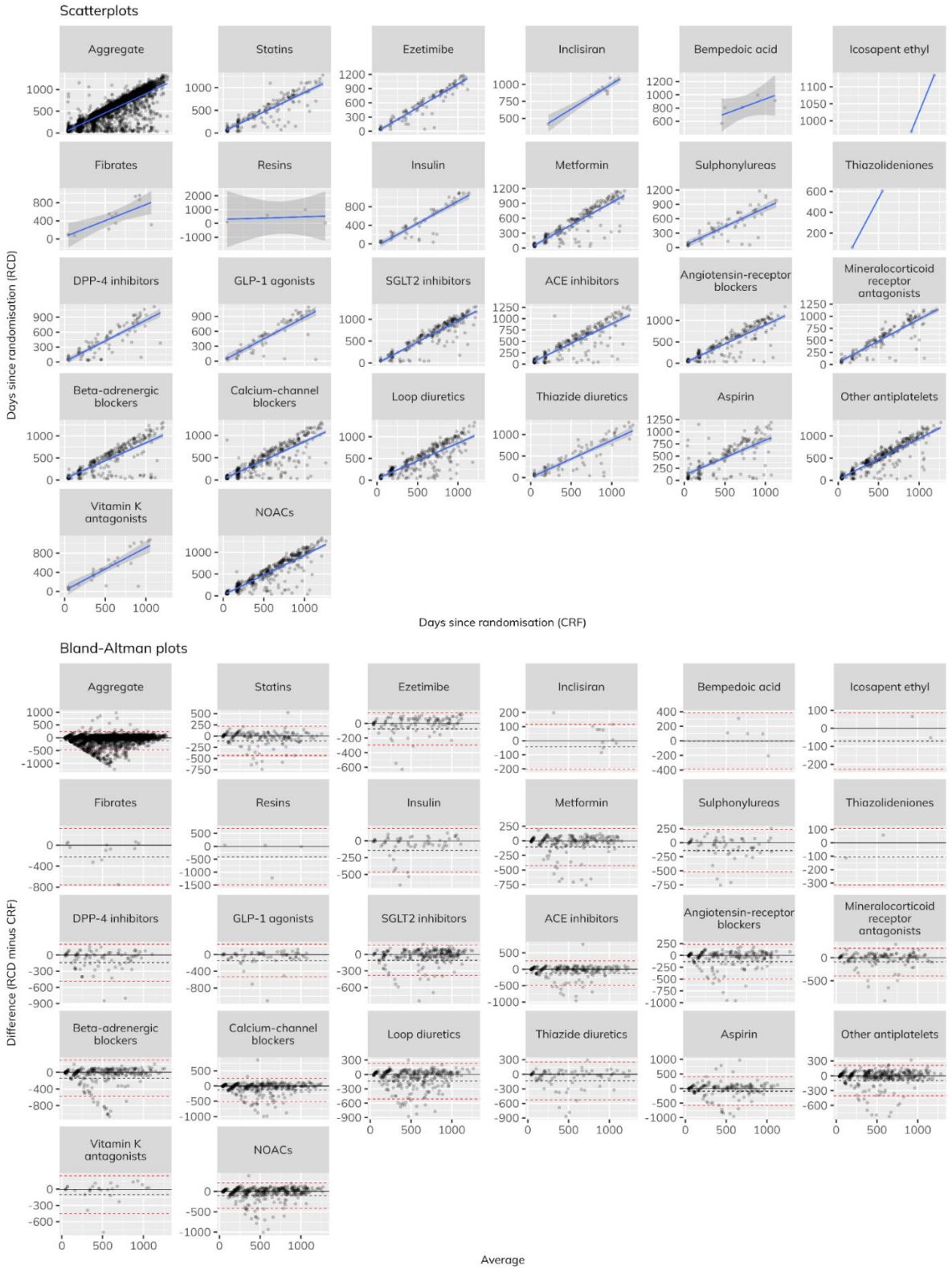


Figure legend: Each dot represents one participant. Scatterplots include linear regression lines along with associated 95% CI margins. Bland-Altman plots include horizontal dashed lines for mean difference (black) and 95% LOA (red). CRF dates were imputed as the mid-point between the first follow-up visit with a new drug recorded and the preceding visit. Dispensing record dates imputed as day 15 of each month. Average calculated as arithmetic mean.

1 Table 6.9 - Agreement estimates for time to drug initiation after randomisation, by drug

Drug	Participant counts	ICC (95% CI)	Average difference (95% LOA)[1]	
Aggregate	3,672[2]	0.92 (0.91 : 0.93)	-41.6	(-387.3 : 304.1)
Statins	174	0.94 (0.91 : 0.95)	-36.7	(-349.2 : 275.9)
Ezetimibe	112	0.97 (0.96 : 0.98)	-2.7	(-224.9 : 219.5)
Inclisiran	11	0.95 (0.81 : 0.99)	44.5	(-119.1 : 208)
Bempedoic acid	5	0.81 (0.45 : 0.98)	80.6	(-278 : 439.2)
Icosapent ethyl	2	0.96 (0.08 : 1.00)	6.5	(-155.7 : 168.7)
Fibrates	12	0.84 (0.46 : 0.95)	-119.9	(-583 : 343.2)
Resins	4	0.36 (0.00 : 0.96)	-291.8	(-1516.2 : 932.7)
Insulin	59	0.93 (0.87 : 0.96)	-61.4	(-366.3 : 243.5)
Metformin	202	0.94 (0.92 : 0.95)	-33.9	(-336.9 : 269)
Sulphonylureas	61	0.89 (0.81 : 0.94)	-64.1	(-430.3 : 302.1)
Thiazolidieniones	2	0.98 (0.72 : 1.00)	-27.5	(-261.7 : 206.7)
DPP-4 inhibitors	89	0.90 (0.81 : 0.94)	-75.8	(-423.1 : 271.5)
GLP-1 agonists	64	0.90 (0.83 : 0.94)	-63.7	(-435.3 : 308)
SGLT2 inhibitors	242	0.95 (0.93 : 0.96)	-26.4	(-295.9 : 243)
ACE inhibitors	222	0.92 (0.89 : 0.94)	-45.6	(-398.6 : 307.5)
Angiotensin-receptor blockers	208	0.93 (0.90 : 0.95)	-52.3	(-387.9 : 283.3)
Mineralocorticoid receptor antagonists	145	0.94 (0.91 : 0.95)	-18.1	(-322.2 : 285.9)
Beta-adrenergic blockers	300	0.88 (0.83 : 0.91)	-60.8	(-477.8 : 356.1)
Calcium-channel blockers	363	0.91 (0.87 : 0.93)	-58	(-424.7 : 308.8)
Loop diuretics	316	0.89 (0.85 : 0.92)	-65	(-428.4 : 298.4)
Thiazide diuretics	79	0.91 (0.85 : 0.95)	-71.6	(-463.7 : 320.6)
Aspirin	170	0.83 (0.76 : 0.87)	-23.6	(-513.6 : 466.3)
Other antiplatelets	381	0.94 (0.93 : 0.96)	-19.5	(-304.1 : 265)
Vitamin K antagonists	28	0.91 (0.81 : 0.96)	-32.1	(-395.2 : 331.1)
NOACs	421	0.94 (0.93 : 0.95)	-26.6	(-333.6 : 280.4)

Table legend: [1]Differences calculated as days since randomisation according to the RCD minus days since randomisation according to the CRF (i.e. negative values denote longer times to drug recording in the CRF). [2]Participant-drug pairs (instead of participants) CRF dates imputed as the mid-point between the first follow-up visit with a new drug recorded and the previous visit. Dispensing dates imputed as day 15 of each record month.

2

6.3.9 Use-case scenarios

RCD would generally provide similar or slightly higher absolute overall exposure estimates for all drugs (absolute differences <3%). However, there were large relative differences for some less commonly used drugs (e.g. 54%, 25%, and 20% relative increase in overall prevalence of exposure at randomisation to GLP-1 agonists, DPP-4 inhibitors, and SGLT2 inhibitors, respectively) - Figure 6.12 and Table 6.10. The only two drug groups for which exposure would be underestimated were aspirin (absolute 3% and relative 5% underestimation at both randomisation and follow-up), and statins (absolute 1.4% and 0.8% and relative 1.7% and 1% underestimation at screening and randomisation, respectively).

Figure 6.12 – Absolute differences in aggregate drug exposure estimates between routinely-collected data (RCD) and case report form (CRF), by study visit and drug group

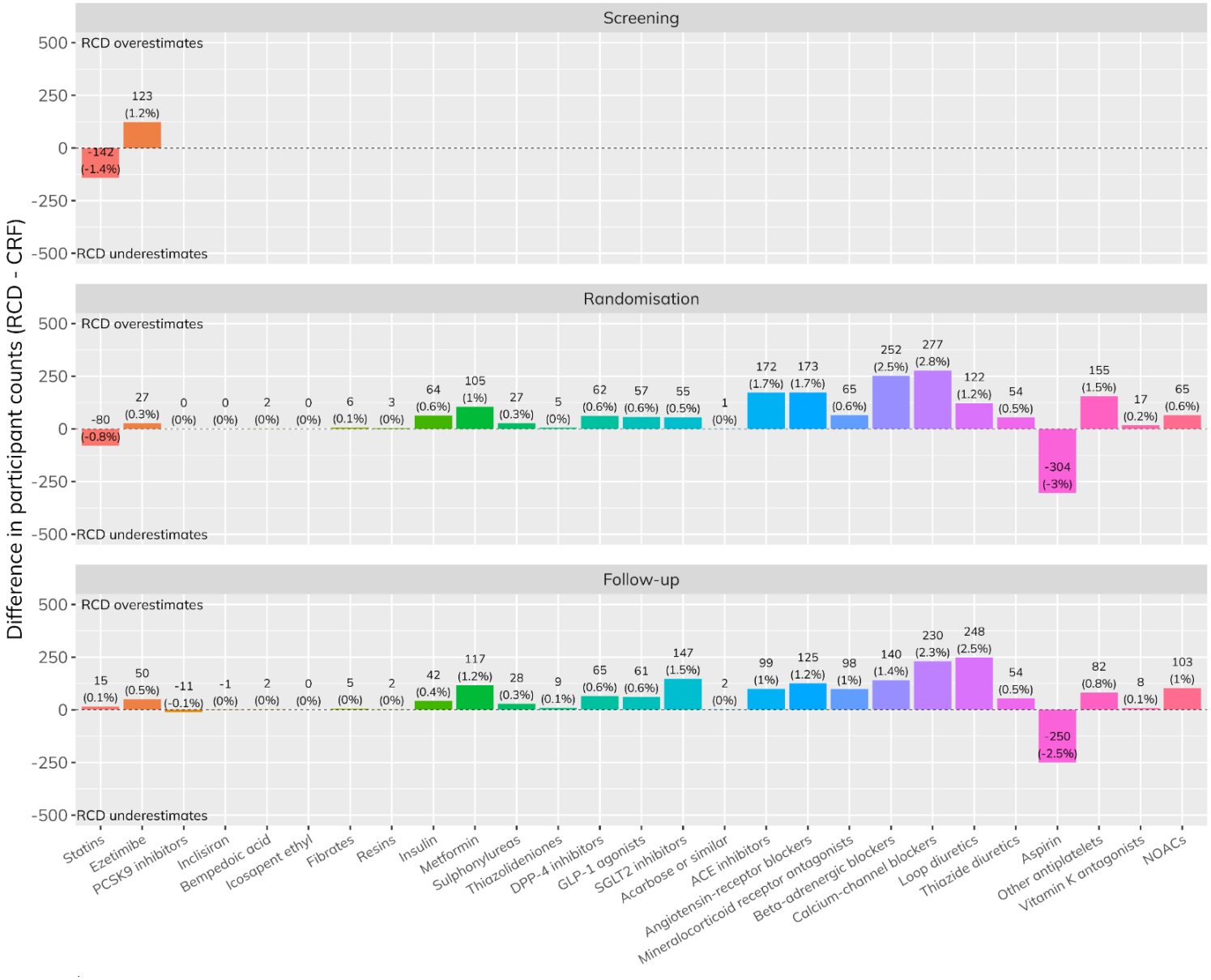


Figure legend: Each panel depicts one visit type (screening, randomisation, or follow-up). Bars above the dashed line represent overestimation by RCD (i.e. more participants identified as exposed vs CRF; e.g. SGLT2 inhibitors), and bars beneath the dashed line represent underestimation (e.g. statins, aspirin). Text labels depict absolute differences in participant counts and proportions (calculated as RCD minus CRF). Only statins and ezetimibe were reliably recorded at the screening visit. Underlying counts, as well as proportions and relative differences, are shown in Table 6.10 in the next page.

Table 6.10 - Differences in aggregate drug exposure between routinely-collected data (RCD) and case report form (CRF), by study visit and drug group

Drug group	Visit	CRF		RCD		Difference in days (RCD-CRF)	Absolute difference	Relative difference
		Counts	%	Counts	%			
Statins	Screening	8,372	83.6%	8,230	82.1%	-142	-1.42%	-1.7%
	Randomisation	8,155	81.4%	8,075	80.6%	-80	-0.8%	-0.98%
	Follow-up	7,123	71.1%	7,138	71.2%	15	0.15%	0.21%
Ezetimibe	Screening	500	5%	623	6.2%	123	1.23%	24.6%
	Randomisation	645	6.4%	672	6.7%	27	0.27%	4.19%
	Follow-up	684	6.8%	734	7.3%	50	0.5%	7.31%
PCSK9 inhibitors	Randomisation	0	0%	0	0%	0	0%	N/A
	Follow-up	11	0.1%	0	0%	-11	-0.11%	-100%
Inclisiran	Randomisation	0	0%	0	0%	0	0%	N/A
	Follow-up	13	0.1%	12	0.1%	-1	-0.01%	-7.69%
Bempedoic acid	Randomisation	2	0%	4	0%	2	0.02%	100%
	Follow-up	5	0%	7	0.1%	2	0.02%	40%
Icosapent ethyl	Randomisation	0	0%	0	0%	0	0%	N/A
	Follow-up	2	0%	2	0%	0	0%	0%
Fibrates	Randomisation	171	1.7%	177	1.8%	6	0.06%	3.51%
	Follow-up	174	1.7%	179	1.8%	5	0.05%	2.87%
Resins	Randomisation	18	0.2%	21	0.2%	3	0.03%	16.67%
	Follow-up	20	0.2%	22	0.2%	2	0.02%	10%
Insulin	Randomisation	379	3.8%	443	4.4%	64	0.64%	16.89%
	Follow-up	414	4.1%	456	4.6%	42	0.42%	10.14%
Metformin	Randomisation	1,073	10.7%	1,178	11.8%	105	1.05%	9.79%
	Follow-up	1,135	11.3%	1,252	12.5%	117	1.17%	10.31%
Sulphonylureas	Randomisation	283	2.8%	310	3.1%	27	0.27%	9.54%
	Follow-up	314	3.1%	342	3.4%	28	0.28%	8.92%
Thiazolideniones	Randomisation	28	0.3%	33	0.3%	5	0.05%	17.86%
	Follow-up	27	0.3%	36	0.4%	9	0.09%	33.33%
DPP-4 inhibitors	Randomisation	244	2.4%	306	3.1%	62	0.62%	25.41%
	Follow-up	308	3.1%	373	3.7%	65	0.65%	21.1%
GLP-1 agonists	Randomisation	105	1%	162	1.6%	57	0.57%	54.29%
	Follow-up	164	1.6%	225	2.2%	61	0.61%	37.2%
SGLT2 inhibitors	Randomisation	270	2.7%	325	3.2%	55	0.55%	20.37%
	Follow-up	453	4.5%	600	6%	147	1.47%	32.45%
Acarbose or similar	Randomisation	2	0%	3	0%	1	0.01%	50%
	Follow-up	2	0%	4	0%	2	0.02%	100%
ACE inhibitors	Randomisation	4,630	46.2%	4,802	47.9%	172	1.72%	3.71%
	Follow-up	4,284	42.8%	4,383	43.7%	99	0.99%	2.31%
Angiotensin-receptor blockers	Randomisation	1,991	19.9%	2,164	21.6%	173	1.73%	8.69%
	Follow-up	2,018	20.1%	2,143	21.4%	125	1.25%	6.19%
Mineralocorticoid receptor antagonists	Randomisation	566	5.6%	631	6.3%	65	0.65%	11.48%
	Follow-up	651	6.5%	749	7.5%	98	0.98%	15.05%
Beta-adrenergic blockers	Randomisation	5,324	53.1%	5,576	55.6%	252	2.51%	4.73%
	Follow-up	5,000	49.9%	5,140	51.3%	140	1.4%	2.8%
Calcium-channel blockers	Randomisation	2,458	24.5%	2,735	27.3%	277	2.76%	11.27%
	Follow-up	2,609	26%	2,839	28.3%	230	2.3%	8.82%
Loop diuretics	Randomisation	923	9.2%	1,045	10.4%	122	1.22%	13.22%
	Follow-up	1,169	11.7%	1,417	14.1%	248	2.48%	21.21%
Thiazide diuretics	Randomisation	613	6.1%	667	6.7%	54	0.54%	8.81%
	Follow-up	669	6.7%	723	7.2%	54	0.54%	8.07%
Aspirin	Randomisation	5,643	56.3%	5,339	53.3%	-304	-3.03%	-5.39%
	Follow-up	5,167	51.6%	4,917	49.1%	-250	-2.5%	-4.84%
Other antiplatelets	Randomisation	2,415	24.1%	2,570	25.6%	155	1.55%	6.42%
	Follow-up	2,495	24.9%	2,577	25.7%	82	0.82%	3.29%
Vitamin K antagonists	Randomisation	398	4%	415	4.1%	17	0.17%	4.27%
	Follow-up	402	4%	410	4.1%	8	0.08%	1.99%
NOACs	Randomisation	1,078	10.8%	1,143	11.4%	65	0.65%	6.03%
	Follow-up	1,346	13.4%	1,449	14.5%	103	1.03%	7.65%

Dispensing data provided little value in describing previous statin exposure for participants not taking a statin at screening, with most participants not having any relevant historic statin record (Figure 6.13). There were also similar proportions of participants having one (10-20%) or more than one (1-5%) previous distinct statin records in Dispensing data regardless of the reason provided on the CRF for not taking a statin. Of note, historic Dispensing data was only available from April 2018 onwards, while the average time between the index cardiovascular event and inclusion in the trial for each participant is 7-10 years (data not shown).

Figure 6.13 - Comparison of statin intolerance characterisation between routinely-collected data (RCD) and screening case report form (CRF)

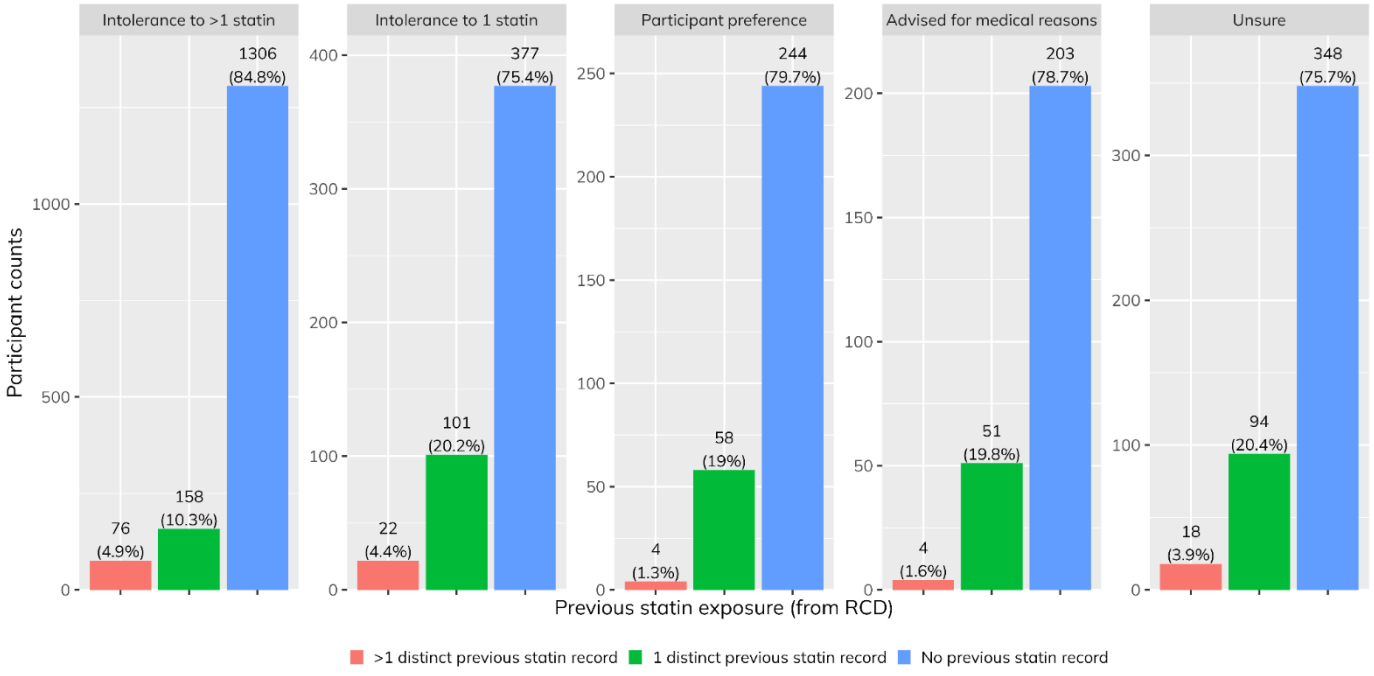


Figure legend: Each panel shows records for a distinct reason for not being on a statin as recorded in the screening CRF. Within each panel, the color coded bars and text labels show distributions of participants with different historic statin exposure patterns as identified from RCD.

Dispensing data would increase the overall proportion of participants classified as being on high-intensity statin (51% to 53%) or no treatment (11% to 16%) at randomisation when compared with the CRF, and decrease those on moderate (31% to 26%) and low intensity treatment (7% to 5%) - Figure 6.14. At an individual participant level, RCD provided the same estimate as the CRF in 85.4% of participants (8,559/10,020), and led to downgrading of treatment intensity category for 889 (8.9%) of participants and upgrading for 572 (5.7%).

Figure 6.14 - Comparison of statin treatment intensity exposure at randomisation between routinely-collected data (RCD) and case report form (CRF)

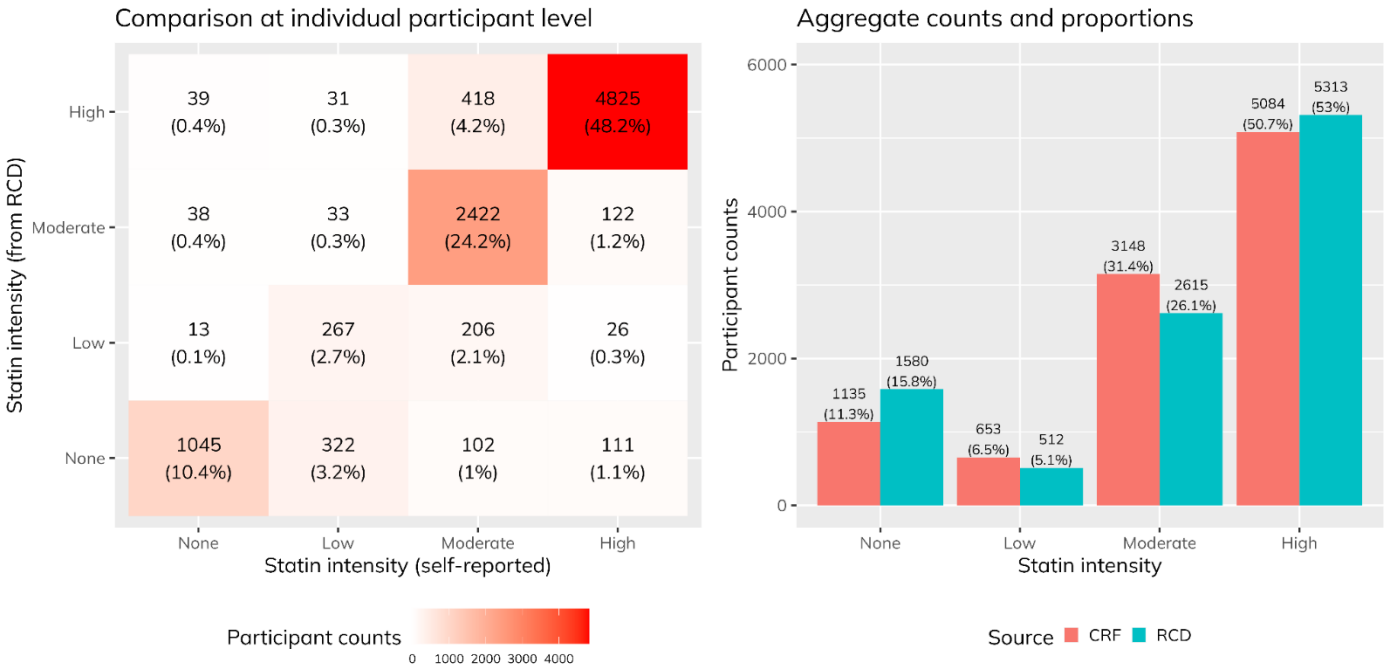


Figure legend: Statin intensity extracted from RCD as the strength of the statin dispensed in the month of randomisation, or that from the most recent previous record.
Statin treatment intensity defined as follows:

- *high intensity: atorvastatin ≥40mg or rosuvastatin ≥20mg*
- *moderate intensity: atorvastatin 10-39 mg, rosuvastatin 5-19 mg, simvastatin ≥20mg, pravastatin ≥40mg, lovastatin ≥40mg, fluvastatin 80mg, pitavastatin ≥1mg*
- *low-intensity: any other lower dose statin regimen*

6.4 Discussion

Randomised trials frequently use self-reported medication lists to collect information on concomitant treatments. This approach may be limited by either failed recall of relevant medications, or inaccurate reporting of medications no longer being used, and is labour intensive (particularly if more details such as dose or length of use are required). Accordingly, objective sources of information on drug exposure (such as electronic monitors, or RCD from insurance claims, pharmacy records, or prescription registries) are generally considered superior and frequently used as a reference to validate self-report medications in medication adherence and pharmacoepidemiological studies. Nonetheless, the use of RCD on medications in RCTs has been limited, both globally and in the UK. The recent release of individual level, nationwide dispensing data from NHSE for research purposes offers an opportunity to tackle these challenge, but its implementation within trials requires appropriate validation and development.

Together with the analyses presented in chapter 3 for the RECOVERY trial of inpatient COVID-19 treatments, this chapter describes the first use of the Medicines Dispensed in the Community (Dispensing) dataset within an RCT, applied here to a large phase-III trial testing a cholesterol lowering medication in the outpatient setting. Dispensing data captured 99.8% of the intended population of ORION-4 participants recruited in England. There were generally high levels of agreement with self-reported medications recorded in the study CRFs across multiple cardiovascular drugs, leading to mostly similar estimates of overall drug exposure prevalence. However, Dispensing data failed to adequately capture exposure to two prohibited medications, provided limited value in characterising historic statin exposure, and resulted in moderate changes to baseline statin treatment intensity estimates. Taken together, these results suggest that Dispensing data is an accurate and complete source of drug exposure in the community setting for most

cardiovascular medications. However, these data may need to be complemented with self-reporting in some particular cases.

Dispensing data captured almost all ORION-4 participants recruited in England (10,020/10,038). This likely reflects the unified nature of the NHS in England, the compulsory nature of dispensing record submission for reimbursement to pharmacies, and the focus on cardiovascular medications generally managed in the community setting. Participants not included in this dataset had many medications recorded in the study CRF which should have been captured by Dispensing data collection, but there were no obvious geographical or temporal gaps. While it is not clear why these participants were not captured, they may possibly collect their medications from dispensing sources not captured in the Dispensing dataset (such as hospital pharmacies). Nonetheless, the very high capture levels strengthen the use of Dispensing data in this setting. The 3-month lag in data dissemination seen here, albeit short when compared with other RCD sources,[247] hinders the use of Dispensing data for real-time assessments of medication exposure (similarly to the Scottish dispensing data assessed in the previous chapter). This is particularly the case when information on medication exposure is required to determine study eligibility, and for those that may pose a safety risk (namely prohibited medications, such as PCSK9 inhibitors and non-study inclisiran). Lack of coverage of these two medications likely results from the need for subcutaneous administration, and management mostly or partly in secondary care. While it is possible that the dissemination lag here may be change in future data extracts, capture of these prohibited drugs in the Dispensing data is unlikely to improve unless their management is shifted towards primary care.

To implement Dispensing data, the performance of different possible algorithms to define cross-sectional drug exposure was tested against CRF records at randomisation. This

exercise showed that agreement was maximized by using a 3-6 month lookback window (κ=0.93), and further improved slightly if also including records in the month of randomisation (κ=0.94). This window aligns with the underlying average dispensing intervals of 30 days found for most medications, and replicates results from the previous chapters 3 (Dispensing dataset in the RECOVERY trial) and 5 (Scottish dispensing data in the LENS trial), as well as reports from other countries.[258,318,327,328] Nonetheless, the performance of different possible approaches will depend on the setting considered, highlighting the importance of appropriate, specific algorithm development and testing.[164,258,326,329–331]

In the absence of exact Dispensing dates, the impact of three different imputation approaches was assessed by comparing resulting estimates of time to medication initiation from randomisation with those obtained from the date of the first visit in which the corresponding medication was recorded in the CRF, all yielding similarly high levels of agreement. However, it is likely that the new medication would in fact have been started somewhere between the visit when it was first recorded and the previous visit. Accordingly, agreement between Dispensing and CRF estimates improved if the CRF date was imputed as the mid-point between the two visits, suggesting that this is the best approach to extract exact medication initiation dates from CRF records (with no further differences in agreement yielded from different Dispensing date imputation approaches).

Besides fact of medication use, daily dose may also be an important aspect of medication use in some trial settings. This information is frequently missing from RCD sources such as the Dispensing dataset, thus requiring use of assumptions which may greatly influence estimates of drug adherence or exposure.[205,266,332–334] CRF records were used here to validate extraction of daily statin dose from the dispensed strength, showing that daily doses could be reliably extracted from Dispensing for most participants (~85-90%),

similarly to previous reports.[335] Given the importance of statin dose in the ORION-4 trial, CRF records may still be considered the preferred source of such information. However these results generally support extracting statin dose from Dispensing data if it is impractical to request that information from participants (e.g. linkage-based post-trial follow-up, or observational population-level analyses).

Using the best performing algorithm identified above, aggregate agreement between RCD and CRF was very high at all study visits (restricted to statins and ezetimibe at screening). Agreement between RCD and CRF for statins and ezetimibe improved from screening to randomisation, due to improvements in sensitivity, specificity, PPV, and NPV. Drug exposure derivation from Dispensing data was performed using the same algorithm for all visits. If assuming that drug capture in the Dispensing dataset is consistent over time, and that the randomisation CRF is likely to be more accurate than the screening CRF (given a stronger focus on recording all medications, and a higher familiarity of participants with the study at this point), this suggests that improved agreement between sources from screening to randomisation is due to improved CRF recording accuracy.

Medications with lowest agreement were those used to treat diabetes and diuretic medications, mostly due to lower PPVs (<85%). This pattern may result from frequent switches or stopping of these medications in response to fluctuations in glycaemic control or fluid status (so that they are captured in the preceding 3-month interval in Dispensing dataset but recorded as stopped in the CRF), and the low overall prevalence of some of these drugs (which negatively impacts PPV).

When considering estimation of time to drug initiation after randomisation, Dispensing data provided very similar average estimates to those from the CRF, as evidenced by generally high ICCs. Dispensing records anticipated those from the CRF by 41 days on average (although the relatively small number of participants included in this analysis

resulted in wide 95% LOAs). This overall estimate is consistent with the likely event sequence (i.e. a medication is first dispensed to a participant from a pharmacy, and only sometime after recorded at a study visit) and the monthly Dispensing data recording schedule. Interestingly, Dispensing data anticipated CRF records for all drugs, except inclisiran, bempedoic acid, and icosapent ethyl, which are more likely to be managed in secondary care (and thus not captured as accurately in Dispensing data).

Finally, the use-case scenarios tested showed that Dispensing data provided very little useful information to characterise historic statin exposure for participants not on a statin at screening. This is possibly due to the relatively recent data collection span (from April 2018 onwards), highlighting a potential use-case where these data may not be helpful. By contrast, Dispensing data would provide similar estimates of overall drug exposure when compared with the CRF for most medications at randomisation and during follow-up, but generally overestimating exposure to anti-hypertensive and antidiabetic drugs and underestimating for aspirin (the latter likely the result of over-the-counter dispensing) and statins. While absolute differences were small, relative differences could be large for some less commonly used antidiabetic drugs. Similarly, Dispensing data yielded overall similar estimates of statin treatment intensity category as CRF data, but with a moderate degree of reclassification at an individual level. While not identical, these results suggest that Dispensing data may be a reasonable alternative (or complement) to CRF data collection for uses where some degree of potential misclassification may be tolerable, such as characterising the study population overall (particularly if accepting that CRF drug recording may also not be perfectly accurate). Nonetheless, the issues identified with data dissemination timeliness, lack of capture of some prohibited drugs, and potential misclassification of drug exposure (even if in small numbers) imply that Dispensing data should not be used for circumstances that require recording of drug exposure in real-time

or where small errors may be consequential (such as for immediate eligibility or safety assessments), or for analyses produced at a relatively rapid pace (such as monitoring baseline population characteristics as trial recruitment progresses).

Taken together, the high agreement observed between RCD and CRF records across a diverse range of drug exposures and time points suggest that Dispensing data is a highly accurate and complete source of information on exposure to most concomitant cardiovascular medications of interest to the ORION-4 trial. Simultaneously, the high agreement found helps to reaffirm the accuracy of medication recording in the study CRF, through triangulation with a collateral, objective data collection method. However, some potential limitations of Dispensing data were found, namely the relatively long time to data dissemination and lack of adequate capture of some prohibited drugs. Similarly to previous chapters, these results support the use of a hybrid data collection approach, wherein RCD are used to extract information for most medications, and complemented with drug recording by study staff at each visit for specific drugs which may be time-critical or considered high-risk. If implemented in future trials, this approach could reduce the time and cost required to collect accurate information on concomitant drug exposure, allowing research staff to focus on accurately capturing other information which may be more critical for study quality. Such an approach would simultaneously help to foster *quality-by-design* principles[51,322,336] while potentially improving overall medication data accuracy.

6.4.1 Limitations

The results presented here have a number of important limitations. First, this analysis is restricted to ORION-4 participants recruited in England, as no other source of RCD on medications was available within this study. It is possible that data quality assessments for other datasets might yield distinct results. However, the assessment for Scottish dispensing data performed in the previous chapter produced similar conclusions. Second,

the analyses presented here were restricted to cardiovascular medications of interest to ORION-4. Although this included a large number of drugs, these results may not translate to other types of medications (e.g. oncology medications managed primarily in secondary care, or those dispensed mostly over-the-counter). Third, self-reported medications (as captured in the study CRF) were used here as reference to both develop and validate the use of Dispensing data. However, self-reported medication use is a subjective measure and unlikely to be a perfect estimate of drug exposure. Conversely, Dispensing data capture submission of medication dispensing for reimbursement purposes, rather than actual medication intake, and are therefore also not a perfect source of drug exposure. Therefore, direct comparisons of Dispensing data and CRF records should be made with caution, as both sources may capture different aspects of the same underlying true exposure. The lack of exact dispensing event dates in the data currently available also raises some methodological challenges with handling records occurring close to a particular time point of interest (such as randomisation). However, this can be circumvented using the algorithmic approaches described, and, according to recent communication from NHSE, exact dates may be provided in the near future. Finally, the main agreement metrics used here were Cohen's κ coefficients, as they provide an overall measure of agreement that takes into account the agreement expected by chance alone. As discussed in previous chapters, κ coefficients have a number of limitations, and their interpretation may be challenging. For this reason, alternative measures of agreement (sensitivity, specificity, and PPV/NPV) were used to both complement κ coefficients and investigate underlying reasons for disagreement.

6.5 Conclusion

Dispensing data is a highly complete and accurate source of information on exposure to most cardiovascular medications of interest in the ORION-4 trial when compared with

self-reported medication lists. A hybrid approach that combines both data collection methods for different purposes may be used in similar trials to maximise both data quality and cost-efficiency.

CHAPTER 7: Overall discussion

7.1 Chapter summary

Results from this thesis show that RCD on medications in England and Scotland are a valuable and reliable source of information on drug exposure in the community setting for trials in cardiovascular disease and COVID-19. Repurposing these data may help streamline future trials and facilitate development and deployment of new therapies. Further research is now needed to expand these findings to broader trial populations and medications. The experience gained from working with these datasets may also help data controllers improve the quality of the documentation and data provided, facilitating expanded use of these resources. Full realisation of the potential of these data sources to advance clinical research and improve public health will require collaborative efforts between medical professionals, regulators, and government.

7.2 Introduction

RCTs are the gold-standard for the reliable assessment of benefit and harm of medical interventions.[1,2] However, trials have become increasingly costly and complex as a result of more intricate study designs, excessively restrictive eligibility criteria, unnecessary bureaucracy, inefficient data collection and/or verification, and the impact of declining mortality and morbidity rates on trial size and length.[13–15] Use of RCD has been frequently suggested as a potential solution to this problem, given their similarity to that collected in trials. However, widespread use of RCD within observational studies, even when complex statistical adjustment approaches are employed, is not a substitute for the conduct of well-designed and conducted trials.[17,26] Instead, RCD should be integrated within RCTs so that more, better, and faster trials are undertaken, and the necessary robust evidence is generated.[32–34,36]

For this to happen, RCD first needs to be carefully validated to ensure that they are of adequate quality to be repurposed within trials. In the UK, a number of RCD sources have been available for research purposes for some years,[33,61,64] and occasionally used in trials.[61,62] However, sources of RCD on medications suitable for linkage to randomised trial cohorts have only become available in England since 2020.[158,230,337] A similar dataset has been available in Scotland since 2009, but has only been used in two RCTs.[166,167]

The aims of this thesis were to:

1) develop the methodology required to implement 4 different sources of RCD on medication use in the UK to identify drug exposures of interest within 3 distinct, large randomised trials in COVID-19 (in England) and cardiovascular disease (in England and Scotland);

2) ascertain whether those RCD can adequately capture the intended populations and drug exposures of interest in each case; and

3) quantify the impact and test the feasibility of using these data for specific use-case scenarios of direct relevance to each trial.

7.3 Main findings

7.3.1 Previous studies

The scoping literature review performed retrieved a large number of studies comparing RCD with other sources of information on medication exposure, namely 18 within RCTs and 174 within observational studies. Identified studies spanned a wide range of geographies, study populations, and medication groups. Multiple types of RCD sources, drug exposure outcomes, and agreement metrics were used. Importantly, RCD were the reference most commonly used to validate other sources (rather than the opposite), followed by studies where both RCD and the collateral source were considered similar. Agreement results were generally higher or similar for comparisons conducted within trials

versus observational studies. Most studies had positive agreement results (as interpreted by the individual study authors), but with large variability in interpretations of similar estimates across studies.

This review concluded that a substantial amount of research had already been performed comparing RCD on medications with other sources in a multitude of settings. The studies identified revealed promising overall results in terms of agreement between RCD and other sources, while generally illustrating the value that multiple investigators see in the use of RCD for clinical research (including as a reference standard used to validate other sources). This review also identified important heterogeneity in the methodological approaches used, and interpretation of the results obtained. Overall, these results highlight the importance of validating the use of RCD on medications in the specific contexts in which they may be incorporated.

7.3.2 Data on medication use in the community within the RECOVERY trial

Chapter 3 assessed the potential value of incorporating two distinct novel RCD sources in England to identify drug exposure for participants recruited in England within the RECOVERY trial of treatments of acute COVID-19. These were 1) the General Practice Extraction Service Data for Pandemic Planning and Research (GDPPR), or "GP" dataset, which is extracted from primary care records, and 2) the Medicines Dispensed in the Community, or "Dispensing" dataset, generated from claims for reimbursement for medications dispensed in the community. These data were compared with those captured in the trial CRF completed upon randomisation (used as reference).

Both datasets captured a substantial proportion of the population considered, and most of a broad range of 17 drug groups of interest to the trial (including cardiovascular, respiratory, immunosuppressive, and mental health medications), with Dispensing data

capturing a wider range of medications and more participants than GP data. Dispensing data were more stable across consecutive extracts, but there was a 2-month lag between an event and its inclusion in the disseminated data (contrary to fortnightly dissemination in the GP dataset, but with inferior data stability). While some discrepancies between RCD and the CRF would be expected (as the two RCD sources capture medication use in the community, while the CRF captured medication use upon study entry while in-hospital), the overall agreement observed between RCD and CRF was high. However, RCD generated lower overall estimates of drug exposure at randomisation for antiplatelets and oral anticoagulation versus the CRF. When compared with the use of Dispensing data only, the addition of GP data only slightly improved agreement for drug exposure at randomisation with the CRF. Overall estimates of drug exposure at randomisation and initiation after randomisation were similar between both RCD sources across most medications studied.

Taken together, these results show that, albeit not perfect substitutes to CRF data collection in-hospital, the RCD datasets assessed here may be considered reliable sources of information on drug exposure in the community for trials conducted in the acute setting. Expanded use of these data could allow more comprehensive characterisation of the characteristics and outcomes of trial participants (both before and after the acute illness period which the study is focused on), without unnecessarily complicating trial data collection within busy clinical settings (such as those in which RECOVERY took place). These data sources may still need to be complemented with existing data collection methods for some particular drug exposures of vital importance, such as those required to immediately determine eligibility to the trial intervention upon randomisation.

7.3.3 Inpatient medication data in the RECOVERY trial

Chapter 4 tested the use of RCD on inpatient medications within the RECOVERY trial using data from a subset of hospital trusts in England, as captured in the Electronic Prescribing and Medicines Administration (EPMA) dataset. These data were compared with those obtained from the RECOVERY trial CRFs at randomisation and follow-up (used as reference).

EPMA data captured a representative sample of approximately 10% of RECOVERY participants recruited in England. Due to distinct data reporting schedules between contributing hospitals, a two-month lag between the end of the analysis period and data reception was necessary. Agreement between EPMA and the CRF was high for drug exposure at randomisation and during follow-up across all drugs considered, with both sources generating mostly similar overall drug exposure estimates. The underlying reasons for discrepancies between EPMA and CRF varied based on the specific drug considered. Specifically, these were mostly due to low sensitivity overall (resulting from poor capture of medication use in intensive care units), but low specificity for corticosteroids (possibly as a result of automated electronic prescriptions being created for COVID-19 patients upon admission, but later not administered). Agreement for estimates of quantitative drug exposure during follow-up was modest.

Overall, these results show that the EPMA dataset may be a reasonable source of information on inpatient medication use in England at the population level, or for research at an individual patient level if some degree of misclassification is acceptable. However, the inclusion of a small subset of hospitals, data dissemination lags, poor coverage of prescriptions within intensive care units, and issues with suboptimal specificity limit their use as the main or single source to determine inpatient drug exposure in trials.

7.3.4 Community dispensing data in Scotland in the LENS trial

Chapter 5 validated the use of RCD on medication dispensing in the community setting in Scotland for participants in the LENS trial of fenofibrate for prevention of diabetic retinopathy progression, using the Prescribing Information Services (PIS) dataset.

Almost all LENS participants were captured in PIS data, and there was very high agreement for drug exposure at randomisation between PIS and the LENS CRF for the 18 cardiovascular medications selected for analysis. There was a 3-month lag between the occurrence of an event and it being included in the data available for analysis. Both sources yielded similar overall proportions of exposed to each drug of interest at randomisation, and of participants exposed to some prohibited drugs after randomisation. PIS data also agreed with CRF records for daily statin dose for 98-100% of participants (depending on the specific statin). Finally, PIS data complemented CRF data collection by capturing changes occurring post-randomisation to non-prohibited drugs recorded in the randomisation CRF (as the follow-up CRFs only collected information on prohibited drugs). Overall, results from this chapter support the expanded use of PIS data to extract information on exposure to cardiovascular drugs in future trials in Scotland, combined with real-time collection of information on prohibited medications from study participants in a hybrid approach.

7.3.5 Community dispensing data in England in the ORION-4 trial

Chapter 6 investigated the potential role of using RCD on medication dispensing in the community setting in England (from the Dispensing dataset, also used in Chapter 3) within the ORION-4 trial of inclisiran for secondary cardiovascular disease prevention. Differently from previous chapters, inclisiran was still an unlicensed medication upon launch of the trial.

Almost all participants recruited in England were included in Dispensing data, which had a 3 month lag between medication dispensing events and their inclusion in the disseminated dataset. Overall agreement between the study CRF and Dispensing data was high at all study visits, with both sources yielding generally similar overall estimates of drug exposure for most drugs (with small absolute differences). However, Dispensing data failed to adequately capture exposure to a small number of participants using two concomitant medications prohibited by the trial protocol (inclisiran and PCSK9 inhibitors), but these were not expected to be captured in Dispensing data as they are usually managed outside the community setting. Agreement for time to drug initiation after randomisation was also high. Dispensing data and the CRF agreed on daily statin dose for most participants, and generated broadly similar estimates of statin treatment intensity exposure at randomisation.

Results from this chapter support using Dispensing data to identify exposure to a wide range of cardiovascular medications, including in a trial of an unlicensed medication. However, given the data dissemination lags and incomplete capture of prohibited medications, these drugs should continue to be recorded in real-time using existing data collection pathways. Similarly to previous chapters, these findings suggest that both data collection methods can be combined in a hybrid model that harnesses the convenient, comprehensive, and accurate nature of the RCD sources investigated, while safeguarding prompt and complete capture of exposure to specific drugs of paramount importance.

7.4 Strengths and limitations

The analyses presented in this thesis have several strengths. First, they were focused on nationwide RCD sources which were either released recently or have had little use within trials. Thus, the results of the validation exercises performed are both novel and highly relevant to support their implementation within current and future RCTs, but also for the

broader UK clinical research ecosystem. Of note, all trials considered here were still ongoing at the time this work was carried out, demonstrating the feasibility of rapid evaluation and implementation of these data sources during the main trial lifecycle. Secondly, this thesis assessed a diverse range of RCD sources, focused on medication use in the community (in England and Scotland) and in-hospital (in England), and also within a diverse range of clinical trial settings (including trials in cardiovascular disease and COVID-19 populations, conducted in the acute and the community setting, and studying either new or well-established drugs). For each trial, rather than focusing on any particular drug, a large selection of drugs of potential interest was selected for assessment in collaboration with the trial investigators to ensure comprehensive validation of any potential benefits or pitfalls of using RCD. These also spanned non-study medications and medications being studied in the trial (e.g. corticosteroids and tocilizumab in RECOVERY). Assessment of such a large number of drugs was also facilitated by the large nature of the trials considered, with this thesis including head-to-head comparisons between RCD and trial data on >47,000 distinct individuals.

Importantly, validation of the multiple RCD sources tested was performed using a comprehensive, multi-dimensional approach, developed based on the literature on data quality assessments and relevant recommendations issued from regulatory bodies in the US and the UK, specifically focused on the use of RCD within trials. Given the novelty of the datasets addressed, and the lack of any specific guidance on how to use them to identify drug exposure, an agnostic data-driven approach was employed across all chapters to identify the best methodological approach for this purpose (in comparison with the existing data collection procedure in each trial, considered as the optimal reference). This approach avoided the need for assumptions about underlying medication use patterns, and how best to replicate those in the RCD sources. At the same time, the testing

performed allowed simplification of some proposed derivation procedures, by showing that combination of multiple datasets or data fields within the same dataset did not make a material difference in the estimates obtained (as shown in Chapters 3 and 4). Similar results were found for some of the proposed missing data imputation approaches.

The drug exposure endpoints selected for RCD validation were based on those being collected in each trial, which included not only fact of drug exposure, but also quantity, dose, and timing. The broadly positive agreement results found for these parameters highlight the potential value of using RCD for more detailed drug exposure assessments, and particularly those for which existing data collection methods (such as self-report) may be troublesome or time-consuming. Finally, the multiple use-case scenarios tested helped demonstrate the feasibility of incorporating the RCD sources assessed in concrete analyses of direct relevance to each trial, and the potential pitfalls of doing so.

Some limitations should also be discussed. First, the datasets assessed here are restricted to the UK (specifically England and Scotland), and the underlying data collection pathways are highly variable from one RCD resource to another. Therefore, the validation results obtained should not be directly translated to other RCD sources, either within or outside the UK. Importantly, two of the data sources considered here (primary care records in the GP dataset, and secondary care medications in the EPMA dataset) are currently only available for COVID-19 research. Moreover, EPMA data collection only covers a small sample of hospitals in England. NHSE have nonetheless signalled that both these constraints might change in the future.[158,238] Results on data dissemination timeliness should be seen as depicting the experience with the particular data extracts used, rather than a fixed property of the underlying data source (as this may vary over time for the same dataset, for example due to changes within data processing pipelines).

The datasets assessed in this thesis were validated within trials in cardiovascular disease (in the community setting) and COVID-19 (in the acute setting), and restricted to specific drugs. Extrapolation of these results to trials focused on other study populations or medications should be performed with caution. Ideally, use of these RCD sources to identify other drug exposures should follow similar development and validation pipelines as the ones used here. These included preliminary data quality and coverage assessments, exploratory data analysis to identify relevant underlying medication use patterns, identification and testing of candidate algorithms, and quantification of the impact of using these new data sources vs existing methods.

The validation exercises presented across this thesis were performed against data generated by existing procedures in each trial (i.e. medical records review in RECOVERY, and self-report in LENS and ORION-4). While these references may be considered both optimal and the most relevant comparators in each trial, it is recognised that they may not provide a perfect depiction of the underlying, true medication exposure which they aim to capture. It follows that optimising drug exposure algorithms in order to maximise agreement between RCD and the reference source may, in fact, lead to estimates from RCD better representing incorrect records occurring in the reference data, rather than true drug exposure. Nonetheless, given their use as standard data collection methods in each trial, and in the absence of alternative ways to verify CRF data (e.g. individual review of medical records for all participants), CRF records were therefore chosen in this thesis as the best possible reference standard.

Finally, agreement between sources was primarily assessed in this thesis using κ coefficients. As shown in the literature review, these are widely used in studies comparing RCD against other sources, facilitating contextualisation of these results. It is nevertheless acknowledged that κ coefficients were not developed for the purpose of comparing two

data sources, that they may exhibit paradoxical behaviours in particular settings, and that their interpretation may be challenging. For that reason, complementary statistics were presented alongside κ coefficients to aid interpretation, and to investigate underlying reasons for disagreement between sources.

7.5 Implications for future clinical trials

Results from this thesis highlight the value of using RCD on medications within trials in the UK. In particular, the Dispensing and PIS datasets are currently available for trials in England and Scotland and 1) have published robust internal data quality assessments,[162,239] 2) provide excellent coverage of individuals with acute COVID-19, established cardiovascular disease, and diabetic retinopathy, 3) have short data dissemination lags (for extracts assessed during the course of this thesis), and 4) were shown to be highly complete and accurate to identify exposure to a wide range of cardiovascular medications when compared with self-reports (including for statin doses). Conversely, the high agreement observed between trial data collection methods (self-reported or medical records review) and a centralised, objective source of drug exposure (such as RCD) also helps to strengthen confidence in the existing trial procedures (particularly for self-reports when no confirmation is sought from medical or prescription records). Overall, it should be reasonable to extrapolate the results of the validation exercises performed here for trials focused in the same populations, medications, and clinical settings. The datasets evaluated may then help streamline data collection for most cardiovascular medications, provided that the limitations identified are acknowledged and addressed (namely ensuring alternative procedures to promptly collect information on prohibited medications, or others of particular or time-critical importance). Future (and ongoing) cardiovascular trials in the UK should therefore strongly consider applying for

linkage to these datasets, bearing in mind that such applications may be costly and time-consuming.[33,61]

For this purpose, trialists need first to ensure that consent materials foresee the possibility of data linkage, and appropriately explain what this involves. It should also be acknowledged that different methods of collecting information on drug exposure may provide complementary information about the underlying true drug exposure, so that the optimal approach may indeed involve triangulation between distinct sources (e.g. using RCD to capture exposure to most prescription medications, and self-report to capture prohibited drugs, quantify adherence, and characterise historic exposure or reasons for non-adherence). Importantly, linkage requests should include provision of exact medication event dates wherever possible, as lack of this information generates additional hurdles at the analysis stage (as shown in Chapters 3, 5, and 6). Trialists should also request a reasonable length of historic data preceding inclusion in the trial to allow adequate characterisation of both past and current exposure (and address this in consent materials). This may help improve identification of medication use upon study entry, particularly if average dispensing intervals are long (such as for Scottish dispensing data in Chapter 5, with the LENS trial only requesting linked data from screening onwards). Collection of historic data may also help better explore patterns of past medication use, which was not possible for participants not on statin in ORION-4 in Chapter 6 (likely due to the recency of the Dispensing data collection). Trialists aiming to validate the use of RCD on medications in other settings should consider what is the specific purpose of using RCD in their trial, the impact that misclassification might have in their context, and therefore whether an acceptable error margin can be identified and what that margin may be (as it is unlikely that RCD will perfectly capture drug exposure in all cases, and the impact of discrepancies will differ across specific scenarios). For example, the acceptable

degree of misclassification of drug exposure to concomitant drugs not known to pose particular safety concerns should be much higher than for those in which such concerns exist (e.g. warfarin in LENS, and PCSK9 inhibitors in ORION-4, as further discussed in Chapters 5-6). Future trials aiming to use RCD as the sole or primary method to identify drug exposure should ensure that adequate validation is performed before the start of their trial, as regulatory authorities will require reassurance that the proposed data collection approach does not jeopardise participant safety or data integrity.[24,36] Alternatively, preliminary validation may be performed at the beginning of the trial, with subsequent results informing changes to the study data collection procedures. Finally, trialists should take into consideration that, while use of RCD may reduce overall trial costs (due to less time spent manually collecting data), it nevertheless requires redistribution of resources within the trial towards data management and analysis. This includes ensuring that appropriate infrastructure and personnel are available, and should be carefully considered a priori given the need for (and potential challenges in) recruitment of highly-specialised staff.

Analysts making use of these data sources to identify similar drug exposures may repurpose the optimal derivation algorithms identified here. Of note, the development exercises performed all suggested that a 3-month lookback period could be considered optimal to identify baseline drug exposure upon study entry, using either English or Scottish data, and for both COVID-19 and cardiovascular disease populations. However, different approaches may be needed for other drugs or populations (e.g. oncological medications, topical formulations, or inhalers). Trials interested in other drug exposures or focused in other populations or clinical settings may also repurpose the comprehensive methodological development pipeline used here. Importantly, users of these or similar RCD sources should carefully consider the implications of specific features of each dataset in

their analysis pipelines and for their intended uses. For example, data provided in real or near real time may not become stable for a period of months. The specific clinical care pathways underlying data collection for each resource should also be well understood. This is particularly the case for prescribing and/or management of certain medications in primary vs secondary care (and subsequent capture in the RCD collections), so that absence of a relevant drug record may not necessarily mean that that drug is not being used. Although RCD may be provided in multiple data tables, and dozens of possibly relevant data fields, it is likely that the information required to adequately characterise drug exposure could be extracted using only two or three variables (as shown here). Nonetheless, appropriate validation of the proposed derivation procedures should be performed a priori, particularly as the most obvious choices may not always prove to be the most accurate (as exemplified with inpatient EMPA data in Chapter 4). Finally, procedures used to derive drug exposure should be accurately and publicly reported to allow inspection, replication, and repurposing where appropriate.

7.6 Implications for future research

This thesis has focused on a limited number of medications of interest to trials in cardiovascular disease and COVID-19. Future research should expand these analyses to other population settings and medications. It would also be helpful if other researchers could replicate this work in similar settings. At the same time, future work could explore the possibility of using information on drug exposure collected from these RCD sources to identify potential safety or efficacy events which may be missed by existing trial data collection procedures (e.g. start of donepezil, rivastigmine, or galantamine as a surrogate for dementia, of an antibacterial for infection, or of insulin or other hypoglycaemic drugs for new-onset diabetes).

While the main focus of this thesis was repurposing of RCD on medications within trials, these data may also be used to better investigate prescription patterns, medication adherence, clinical care pathways, and healthcare associated costs, particularly when linked to other RCD sources at an individual level. Importantly, the validation work performed here will help to strengthen confidence in the use of these data by researchers in settings where alternative data collection methods (such as self-report) would be impractical. Alternatively, these RCD may help develop and validate self-reported measures of medication exposure and adherence (as described in Chapter 2).

7.7 Implications for data controllers and other stakeholders

This thesis made use of RCD sources generated as part of clinical care, and managed and provided by NHSE and Public Health Scotland. The way these organisations collect, maintain and supply these resources has important implications on their value for research purposes and the feasibility of harnessing them in that context.

For example, the GDPPR dataset (primary care records) could greatly expand the breadth of RCD currently available for use in trials in England. However, this data collection, derived from the long-standing GPES platform for planning and commissioning purposes, was only made available for COVID-19 research and is focused on a limited set of clinical features specified by NHSE, which greatly limit its potential value. The data provided also had important issues, identified as part of the exploratory analyses performed (not all shown in this thesis), including 1) suboptimal documentation (namely poor explanation of data provenance, data extraction timelines, and the expected use of each data field), 2) underlying clinical codes not correctly mapped (and therefore shown as missing), 3) variable uses of some data fields depending on the electronic GP record system provider,

4) no capture of medication data preceding 2018, 5) introduction of duplicates when patients moved residence or GP practice, or their practice implements a different electronic record system, 6) issues with data handling within internal NHSE systems (so that some data may be lost along the process), and 7) poor stability over consecutive extracts due to changes in the underlying data extraction procedures (as shown in Chapter 3). Moreover, the existing documentation is scattered across multiple locations within NHSE's website, and no updates were provided to users when either the documentation or the underlying data extraction processes changed. It is recognised that the GDPPR dataset was generated for the purpose of rapidly providing primary care data for use in COVID-19 research at the height of the initial stages of the pandemic, and that the data collected is influenced by how different electronic system providers store and process the information recorded by GPs. However, the challenges described greatly hindered expanded use of these data in this thesis (namely for its specified purpose of COVID-19 research). Therefore, these limitations should be addressed when, and if, a similar dataset becomes available for other research purposes in the future - an aspiration which has been repeatedly deferred.[238,338] These findings have been communicated to NHSE by the candidate, with advice subsequently received on how best to use these data.

The EPMA dataset (inpatient medications) would also be of great value in expanding the current landscape of RCD available for trials in England. However, as shown in Chapter 4 (and publicly disclosed by NHSE), the underlying data collection only spans a small subset of hospital trusts in England, and does not adequately capture medication administration in particular settings (such as intensive care and paediatric wards). This, together with the current restriction to COVID-19 research, also hinder broader use of EPMA data in trials. Similarly to the GDPPR dataset, there were some issues with the documentation provided for some fields, and provision of clinical codes which were not correctly mapped from the

underlying electronic record, thus complicating analysis and development procedures. These challenges have been communicated to NHSE by the candidate.

The Dispensing dataset was found to be highly stable, accurate, and complete to identify medication use in the community. Some challenges were identified nevertheless, namely the lack of exact event dates. This dataset also included a large number of fields (e.g. prescribed and paid versions for dozens of fields), once again with suboptimal accompanying documentation. These findings were communicated to NHSE, who provided further guidance for data analysis. NHSE have also since informed that they are working to provide specific dates as part of improvement work currently being undertaken. The candidate was later invited to provide detailed feedback for a user consultation performed by NHSE, with the aim of improving their internal processes and promoting expanded use of their datasets through repurposing as data products. Preliminary work has also began on a joint publication including the candidate and the Prescribing team at NHSE. This data resource profile will outline the provenance and characteristics of the Dispensing dataset and the validation results described here, fostering expanded use by other researchers (similarly to publications on other RCD sources).[162]

Similarly to the Dispensing dataset, Scottish PIS data were found to be highly accurate and complete to identify drug exposure, and a valuable resource for use in trials. Contrary to the datasets mentioned previously, PIS data are provided alongside detailed recommendations on how to best to use them to extract information on drug exposure, greatly facilitating the development and analysis processes. PIS data have been extensively used in research, including within trials.[163] However, members of the electronic Data Research and Innovation Service (eDRIS) team at Public Health Scotland informed the candidate (in personal communication) that they did not consider PIS data to be amenable for use in randomised trials, as it was a record of routine NHS care. Results from

this thesis show that this is not the case, and that PIS data may indeed be used more frequently in trials, including of pharmacological interventions. This interaction highlights the importance of involving data controllers in efforts towards implementing RCD in trials, and providing evidence of the great potential value of the resources they manage.

On a general note, it is recommended that NHSE improve the publicly-available documentation which accompanies their data products (namely to explain data provenance, transformations to the data performed at each processing step, and the exact meaning of each data field). They should also ensure that this information is provided to users upon data dissemination, along with any updates, for example by including such documents in a release package alongside the dataset itself. NHSE may also wish to provide general guidance on how best to handle the data they publish, with future recommendations potentially informed by the work presented here. Importantly, the combination of better documentation and specific guidance would prevent researchers having to request data that they do not need, thus contributing to data minimisation (in line with General Data Practice Regulation principles) and reducing data processing and storage costs for both NHSE and end users. Overall, implementation of these recommendations would help increase confidence in use of RCD, simplify and accelerate data analysis, and foster more replicable and translatable analysis pipelines.

Regulators should consider the potential value of incorporating RCD such as those assessed within this thesis in future clinical trial applications, particularly given the potential impact on reduced trial costs, length, and complexity, and thus faster and more efficient development and deployment of new medications.[48] While it should be acknowledged that RCD were not generated to address a particular research question, regulators should be open to their incorporation in trials, either to complement or entirely replace traditional data collection methods, provided that adequate validation has been

performed. The recent issue of guidance specifically focused on this issue by both the MHRA and the FDA,[24,36] and successful experiences of using RCD within trials in both countries,[41,51] suggest that the path is already being trodden.

Finally, for the potential value of harnessing RCD in research to fully materialise, public trust in the management and use of linked health data must be established and preserved. These efforts must encompass the NHS, medical societies, regulators, and government. Initiates such as Data Saves Lives[339], the Goldacre review,[340] and the UK Government Life Sciences Vision[341] are moving this agenda forward, particularly by capitalising on the heightened awareness of the importance of clinical data and research following the COVID-19 response. Alas, previous setbacks such as the abandoned care.data[338] project and the postponement of plans to facilitate collation and sharing of primary care data in 2021 (notably fuelled by criticism from medical professionals)[238] show that much still needs to be done.

7.8 Final summary

Results from this thesis show that RCD sources on medications within England and Scotland are a feasible, accurate, and complete method to identify drug exposure in the community setting for trials in cardiovascular disease and COVID-19. These findings support expanded use of these data in current and future trials, while providing a framework to guide validation studies in other settings. Integration of these data in RCTs, either to complement or replace existing data collection methods, could simultaneously help streamline trial procedures and improve data quality. Such a shift may bring about important benefits to trial participants, researchers, and society more broadly by allowing more trials to be conducted at greater speed and lower cost, thus making better use of finite resources. Findings from this thesis are being used to guide implementation of these RCD sources within the trials discussed, have helped to improve internal procedures at

NHSE, and will be made publicly available through publications in open-access, peer-reviewed journals.

References

1. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol*. 2011;64(4):383-394. doi:10.1016/j.jclinepi.2010.04.026

2. Concato J, Shah N, Horwitz RI. Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. *N Engl J Med*. 2000;342(25):1887-1892. doi:10.1056/NEJM200006223422507

3. Childs D, Kansagra S. 10 Health Advances That Changed the World. abc News. https://abcnews.go.com/Health/TenWays/story?id=3605442

4. Banerjee A, Duflo E, Goldberg N, et al. Development economics. A multifaceted program causes lasting progress for the very poor: evidence from six countries. *Science*. 2015;348(6236):1260799. doi:10.1126/science.1260799

5. Bothwell LE, Podolsky SH. The Emergence of the Randomized, Controlled Trial. *N Engl J Med*. 2016;375(6):501-504. doi:10.1056/NEJMp1604635

6. Otte WM, Vinkers CH, Habets PC, van IJzendoorn DGP, Tijdink JK. Analysis of 567,758 randomized controlled trials published over 30 years reveals trends in phrases used to discuss results that do not reach statistical significance. Boutron I, ed. *PLOS Biol*. 2022;20(2):e3001562. doi:10.1371/journal.pbio.3001562

7. Getz KA, Campo RA. Trends in clinical trial design complexity. *Nat Rev Drug Discov*. 2017;16(5):307-307. doi:10.1038/nrd.2017.65

8. Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? *Nat Rev Drug Discov*. 2017;16(6):381-382. doi:10.1038/nrd.2017.70

9. Lauer MS, Gordon D, Wei G, Pearson G. Efficient design of clinical trials and epidemiological research: is it possible? *Nat Rev Cardiol*. 2017;14(8):493-501. doi:10.1038/nrcardio.2017.60

10. Eapen ZJ, Vavalle JP, Granger CB, Harrington RA, Peterson ED, Califf RM. Rescuing clinical trials in the United States and beyond: A call for action. *Am Heart J*. 2013;165(6):837-847. doi:10.1016/j.ahj.2013.02.003

11. Roth GA, Mensah GA, Johnson CO, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J Am Coll Cardiol*. 2020;76(25):2982-3021. doi:10.1016/j.jacc.2020.11.010

12. Global Burden of Disease 2019 Cancer Collaboration, Kocarnik JM, Compton K, et al. Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019. *JAMA Oncol*. 2022;8(3):420. doi:10.1001/jamaoncol.2021.6987

13. Sertkaya A, Wong HH, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. *Clin Trials*. 2016;13(2):117-126. doi:10.1177/1740774515625964

14. Califf RM. Clinical trials bureaucracy: unintended consequences of well-intentioned policy. *Clin Trials Lond Engl*. 2006;3(6):496-502. doi:10.1177/1740774506073173

15. Eisenstein EL, Collins R, Cracknell BS, et al. Sensible approaches for reducing clinical trial costs. *Clin Trials J Soc Clin Trials*. 2008;5(1):75-84. doi:10.1177/1740774507087551

16. Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. *The Lancet*. 2001;357(9253):373-380. doi:10.1016/S0140-6736(00)03651-5

17. Collins R, Bowman L, Landray M, Peto R. The Magic of Randomization versus the Myth of Real-World Evidence. *N Engl J Med*. 2020;382(7):674-678. doi:10.1056/NEJMsb1901642

18. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*. 2014;383(9913):267-276. doi:10.1016/S0140-6736(13)62228-X

19. ICH Harmonised Tripartite Guideline. ICH Guideline E3: Structure and Content of Clinical Study Reports. Published online November 30, 1995. https://database.ich.org/sites/default/files/E3_Guideline.pdf

20. Food and Drug Administration. FDA Guidance for Industry and Investigators: Safety Reporting Requirements for INDs and BA/BE Studies. Published online December 2012. https://www.fda.gov/media/79394/download

21. Food and Drug Administration. FDA Guidance for Industry: Premarketing Risk Assessment. Published online March 2005. https://www.fda.gov/media/71650/download

22. Monti S, Grosso V, Todoerti M, Caporali R. Randomized controlled trials and real-world data: differences and similarities to untangle literature data. *Rheumatology*. 2018;57(Supplement_7):vii54-vii58. doi:10.1093/rheumatology/key109

23. Getz KA. With clinical data, less is more. *Appl Clin Trials*. 2010;19:28-30.

24. Food and Drug Agency. FDA draft guidance for Industry: Real-World Data - Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. Published online September 2021.

25. Eapen ZJ, Lauer MS, Temple RJ. The Imperative of Overcoming Barriers to the Conduct of Large, Simple Trials. *JAMA*. 2014;311(14):1397. doi:10.1001/jama.2014.1030

26. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence — What Is It and What Can It Tell Us? *N Engl J Med*. 2016;375(23):2293-2297. doi:10.1056/NEJMsb1609216

27. Ramagopalan SV, Simpson A, Sammon C. Can real-world data really replace randomised clinical trials? *BMC Med*. 2020;18(1):13. doi:10.1186/s12916-019-1481-8

28. Ioannidis JPA. Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies. *JAMA*. 2001;286(7):821. doi:10.1001/jama.286.7.821

29. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol*. 2017;106(1):1-9. doi:10.1007/s00392-016-1025-6

30. Schaffner W. Surveillance of Antibiotic Prescribing in Office Practice. *Ann Intern Med*. 1978;89(5_Part_2):796. doi:10.7326/0003-4819-89-5-796

31. Adami HO. A paradise for epidemiologists? *The Lancet*. 1996;347(9001):588-589. doi:10.1016/S0140-6736(96)91279-9

32. Mc Cord KA, Al-Shahi Salman R, Treweek S, et al. Routinely collected data for randomized trials: promises, barriers, and implications. *Trials*. 2018;19(1):29. doi:10.1186/s13063-017-2394-5

33. Sydes MR, Barbachano Y, Bowman L, et al. Realising the full potential of data-enabled trials in the UK: a call for action. *BMJ Open*. 2021;11(6):e043906. doi:10.1136/bmjopen-2020-043906

34. Food and Drug Administration. Real-World Evidence. Published online July 16, 2021. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

35. European Union. The TRANSFoRm project. https://www.transformproject.eu/

36. Medicines & Healthcare products, Regulatory Agency, MHRA. MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions. Published online December 16, 2021. Accessed January 23, 2023. https://www.gov.uk/government/publications/mhra-guidance-on-the-use-of-real-world-data-in-clinical-studies-to-support-regulatory-decisions/mhra-guidance-on-the-use-of-real-world-data-in-clinical-studies-to-support-regulatory-decisions

37. Pharmaceuticals and Medical Devices Agency. Real World Data Working Group. Published 2023. https://www.pmda.go.jp/english/rs-sb-std/rs/0023.html

38. Pharamceutical Technology. FDA expands Ibrance's label to include male breast cancer patients. Published April 2019. Accessed July 27, 2020. https://www.pharmaceutical-technology.com/news/male-breast-cancer-ibrance-label-expansion/

39. Jenni Spinner. Medidata synthetic control arm lands FDA approval for cancer trial. Published November 19, 2020. https://www.outsourcing-pharma.com/Article/2020/11/19/Synthetic-control-arm-lands-FDA-approval-for-cancer-trial

40. IMplementation of an RCT to imProve Treatment With Oral AntiCoagulanTs in Patients With Atrial Fibrillation IMPACT-Afib (IMPACT-AFib). Published July 26, 2021. https://clinicaltrials.gov/ct2/show/NCT03259373.

41. Jones WS, Mulder H, Wruck LM, et al. Comparative Effectiveness of Aspirin Dosing in Cardiovascular Disease. *N Engl J Med*. 2021;384(21):1981-1990. doi:10.1056/NEJMoa2102137

42. De Moor G, Sundgren M, Kalra D, et al. Using electronic health records for clinical research: the case of the EHR4CR project. *J Biomed Inform*. 2015;53:162-173. doi:10.1016/j.jbi.2014.10.006

43. Academy of Managed Care Pharmacy null, American Pharmacists Association null. Medicare star ratings: stakeholder proceedings on community pharmacy and managed care partnerships in quality. *J Am Pharm Assoc JAPhA*. 2014;54(3):228-240. doi:10.1331/JAPhA.2014.13180

44. Luchenski SA, Reed JE, Marston C, Papoutsi C, Majeed A, Bell D. Patient and Public Views on Electronic Health Records and Their Uses in the United Kingdom: Cross-Sectional Survey. *J Med Internet Res*. 2013;15(8):e160. doi:10.2196/jmir.2701

45. Tully MP, Hassan L, Oswald M, Ainsworth J. Commercial use of health data-A public "trial" by citizens' jury. *Learn Health Syst*. 2019;3(4):e10200. doi:10.1002/lrh2.10200

46. da Silva MEM, Coeli CM, Ventura M, et al. Informed consent for record linkage: a systematic review. *J Med Ethics*. 2012;38(10):639-642. doi:10.1136/medethics-2011-100208

47. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Routinely collected data and comparative effectiveness evidence: promises and limitations. *Can Med Assoc J*. 2016;188(8):E158-E164. doi:10.1503/cmaj.150653

48. Llewellyn-Bennett R, Edwards D, Roberts N, Hainsworth AH, Bulbulia R, Bowman L. Post-trial follow-up methodology in large randomised controlled trials: a systematic review. *Trials*. 2018;19(1):298. doi:10.1186/s13063-018-2653-0

49. Mc Cord KA, Ewald H, Ladanie A, et al. Current use and costs of electronic health records for clinical trial research: a descriptive study. *CMAJ Open*. 2019;7(1):E23-E32. doi:10.9778/cmajo.20180096

50. Li G, Sajobi TT, Menon BK, et al. Registry-based randomized controlled trials- what are the advantages, challenges, and areas for future research? *J Clin Epidemiol*. 2016;80:16-24. doi:10.1016/j.jclinepi.2016.08.003

51. Pessoa-Amorim G, Campbell M, Fletcher L, et al. Making trials part of good clinical care: lessons from the RECOVERY trial. *Future Healthc J*. 2021;8(2):e243-e250. doi:10.7861/fhj.2021-0083

52. Zannad F, Pfeffer MA, Bhatt DL, et al. Streamlining cardiovascular clinical trials to improve efficiency and generalisability. *Heart*. 2017;103(15):1156-1162. doi:10.1136/heartjnl-2017-311191

53. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic Screening Improves Efficiency in Clinical Trial Recruitment. *J Am Med Inform Assoc*. 2009;16(6):869-873. doi:10.1197/jamia.M3119

54. Wachtell K, Lagerqvist B, Olivecrona GK, James SK, Fröbert O. Novel Trial Designs: Lessons Learned from Thrombus Aspiration During ST-Segment Elevation Myocardial Infarction in Scandinavia (TASTE) Trial. *Curr Cardiol Rep*. 2016;18(1):11. doi:10.1007/s11886-015-0677-6

55. Butala NM, Faridi KF, Secemsky EA, et al. Prognosis of Claims- Versus Trial-Based Ischemic and Bleeding Events Beyond 1 Year After Coronary Stenting. *J Am Heart Assoc*. 2021;10(6). doi:10.1161/JAHA.120.018744

56. Mueller S, Gottschalk F, Groth A, et al. Primary data, claims data, and linked data in observational research: the case of COPD in Germany. *Respir Res*. 2018;19(1):161. doi:10.1186/s12931-018-0865-1

57. Blake SR, Roome C, Shahzad A, et al. A comparison of hospital episode statistics and traditional methods to identify outcomes in a randomized trial; a sub-study of HEAT-PPCI. *J Public Health*. Published online January 4, 2019. doi:10.1093/pubmed/fdy225

58. Strom JB, Tamez H, Zhao Y, et al. Validating the use of registries and claims data to support randomized trials: Rationale and design of the Extending Trial-Based Evaluations of Medical Therapies Using Novel Sources of Data (EXTEND) Study. *Am Heart J*. 2019;212:64-71. doi:10.1016/j.ahj.2019.02.007

59. Rogers JR, Lee J, Zhou Z, Cheung YK, Hripcsak G, Weng C. Contemporary use of real-world data for clinical trial conduct in the United States: a scoping review. *J Am Med Inform Assoc*. 2021;28(1):144-154. doi:10.1093/jamia/ocaa224

60. Mc Cord KA, Ewald H, Agarwal A, et al. Treatment effects in randomised trials using routinely collected data for outcome assessment versus traditional trials: meta-research study. *BMJ*. Published online March 3, 2021:n450. doi:10.1136/bmj.n450

61. Lensen S, Macnair A, Love SB, et al. Access to routinely collected health data for clinical trials – review of successful data requests to UK registries. *Trials*. 2020;21(1):398. doi:10.1186/s13063-020-04329-8

62. McKay AJ, Jones AP, Gamble CL, Farmer AJ, Williamson PR. Use of routinely collected data in a UK cohort of publicly funded randomised clinical trials. *F1000Research*. 2020;9:323. doi:10.12688/f1000research.23316.3

63. Mc Cord KA, Hemkens LG. Using electronic health records for clinical trials: Where do we stand and where can we go? *Can Med Assoc J*. 2019;191(5):E128-E133. doi:10.1503/cmaj.180841

64. JACKIE MACARTHUR, LYNN MORRICE, CATHIE SUDLOW, MATTHEW SYDES. Data-enabled trials survey report: the British Heart Foundation Data Science Center. Published online July 2021. https://zenodo.org/record/5079178#.YRTo19NKhhE

65. Gossec L, Dougados M, Tubach F, Ravaud P. Reporting of Adherence to Medication in Recent Randomized Controlled Trials of 6 Chronic Diseases: A Systematic

Literature Review. *Am J Med Sci*. 2007;334(4):248-254. doi:10.1097/MAJ.0b013e318068dde8

66. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA*. 2007;297(11):1233-1240. doi:10.1001/jama.297.11.1233

67. Modi ND, Tan JQE, Rowland A, et al. The Influence of Pre-Existing Beta-Blockers Use on Survival Outcomes in HER2 Positive Advanced Breast Cancer: Pooled Analysis of Clinical Trial Data. *Front Oncol*. 2020;10:1130. doi:10.3389/fonc.2020.01130

68. Pinho-Gomes AC, Azevedo L, Ahn JM, et al. Compliance With Guideline-Directed Medical Therapy in Contemporary Coronary Revascularization Trials. *J Am Coll Cardiol*. 2018;71(6):591-602. doi:10.1016/j.jacc.2017.11.068

69. Bethel MA, Stevens SR, Buse JB, et al. Exploring the Possible Impact of Unbalanced Open-Label Drop-In of Glucose-Lowering Medications on EXSCEL Outcomes. *Circulation*. 2020;141(17):1360-1370. doi:10.1161/CIRCULATIONAHA.119.043353

70. Study of the effectiveness of additional reductions in cholesterol and homocysteine (SEARCH): Characteristics of a randomized trial among 12064 myocardial infarction survivors. *Am Heart J*. 2007;154(5):815-823.e6. doi:10.1016/j.ahj.2007.06.034

71. Borad MJ, Curtis KK, Babiker HM, et al. The Impact of Concomitant Medication Use on Patient Eligibility for Phase I Cancer Clinical Trials. *J Cancer*. 2012;3:345-353. doi:10.7150/jca.4714

72. RECOVERY Collaborative Group, Horby P, Lim WS, et al. Dexamethasone in Hospitalized Patients with Covid-19. *N Engl J Med*. 2021;384(8):693-704. doi:10.1056/NEJMoa2021436

73. the CONSORT Group, Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Trials*. 2010;11(1):32. doi:10.1186/1745-6215-11-32

74. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 2020;584(7821):430-436. doi:10.1038/s41586-020-2521-4

75. ICH Harmonised Tripartite Guideline. ICH Guideline E10: Choice of Control Group and Related Issues in Clinical Trials. Published online July 20, 2000. https://database.ich.org/sites/default/files/E10_Guideline.pdf

76. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*. 2014;348(mar07 3):g1687-g1687. doi:10.1136/bmj.g1687

77. Tesser J, Fleischmann R, Dore R, et al. Concomitant medication use in a large, international, multicenter, placebo controlled trial of anakinra, a recombinant interleukin 1 receptor antagonist, in patients with rheumatoid arthritis. *J Rheumatol*. 2004;31(4):649-654.

78. Schwarting A, Dooley MA, Roth DA, Edwards L, Thompson A, Wilson B. Impact of concomitant medication use on belimumab efficacy and safety in patients with systemic lupus erythematosus. *Lupus*. 2016;25(14):1587-1596. doi:10.1177/0961203316655215

79. Anker SD, Butler J, Filippatos G, et al. Baseline characteristics of patients with heart failure with preserved ejection fraction in the EMPEROR-Preserved trial. *Eur J Heart Fail*. 2020;22(12):2383-2392. doi:10.1002/ejhf.2064

80. RECOVERY Collaborative Group. Baricitinib in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial and updated meta-analysis. *Lancet Lond Engl*. 2022;400(10349):359-368. doi:10.1016/S0140-6736(22)01109-6

81. Sackett DL. Clinician-trialist rounds: 5. Cointervention bias – how to diagnose it in their trial and prevent it in yours. *Clin Trials*. 2011;8(4):440-442. doi:10.1177/1740774511410995

82. Beiderbeck AB, Sturkenboom MCJM, Coebergh JWW, Leufkens HGM, Stricker BHC. Misclassification of exposure is high when interview data on drug use are used as a proxy measure of chronic drug use during follow-up. *J Clin Epidemiol*. 2004;57(9):973-977. doi:10.1016/j.jclinepi.2003.12.017

83. Fogelman I, Lim L, Bassett R, et al. Prevalence and patterns of use of concomitant medications among participants in three multicenter human immunodeficiency virus type I clinical trials. AIDS Clinical Trials Group (ACTG). *J Acquir Immune Defic Syndr*. 1994;7(10):1057-1063.

84. Pagnoux C, Dechartres A, Giraudeau B, Seror R, Guillevin L, Ravaud P. Reporting of corticosteroid use in systemic disease trials: Evidence from a systematic review of the potential impact on treatment effect. *Arthritis Care Res*. 2010;62(7):1002-1008. doi:10.1002/acr.20139

85. Mokoka MC, McDonnell MJ, MacHale E, et al. Inadequate assessment of adherence to maintenance medication leads to loss of power and increased costs in trials of severe asthma therapy: results from a systematic literature review and modelling study. *Eur Respir J*. 2019;53(5):1802161. doi:10.1183/13993003.02161-2018

86. Shiovitz TM, Bain EE, McCann DJ, et al. Mitigating the Effects of Nonadherence in Clinical Trials. *J Clin Pharmacol*. 2016;56(9):1151-1164. doi:10.1002/jcph.689

87. McBrien CN, Menzies-Gow A. Less is more: the impact of maintenance treatment adherence in severe asthma clinical trials. *Eur Respir J*. 2019;53(5):1900599. doi:10.1183/13993003.00599-2019

88. van Onzenoort HAW, Menger FE, Neef C, et al. Participation in a clinical trial enhances adherence and persistence to treatment: a retrospective cohort study. *Hypertens Dallas Tex 1979*. 2011;58(4):573-578. doi:10.1161/HYPERTENSIONAHA.111.171074

89. Duff JM, Leather H, Walden EO, LaPlant KD, George TJ. Adequacy of Published Oncology Randomized Controlled Trials to Provide Therapeutic Details Needed for Clinical Application. *JNCI J Natl Cancer Inst*. 2010;102(10):702-705. doi:10.1093/jnci/djq117

90. Mahfoud F, Böhm M, Baumhäkel M. Inadequate reporting of concomitant drug treatment in cardiovascular interventional head-to-head trials. *Clin Cardiol*. 2012;35(4):255-256. doi:10.1002/clc.21958

91. Iudici M, Riveros C, Gabay C, Puéchal X, Berthelot JM, Ravaud P. Inadequate reporting of concomitant medications in rheumatology randomized controlled trials of pharmacologic interventions. *Semin Arthritis Rheum*. 2019;49(3):453-458. doi:10.1016/j.semarthrit.2019.03.009

92. Boussageon R, Supper I, Erpeldinger S, et al. Are concomitant treatments confounding factors in randomized controlled trials on intensive blood-glucose control in type 2 diabetes? a systematic review. *BMC Med Res Methodol*. 2013;13:107. doi:10.1186/1471-2288-13-107

93. Patel MX, Collins S, Hellier J, Bhatia G, Murray RM. The quality of reporting of phase II and III trials for new antipsychotics: a systematic review. *Psychol Med*. 2015;45(3):467-479. doi:10.1017/S0033291714001214

94. Rudd P. In Search of the Gold Standard for Compliance Measurement. *Arch Intern Med*. 1979;139(6):627. doi:10.1001/archinte.1979.03630430009004

95. Lehmann A, Aslani P, Ahmed R, et al. Assessing medication adherence: options to consider. *Int J Clin Pharm*. 2014;36(1):55-69. doi:10.1007/s11096-013-9865-x

96. Forbes CA, Deshpande S, Sorio-Vilela F, et al. A systematic literature review comparing methods for the measurement of patient persistence and adherence. *Curr Med Res Opin*. 2018;34(9):1613-1625. doi:10.1080/03007995.2018.1477747

97. Demonceau J, Ruppar T, Kristanto P, et al. Identification and assessment of adherence-enhancing interventions in studies assessing medication adherence through electronically compiled drug dosing histories: a systematic literature review and meta-analysis. *Drugs*. 2013;73(6):545-562. doi:10.1007/s40265-013-0041-3

98. Lam WY, Fresco P. Medication Adherence Measures: An Overview. *BioMed Res Int*. 2015;2015:217047. doi:10.1155/2015/217047

99. Farmer KC. Methods for measuring and monitoring medication regimen adherence in clinical trials and clinical practice. *Clin Ther*. 1999;21(6):1074-1090. doi:10.1016/S0149-2918(99)80026-5

100. Spangler L, Ichikawa LE, Hubbard RA, et al. A comparison of self-reported oral contraceptive use and automated pharmacy data in perimenopausal and early postmenopausal women. *Ann Epidemiol*. 2015;25(1):55-59. doi:10.1016/j.annepidem.2014.10.009

101. Garfield S, Clifford S, Eliasson L, Barber N, Willson A. Suitability of measures of self-reported medication adherence for routine clinical use: a systematic review. *BMC Med Res Methodol*. 2011;11:149. doi:10.1186/1471-2288-11-149

102. Kwan YH, Weng SD, Loh DHF, et al. Measurement Properties of Existing Patient-Reported Outcome Measures on Medication Adherence: Systematic Review. *J Med Internet Res*. 2020;22(10):e19179. doi:10.2196/19179

103. Fletcher SW, Pappius EM, Harper SJ. Measurement of medication compliance in a clinical setting. Comparison of three methods in patients prescribed digoxin. *Arch Intern Med*. 1979;139(6):635-638.

104. Vik SA, Maxwell CJ, Hogan DB. Measurement, correlates, and health outcomes of medication adherence among seniors. *Ann Pharmacother*. 2004;38(2):303-312. doi:10.1345/aph.1D252

105. Garber MC, Nau DP, Erickson SR, Aikens JE, Lawrence JB. The Concordance of Self-Report With Other Measures of Medication Adherence: A Summary of the Literature. *Med Care*. 2004;42(7):649-652. doi:10.1097/01.mlr.0000129496.05898.02

106. de Denus S, O'Meara E, Desai AS, et al. Spironolactone Metabolites in TOPCAT — New Insights into Regional Variation. *N Engl J Med*. 2017;376(17):1690-1692. doi:10.1056/NEJMc1612601

107. Allen EN, Mushi AK, Massawe IS, et al. How experiences become data: the process of eliciting adverse event, medical history and concomitant medication reports in antimalarial and antiretroviral interaction trials. *BMC Med Res Methodol*. 2013;13:140. doi:10.1186/1471-2288-13-140

108. Gama H, Correia S, Lunet N. Questionnaire design and the recall of pharmacological treatments: a systematic review. *Pharmacoepidemiol Drug Saf*. 2009;18(3):175-187. doi:10.1002/pds.1703

109. Meyer C, Stern M, Woolley W, Jeanmonod R, Jeanmonod D. How reliable are patient-completed medication reconciliation forms compared with pharmacy lists? *Am J Emerg Med*. 2012;30(7):1048-1054. doi:10.1016/j.ajem.2011.06.038

110. Kwon A, Bungay KM, Pei Y, et al. Antidepressant use: concordance between self-report and claims records. *Med Care*. 2003;41(3):368-374. doi:10.1097/01.MLR.0000053019.79054.B6

111. Konstantinou P, Kasinopoulos O, Karashiali C, et al. A Scoping Review of Methods Used to Assess Medication Adherence in Patients with Chronic Conditions. *Ann Behav Med*. 2022;56(12):1201-1217. doi:10.1093/abm/kaab080

112. Rand CS, Wise RA, Nides M, et al. Metered-dose inhaler adherence in a clinical trial. *Am Rev Respir Dis*. 1992;146(6):1559-1564. doi:10.1164/ajrccm/146.6.1559

113. Wetzels GE, Nelemans PJ, Schouten JS, van Wijk BL, Prins MH. All that glisters is not gold: a comparison of electronic monitoring versus filled prescriptions – an observational study. *BMC Health Serv Res*. 2006;6(1):8. doi:10.1186/1472-6963-6-8

114. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323-337. doi:10.1016/j.jclinepi.2004.10.012

115. Fanaroff AC, Peterson ED, Kaltenbach LA, et al. Agreement and Accuracy of Medication Persistence Identified by Patient Self-report vs Pharmacy Fill: A Secondary Analysis of the Cluster Randomized ARTEMIS Trial. *JAMA Cardiol*. 2020;5(5):532. doi:10.1001/jamacardio.2020.0125

116. Toscos T, Drouin M, Pater JA, et al. Medication adherence for atrial fibrillation patients: triangulating measures from a smart pill bottle, e-prescribing software, and patient communication through the electronic health record. *JAMIA Open*. 2020;3(2):233-242. doi:10.1093/jamiaopen/ooaa007

117. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46(5):830-836. doi:10.1016/j.jbi.2013.06.010

118. Nowalk MP, D'Agostino HEA, Zimmerman RK, et al. Agreement among sources of adult influenza vaccination in the age of immunization information systems. *Vaccine*. 2021;39(47):6829-6836. doi:10.1016/j.vaccine.2021.10.041

119. Plaza V, Giner J, Curto E, et al. Assessing Adherence by Combining the Test of Adherence to Inhalers With Pharmacy Refill Records. *J Investig Allergol Clin Immunol*. 2021;31(1):58-64. doi:10.18176/jiaci.0461

120. Van der Heyden J, Berete F, Renard F, et al. Assessing polypharmacy in the older population: Comparison of a self-reported and prescription based method. *Pharmacoepidemiol Drug Saf*. 2021;30(12):1716-1726. doi:10.1002/pds.5321

121. Lam JH, Singh S, Kuo GM. Comparisons of immunization records between a community pharmacy, a regional registry, and a health system. *J Am Pharm Assoc JAPhA*. 2019;59(1):30-34. doi:10.1016/j.japh.2018.09.007

122. National Institute for Health and Care Excellence. Medicines optimisation: the safe and effective use of medicines to enable the best possible outcomes. Published online 2015. Medicines optimisation: the safe and effective use of medicines to enable the best possible outcomes

123. Hall GC, Sauer B, Bourke A, Brown JS, Reynolds MW, Casale RL. Guidelines for good database selection and use in pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf*. 2012;21(1):1-10. doi:10.1002/pds.2229

124. Savitz ST, Stearns SC, Zhou L, et al. A Comparison of Self-reported Medication Adherence to Concordance Between Part D Claims and Medication Possession. *Med Care*. 2017;55(5):500-505. doi:10.1097/MLR.0000000000000701

125. Kelloway JS, Wyatt RA, Adlis SA. Comparison of patients' compliance with prescribed oral and inhaled asthma medications. *Arch Intern Med*. 1994;154(12):1349-1352.

126. Schmidt C. Cancer: Reshaping the cancer clinic. *Nature*. 2015;527(7576):S10-S11. doi:10.1038/527S10a

127. Wallerstedt SM, Bladh L, Ramsberg J. A cost-effectiveness analysis of an in-hospital clinical pharmacist service. *BMJ Open*. 2012;2(1):e000329. doi:10.1136/bmjopen-2011-000329

128. Fiore LD, Brophy M, Ferguson RE, et al. A point-of-care clinical trial comparing insulin administered using a sliding scale *versus* a weight-based regimen. *Clin Trials*. 2011;8(2):183-195. doi:10.1177/1740774511398368

129. Bereznicki BJ, Peterson GM, Jackson SL, Walters EH, Fitzmaurice KD, Gee PR. Data-mining of medication records to improve asthma management. *Med J Aust*. 2008;189(1):21-25. doi:10.5694/j.1326-5377.2008.tb01889.x

130. Hess CN, Rao SV, Kong DF, et al. Embedding a randomized clinical trial into an ongoing registry infrastructure: Unique opportunities for efficiency in design of the Study of Access site For Enhancement of Percutaneous Coronary Intervention for Women (SAFE-PCI for Women). *Am Heart J*. 2013;166(3):421-428.e1. doi:10.1016/j.ahj.2013.06.013

131. MacDonald TM, Mackenzie IS, Wei L, Hawkey CJ, Ford I. Methodology of a large prospective, randomised, open, blinded endpoint streamlined safety study of celecoxib versus traditional non-steroidal anti-inflammatory drugs in patients with osteoarthritis or rheumatoid arthritis: protocol of the standard care versus celecoxib outcome trial (SCOT). *BMJ Open*. 2013;3(1):e002295. doi:10.1136/bmjopen-2012-002295

132. Fitzpatrick T, Perrier L, Shakik S, et al. Assessment of Long-term Follow-up of Randomized Trial Participants by Linkage to Routinely Collected Data: A Scoping Review and Analysis. *JAMA Netw Open*. 2018;1(8):e186019. doi:10.1001/jamanetworkopen.2018.6019

133. Grossberg R, Zhang Y, Gross R. A time-to-prescription-refill measure of antiretroviral adherence predicted changes in viral load in HIV. *J Clin Epidemiol*. 2004;57(10):1107-1110. doi:10.1016/j.jclinepi.2004.04.002

134. Tommelein E, Mehuys E, Van Tongelen I, Brusselle G, Boussery K. Accuracy of the Medication Adherence Report Scale (MARS-5) as a quantitative measure of adherence to inhalation medication in patients with COPD. *Ann Pharmacother*. 2014;48(5):589-595. doi:10.1177/1060028014522982

135. Nau DP, Steinke DT, Williams LK, et al. Adherence analysis using visual analog scale versus claims-based estimation. *Ann Pharmacother*. 2007;41(11):1792-1797. doi:10.1345/aph.1K264

136. Moura CS, Payette Y, Boileau C, Abrahamowicz M, Pilote L, Bernatsky S. Agreement in the CARTaGENE cohort between self-reported medication use and claim data. *Chronic Illn*. Published online January 10, 2021:1742395320985913. doi:10.1177/1742395320985913

137. Cate H, Bhattacharya D, Clark A, Holland R, Broadway DC. A comparison of measures used to describe adherence to glaucoma medication in a randomised

controlled trial. *Clin Trials Lond Engl*. 2015;12(6):608-617. doi:10.1177/1740774515592636

138. Lauffenburger JC, Fontanet CP, Isaac T, et al. Comparison of a new 3-item self-reported measure of adherence to medication with pharmacy claims data in patients with cardiometabolic disease. *Am Heart J*. 2020;228:36-43. doi:10.1016/j.ahj.2020.06.012

139. Lemanska A, Byford RC, Cruickshank C, et al. Linkage of the CHHiP randomised controlled trial with primary care data: a study investigating ways of supplementing cancer trials and improving evidence-based practice. *BMC Med Res Methodol*. 2020;20(1):198. doi:10.1186/s12874-020-01078-9

140. Powell GA, Bonnett LJ, Tudur-Smith C, Hughes DA, Williamson PR, Marson AG. Using routinely recorded data in the UK to assess outcomes in a randomised controlled trial: The Trials of Access. *Trials*. 2017;18(1):389. doi:10.1186/s13063-017-2135-9

141. NHS Digital. Data services. Published October 9, 2019. Accessed October 26, 2019. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services

142. Bowman L, Mafham M. A Randomized Trial Assessing the Effects of Inclisiran on Clinical Outcomes Among People With Cardiovascular Disease (ORION-4). ClinicalTrials.gov. Accessed October 25, 2019. https://clinicaltrials.gov/ct2/show/NCT03705234

143. HM Governement. UK Life Sciences Sector Deal 2, 2018. Published online 2017:35.

144. Health and Social Care Information Centre (Great Britain). *Health and Social Care Information Centre (HSCIC) Annual Report and Accounts 2018-19.*; 2019.

145. Bowman L, Mafham M, Stevens W, et al. ASCEND: A Study of Cardiovascular Events iN Diabetes: Characteristics of a randomized trial of aspirin and of omega-3 fatty acid supplementation in 15,480 people with diabetes. *Am Heart J*. 2018;198:135-144. doi:10.1016/j.ahj.2017.12.006

146. The ASCEND Study Collaborative Group. Effects of Aspirin for Primary Prevention in Persons with Diabetes Mellitus. *N Engl J Med*. 2018;379(16):1529-1539. doi:10.1056/NEJMoa1804988

147. The ASCEND Study Collaborative Group. Effects of n–3 Fatty Acid Supplements in Diabetes Mellitus. *N Engl J Med*. 2018;379(16):1540-1550. doi:10.1056/NEJMoa1804989

148. NHS Digital. Hospital Episode Statistics (HES). Published March 26, 2019. Accessed October 26, 2019. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics

149. NHS Digital. Secondary Uses Service (SUS). Published March 6, 2019. Accessed October 26, 2019. https://digital.nhs.uk/services/secondary-uses-service-sus

150. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol*. 2017;46(4):1093-1093i. doi:10.1093/ije/dyx015

151. Julious SA, Horspool MJ, Davis S, et al. Open-label, cluster randomised controlled trial and economic evaluation of a brief letter from a GP on unscheduled medical contacts associated with the start of the school year: the PLEASANT trial. *BMJ Open*. 2018;8(4):e017367. doi:10.1136/bmjopen-2017-017367

152. electronic Cluster Randomised Trial Research Team eCRT Research Team, Gulliford MC, van Staa T, et al. Cluster randomised trial in the General Practice Research Database: 1. Electronic decision support to reduce antibiotic prescribing in primary care (eCRT study). *Trials*. 2011;12(1):115. doi:10.1186/1745-6215-12-115

153. Elkhenini HF, Davis KJ, Stein ND, et al. Using an electronic medical record (EMR) to conduct clinical trials: Salford Lung Study feasibility. *BMC Med Inform Decis Mak*. 2015;15(1):8. doi:10.1186/s12911-015-0132-z

154. van Staa TP, Dyson L, McCann G, et al. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health Technol Assess*. 2014;18(43). doi:10.3310/hta18430

155. Rockenschaub P, Ansell D, Shallcross L. Linking individual-level data on diagnoses and dispensing for research on antibiotic use: Evaluation of a novel data source from English secondary care. *Pharmacoepidemiol Drug Saf*. 2018;27(2):206-212. doi:10.1002/pds.4367

156. NHSBSA. Prescribing data. Accessed October 26, 2019. https://www.nhsbsa.nhs.uk/prescription-data/prescribing-data

157. NHSBSA. Dispensing data. Accessed October 26, 2019. https://www.nhsbsa.nhs.uk/prescription-data/dispensing-data

158. NHS Digital. Electronic Prescribing and Administration (EPMA) Data in Secondary Care under the COVID-19 Direction. Published May 11, 2022. https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/electronic-prescribing-and-administration-epma-data-in-secondary-care

159. Emanuel G, Henson KE, Broggio J, et al. Endocrine therapy in the years following a diagnosis of breast cancer: A proof of concept study using the primary care prescription database linked to cancer registration data. *Cancer Epidemiol*. 2019;61:185-189. doi:10.1016/j.canep.2019.04.012

160. Datta-Nemdharry P, Thomson A, Beynon J. Opportunities and Challenges in Developing a Cohort of Patients with Type 2 Diabetes Mellitus Using Electronic Primary Care Data. Song Y, ed. *PLOS ONE*. 2016;11(11):e0162236. doi:10.1371/journal.pone.0162236

161. Michel A, González-Pérez A, Sáez ME, García Rodríguez LA. Risk of bleeding events among patients with systemic sclerosis and the general population in the UK: a large population-based cohort study. *Clin Rheumatol*. Published online May 14, 2019. doi:10.1007/s10067-019-04588-0

162. Alvarez-Madrazo S, McTaggart S, Nangle C, Nicholson E, Bennie M. Data Resource Profile: The Scottish National Prescribing Information System (PIS). *Int J Epidemiol*. 2016;45(3):714-715f. doi:10.1093/ije/dyw060

163. Bennie M, Malcolm W, McTaggart S, Mueller T. Improving prescribing through big data approaches-Ten years of the Scottish Prescribing Information System. *Br J Clin Pharmacol*. 2020;86(2):250-257. doi:10.1111/bcp.14184

164. Hafferty JD, Campbell AI, Navrady LB, et al. Self-reported medication use validated through record linkage to national prescribing data. *J Clin Epidemiol*. 2018;94:132-142. doi:10.1016/j.jclinepi.2017.10.013

165. Ford I, Murray H, Packard CJ, Shepherd J, Macfarlane PW, Cobbe SM. Long-Term Follow-up of the West of Scotland Coronary Prevention Study. *N Engl J Med*. 2007;357(15):1477-1486. doi:10.1056/NEJMoa065994

166. Guthrie B, Kavanagh K, Robertson C, et al. Data feedback and behavioural change intervention to improve primary care prescribing safety (EFIPPS): multicentre, three arm, cluster randomised controlled trial. *BMJ*. Published online August 18, 2016:i4079. doi:10.1136/bmj.i4079

167. Marwick CA, Hossain A, Nogueira R, et al. Feedback of Antibiotic Prescribing in Primary Care (FAPPC) trial: results of a real-world cluster randomized controlled trial in Scotland, UK. *J Antimicrob Chemother*. Published online September 29, 2022:dkac317. doi:10.1093/jac/dkac317

168. Erlinge D, Omerovic E, Fröbert O, et al. Bivalirudin versus Heparin Monotherapy in Myocardial Infarction. *N Engl J Med*. 2017;377(12):1132-1142. doi:10.1056/NEJMoa1706443

169. Bender BG, Cvietusa PJ, Goodrich GK, et al. Pragmatic Trial of Health Care Technologies to Improve Adherence to Pediatric Asthma Treatment: A Randomized Clinical Trial. *JAMA Pediatr*. 2015;169(4):317. doi:10.1001/jamapediatrics.2014.3280

170. Meinecke AK, Welsing P, Kafatos G, et al. Series: Pragmatic trials and real world evidence: Paper 8. Data collection and management. *J Clin Epidemiol*. 2017;91:13-22. doi:10.1016/j.jclinepi.2017.07.003

171. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Ann Intern Med*. 2009;151(5):359-360. doi:10.7326/0003-4819-151-5-200909010-00141

172. Morris AD, Boyle DI, MacAlpine R, et al. The diabetes audit and research in Tayside Scotland (darts) study: electronic record linkage to create a diabetes register. *BMJ*. 1997;315(7107):524-528. doi:10.1136/bmj.315.7107.524

173. Mudd KE, Bollinger ME, Hsu VD, Manning A, Tsoukleris MG, Butz AM. Concordance of Medicaid and pharmacy record data in inner-city children with asthma. *Contemp Clin Trials*. 2008;29(1):13-20. doi:10.1016/j.cct.2007.05.002

174. Haerskjold A, Kieler H, Bergman G, Byström C, Linder M. Use of palivizumab is underestimated in the Swedish Prescribed Drug Register – implications for register-based drug studies. *Clin Epidemiol*. Published online December 2014:45. doi:10.2147/CLEP.S73337

175. Bröms G, Söderling J, Sachs MC, et al. Capturing biologic treatment for IBD in the Swedish Prescribed Drug Register and the Swedish National Patient Register – a validation study. *Scand J Gastroenterol*. 2021;56(4):410-421. doi:10.1080/00365521.2021.1884894

176. Garcia CJ, Haynes K, Pokorney SD, et al. Practical challenges in the conduct of pragmatic trials embedded in health plans: Lessons of IMPACT-AFib, an FDA-Catalyst trial. *Clin Trials*. Published online June 26, 2020:174077452092842. doi:10.1177/1740774520928426

177. Overhage JM, Overhage LM. Sensible use of observational clinical data. *Stat Methods Med Res*. 2013;22(1):7-13. doi:10.1177/0962280211403598

178. Kaviani P, Landi SN, McKethan A, Brookhart MA, McGrath LJ. Who are we missing? Underrepresentation of data sources used for pharmacoepidemiology research in the United States. *Pharmacoepidemiol Drug Saf*. 2020;29(11):1494-1498. doi:10.1002/pds.5087

179. Pacurariu A, Plueschke K, McGettigan P, et al. Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. *BMJ Open*. 2018;8(9):e023090. doi:10.1136/bmjopen-2018-023090

180. Ferrer P, Ballarín E, Sabaté M, et al. Sources of European drug consumption data at a country level. *Int J Public Health*. 2014;59(5):877-887. doi:10.1007/s00038-014-0564-8

181. Lai ECC, Man KKC, Chaiyakunapruk N, et al. Brief Report: Databases in the Asia-Pacific Region: The Potential for a Distributed Network Approach. *Epidemiol Camb Mass*. 2015;26(6):815-820. doi:10.1097/EDE.0000000000000325

182. Anderson TS, Xu E, Whitaker E, Steinman MA. A systematic review of methods for determining cross-sectional active medications using pharmacy databases. *Pharmacoepidemiol Drug Saf*. 2019;28(4):403-421. doi:10.1002/pds.4706

183. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. *Med Care*. 2013;51(8 Suppl 3):S80-86. doi:10.1097/MLR.0b013e31829b1d48

184. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Fam Pract*. 2006;23(2):253-263. doi:10.1093/fampra/cmi106

185. Goldacre B, MacKenna B. The NHS deserves better use of hospital medicines data. *BMJ*. Published online July 17, 2020:m2607. doi:10.1136/bmj.m2607

186. Bacon S, Goldacre B. Barriers to Working With National Health Service England's Open Data. *J Med Internet Res*. 2020;22(1):e15603. doi:10.2196/15603

187. Murray ML, Love SB, Carpenter JR, et al. Data provenance and integrity of health-care systems data for clinical trials. *Lancet Digit Health*. 2022;4(8):e567-e568. doi:10.1016/S2589-7500(22)00122-4

188. Ladanie A, Speich B, Briel M, et al. Single pivotal trials with few corroborating characteristics were used for FDA approval of cancer therapies. *J Clin Epidemiol*. 2019;114:49-59. doi:10.1016/j.jclinepi.2019.05.033

189. Heard K, Anderson VE, Dart RC, Green JL. Accuracy of the Structured Medication History Assessment Tool (MedHAT) Compared with Recorded Real-Time Medication Use. *Pharmacotherapy*. 2016;36(5):496-504. doi:10.1002/phar.1750

190. Hess LM, Saboda K, Malone DC, Salasche S, Warneke J, Alberts DS. Adherence assessment using medication weight in a phase IIb clinical trial of difluoromethylornithine for the chemoprevention of skin cancer. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2005;14(11 Pt 1):2579-2583. doi:10.1158/1055-9965.EPI-05-0104

191. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care*. 2012;50 Suppl:S21-29. doi:10.1097/MLR.0b013e318257dd67

192. Weiskopf NG, 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):144-151. doi:10.1136/amiajnl-2011-000681

193. Juran JM, Gryna FM. *Juran's Quality Control Handbook*. 4th edn. McGraw-Hill; 1988.

194. Gentles SJ, Lokker C, McKibbon KA. Health Information Technology to Facilitate Communication Involving Health Care Providers, Caregivers, and Pediatric Patients: A Scoping Review. *J Med Internet Res*. 2010;12(2):e22. doi:10.2196/jmir.1390

195. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, McEwen SA. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Res Synth Methods*. 2014;5(4):371-385. doi:10.1002/jrsm.1123

196. Larsen MD, Cars T, Hallas J. A MiniReview of the use of hospital-based databases in observational inpatient studies of drugs. *Basic Clin Pharmacol Toxicol*. 2013;112(1):13-18. doi:10.1111/j.1742-7843.2012.00928.x

197. Brüne M, Emmel C, Meilands G, et al. Self-reported medication intake vs information from other data sources such as pharmacy records or medical records: Identification and description of existing publications, and comparison of agreement results for publications focusing on patients with cancer - a systematic review. *Pharmacoepidemiol Drug Saf*. 2021;30(5):531-560. doi:10.1002/pds.5210

198. Bian J, Lyu T, Loiacono A, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J Am Med Inform Assoc JAMIA*. 2020;27(12):1999-2010. doi:10.1093/jamia/ocaa245

199. Monnette A, Zhang Y, Shao H, Shi L. Concordance of Adherence Measurement Using Self-Reported Adherence Questionnaires and Medication Monitoring Devices: An Updated Review. *PharmacoEconomics*. 2018;36(1):17-27. doi:10.1007/s40273-017-0570-9

200. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Current use of routinely collected health data to complement randomized controlled trials: a meta-epidemiological survey. *CMAJ Open*. 2016;4(2):E132-E140. doi:10.9778/cmajo.20150036

201. Cameron JI, Tsoi C, Marsella A. Optimizing Stroke Systems of Care by Enhancing Transitions Across Care Environments. *Stroke*. 2008;39(9):2637-2643. doi:10.1161/STROKEAHA.107.501064

202. Levac D, Wishart L, Missiuna C, Wright V. The Application of Motor Learning Strategies Within Functionally Based Interventions for Children with Neuromotor Conditions. *Pediatr Phys Ther*. 2009;21(4):345-355. doi:10.1097/PEP.0b013e3181beb09d

203. Wagner G, Glechner A, Persad E, et al. Risk of Contrast-Associated Acute Kidney Injury in Patients Undergoing Peripheral Angiography with Carbon Dioxide Compared to Iodine-Containing Contrast Agents: A Systematic Review and Meta-Analysis. *J Clin Med*. 2022;11(23):7203. doi:10.3390/jcm11237203

204. Franco JS, Vizcaya D. Availability of secondary healthcare data for conducting pharmacoepidemiology studies in Colombia: A systematic review. *Pharmacol Res Perspect*. 2020;8(5):e00661. doi:10.1002/prp2.661

205. Bartlett LE, Pratt N, Roughead EE. Does tablet formulation alone improve adherence and persistence: a comparison of ezetimibe fixed dose combination versus ezetimibe separate pill combination? *Br J Clin Pharmacol*. 2017;83(1):202-210. doi:10.1111/bcp.13088

206. Bülow C, Noergaard JDSV, Faerch KU, et al. Causes of discrepancies between medications listed in the national electronic prescribing system and patients' actual use of medications. *Basic Clin Pharmacol Toxicol*. 2021;129(3):221-231. doi:10.1111/bcpt.13626

207. Voss SS, Glode Helmuth I, Hiul Suppli C, Valentiner-Branth P. Underreporting of the 5-year tetanus, diphtheria, pertussis and polio booster vaccination in the Danish Vaccination Register. *BMC Public Health*. 2020;20(1):1681. doi:10.1186/s12889-020-09816-w

208. Hashibe M, Ou JY, Herget K, et al. Feasibility of Capturing Cancer Treatment Data in the Utah All-Payer Claims Database. *JCO Clin Cancer Inform*. 2019;3:1-10. doi:10.1200/CCI.19.00027

209. Weerasinghe A, Smith CR, Majpruz V, et al. Validity of Administrative Databases in Comparison to Medical Charts for Breast Cancer Treatment Data. *J Cancer Epidemiol*. 2018;2018:9218595. doi:10.1155/2018/9218595

210. Osterberg L, Blaschke T. Adherence to Medication. *N Engl J Med*. 2005;353(5):487-497. doi:10.1056/NEJMra050100

211. Whalley Buono E, Vrijens B, Bosworth HB, Liu LZ, Zullig LL, Granger BB. Coming full circle in the measurement of medication adherence: opportunities and implications for health care. *Patient Prefer Adherence*. 2017;11:1009-1017. doi:10.2147/PPA.S127131

212. Evandt J, Skurtveit S, Oftedal B, et al. Agreement between self-reported and registry-based use of sleep medications and tranquilizers. *Pharmacoepidemiol Drug Saf*. 2019;28(10):1336-1343. doi:10.1002/pds.4854

213. Cohen JM, Wood ME, Hernandez-Diaz S, Nordeng H. Agreement between paternal self-reported medication use and records from a national prescription database. *Pharmacoepidemiol Drug Saf*. 2018;27(4):413-421. doi:10.1002/pds.4411

214. Norell SE, Boethius G, Persson I. Oral contraceptive use: interview data versus pharmacy records. *Int J Epidemiol*. 1998;27(6):1033-1037. doi:10.1093/ije/27.6.1033

215. Sandini L, Pentti K, Tuppurainen M, Kröger H, Honkanen R. Agreement of self-reported estrogen use with prescription data: an analysis of women from the Kuopio Osteoporosis Risk Factor and Prevention Study. *Menopause N Y N*. 2008;15(2):282-289. doi:10.1097/gme.0b013e3181334b6c

216. King JP, McLean HQ, Belongia EA. Validation of self-reported influenza vaccination in the current and prior season. *Influenza Other Respir Viruses*. 2018;12(6):808-813. doi:10.1111/irv.12593

217. Glassberg MB, Trygstad T, Wei D, Robinson T, Farley JF. Accuracy of Prescription Claims Data in Identifying Truly Nonadherent Patients. *J Manag Care Spec Pharm*. 2019;25(12):1349-1356. doi:10.18553/jmcp.2019.25.12.1349

218. Brunelli SM. Use of prescription drug claims data to identify lipid-lowering medication exposure in pharmacoepidemiology studies: potential pitfalls. *Pharmacoepidemiol Drug Saf*. 2016;25(7):844-846. doi:10.1002/pds.4017

219. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev MCRR*. 2010;67(5):503-527. doi:10.1177/1077558709359007

220. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Med Care*. 2013;51(Supplement 8Suppl 3):S30-S37. doi:10.1097/MLR.0b013e31829b1dbd

221. Schwartz KL, Wilton AS, Langford BJ, et al. Comparing prescribing and dispensing databases to study antibiotic use: a validation study of the Electronic Medical Record Administrative data Linked Database (EMRALD). *J Antimicrob Chemother*. 2019;74(7):2091-2097. doi:10.1093/jac/dkz033

222. Waaseth M, Havelkova M, Forsdahl G, Lund E, Log T. Use of Selective Serotonin Reuptake Inhibitors - Validity of Self-Report versus Plasma Concentrations and Pharmacy Dispensations - A Cross-Sectional Analysis of the Norwegian Women and Cancer Study. *Clin Epidemiol*. 2022;14:815-826. doi:10.2147/CLEP.S366760

223. Cheung K, El Marroun H, Elfrink ME, Jaddoe VWV, Visser LE, Stricker BHC. The concordance between self-reported medication use and pharmacy records in pregnant women. *Pharmacoepidemiol Drug Saf*. 2017;26(9):1119-1125. doi:10.1002/pds.4264

224. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. *J Clin Epidemiol*. 2012;65(3):343-349.e2. doi:10.1016/j.jclinepi.2011.09.002

225. Cramer JA, Silverman SL, Gold DT. Methodological considerations in using claims databases to evaluate persistence with bisphosphonates for osteoporosis. *Curr Med Res Opin*. 2007;23(10):2369-2377. doi:10.1185/030079907X226311

226. Grymonpre R, Cheang M, Fraser M, Metge C, Sitar DS. Validity of a prescription claims database to estimate medication adherence in older persons. *Med Care*. 2006;44(5). doi:10.1097/01.mlr.0000207817.32496.cb

227. Richardson K, Kenny RA, Peklar J, Bennett K. Agreement between patient interview data on prescription medication use and pharmacy records in those aged older than 50 years varied by therapeutic group and reporting of indicated health conditions. *J Clin Epidemiol*. 2013;66(11):1308-1316. doi:10.1016/j.jclinepi.2013.02.016

228. Brien SE, Lorenzetti DL, Lewis S, Kennedy J, Ghali WA. Overview of a formal scoping review on health system report cards. *Implement Sci*. 2010;5(1):2. doi:10.1186/1748-5908-5-2

229. International Conference on Harmonization. ICH E6 - Guideline for good clinical practice (R2). Published online December 15, 2016. Accessed October 16, 2020. https://www.ema.europa.eu/documents/scientific-guideline/ich-e-6-r2-guideline-good-clinical-practice-step-5_en.pdf

230. NHS Digital. Medicines dispensed in Primary Care NHS Business Services Authority data. Published May 26, 2021. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/medicines-dispensed-in-primary-care-nhsbsa-data

231. NHS Digital. GPES Data for Pandemic Planning and Research - Guide for analysts and users of the data. Published September 20, 2021. Accessed September 21, 2021. https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data#code-clusters-and-content

232. EBM DataLab, University of Oxford. OpenPrescribing.net. Published 2019. Accessed October 28, 2019. https://openprescribing.net/

233. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. doi:10.1093/ije/dyv098

234. Garcia CJ, Haynes K, Pokorney SD, et al. Practical challenges in the conduct of pragmatic trials embedded in health plans: Lessons of IMPACT-AFib, an FDA-Catalyst trial. *Clin Trials Lond Engl*. 2020;17(4):360-367. doi:10.1177/1740774520928426

235. Lauer MS, D'Agostino RB. The Randomized Registry Trial — The Next Disruptive Technology in Clinical Research? *N Engl J Med*. 2013;369(17):1579-1581. doi:10.1056/NEJMp1310102

236. James S, Rao SV, Granger CB. Registry-based randomized clinical trials—a new clinical trial paradigm. *Nat Rev Cardiol*. 2015;12(5):312-316. doi:10.1038/nrcardio.2015.33

237. Lachin JM. The role of measurement reliability in clinical trials. *Clin Trials*. 2004;1(6):553-566. doi:10.1191/1740774504cn057oa

238. NHS Digital. Collection of GP Data for Planning and Research Deferred. Published January 9, 2021. Accessed February 16, 2024. https://digital.nhs.uk/news/2021/collection-of-gp-data-for-planning-and-research-to-go-ahead-on-1-september-2021

239. NHSBSA. Quality Assurance of Administrative Data (QAAD) Assessment - Prescription Data. Published online July 2020. https://www.nhsbsa.nhs.uk/sites/default/files/2020-07/QAAD_Assessment_Prescription_Data_v001_0.pdf

240. OpenSAFELY. OpenCodelists. https://www.opencodelists.org/docs/

241. RECOVERY Collaborative Group. RECOVERY trial - Definition and Derivation of Baseline Characteristics and Outcomes. Published online August 23, 2022. https://www.recoverytrial.net/files/recovery-outcomes-definitions-v5-0.pdf

242. Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. *Commun ACM*. 1996;39(11):86-95. doi:10.1145/240455.240479

243. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. *J Manag Inf Syst*. 1996;12(4):5-33. doi:10.1080/07421222.1996.11518099

244. Arts DGT. Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. *J Am Med Inform Assoc*. 2002;9(6):600-611. doi:10.1197/jamia.M1087

245. Love SB, Kilanowski A, Yorke-Edwards V, et al. Use of routinely collected health data in randomised clinical trials: comparison of trial-specific death data in the BOSS trial with NHS Digital data. *Trials*. 2021;22(1):654. doi:10.1186/s13063-021-05613-x

246. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. *Eur J Cancer*. 2009;45(5):756-764. doi:10.1016/j.ejca.2008.11.033

247. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. *Eur J Cancer*. 2009;45(5):747-755. doi:10.1016/j.ejca.2008.11.032

248. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas*. 1960;20(1):37-46. doi:10.1177/001316446002000104

249. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428. doi:10.1037//0033-2909.86.2.420

250. Wang PS, Benner JS, Glynn RJ, Winkelmayer WC, Mogun H, Avorn J. How well do patients report noncompliance with antihypertensive medications?: a comparison of self-report versus filled prescriptions. *Pharmacoepidemiol Drug Saf*. 2004;13(1):11-19. doi:10.1002/pds.819

251. Hoopes M, Angier H, Raynor LA, et al. Development of an algorithm to link electronic health record prescriptions with pharmacy dispense claims. *J Am Med Inform Assoc JAMIA*. 2018;25(10):1322-1330. doi:10.1093/jamia/ocy095

252. Brown C, Barron TI, Bennett K, MacDonald D, Dwane F, Sharp L. Generalisability of pharmacoepidemiological studies using restricted prescription data. *Ir J Med Sci*. 2016;185(3):723-727. doi:10.1007/s11845-015-1360-z

253. Aarnio E, Huupponen R, Martikainen JE, Korhonen MJ. First insight to the Finnish nationwide electronic prescription database as a data source for pharmacoepidemiology research. *Res Soc Adm Pharm RSAP*. 2020;16(4):553-559. doi:10.1016/j.sapharm.2019.06.012

254. Johannesdottir SA, Horváth-Puhó E, Ehrenstein V, Schmidt M, Pedersen L, Sørensen HT. Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions. *Clin Epidemiol*. 2012;4:303-313. doi:10.2147/CLEP.S37587

255. Overbeek JA, Swart KMA, Houben E, Penning-van Beest FJA, Herings RMC. Completeness and Representativeness of the PHARMO General Practitioner (GP) Data: A Comparison with National Statistics. *Clin Epidemiol*. 2023;15:1-11. doi:10.2147/CLEP.S389598

256. Wilke T, Groth A, Mueller S, et al. How to use pharmacy claims data to measure patient nonadherence? The example of oral diabetics in therapy of type 2 diabetes mellitus. *Eur J Health Econ HEPAC Health Econ Prev Care*. 2013;14(3):551-568. doi:10.1007/s10198-012-0410-y

257. Martin BC, Wiley-Exley EK, Richards S, Domino ME, Carey TS, Sleath BL. Contrasting measures of adherence with simple drug use, medication switching, and therapeutic duplication. *Ann Pharmacother*. 2009;43(1):36-44. doi:10.1345/aph.1K671

258. Anderson TS, Jing B, Wray CM, et al. Comparison of Pharmacy Database Methods for Determining Prevalent Chronic Medication Use. *Med Care*. 2019;57(10):836-842. doi:10.1097/MLR.0000000000001188

259. Weisman A, King LK, Mamdani M. Reporting and variability of constructing medication treatment episodes in pharmacoepidemiology studies: A methodologic systematic review using the case study of DPP-4 inhibitors and cardiovascular outcomes. *Pharmacoepidemiol Drug Saf*. 2020;29(8):939-950. doi:10.1002/pds.5071

260. Price M, Bowen M, Lau F, Kitson N, Bardal S. Assessing accuracy of an electronic provincial medication repository. *BMC Med Inform Decis Mak*. 2012;12(1):42. doi:10.1186/1472-6947-12-42

261. Grimes T, Fitzsimons M, Galvin M, Delaney T. Relative accuracy and availability of an Irish National Database of dispensed medication as a source of medication history information: observational study and retrospective record analysis. *J Clin Pharm Ther*. 2013;38(3):219-224. doi:10.1111/jcpt.12036

262. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977;33(1):159. doi:10.2307/2529310

263. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46(5):423-429. doi:10.1016/0895-4356(93)90018-V

264. Nahm ML, Pieper CF, Cunningham MM. Quantifying data quality for clinical trials using electronic data capture. *PloS One*. 2008;3(8):e3049. doi:10.1371/journal.pone.0003049

265. Andrade SE, Bérard A, Nordeng HME, Wood ME, van Gelder MMHJ, Toh S. Administrative Claims Data Versus Augmented Pregnancy Data for the Study of Pharmaceutical Treatments in Pregnancy. *Curr Epidemiol Rep*. 2017;4(2):106-116. doi:10.1007/s40471-017-0104-1

266. Ihle P, Krueger K, Schubert I, et al. Comparison of Different Strategies to Measure Medication Adherence via Claims Data in Patients With Chronic Heart Failure. *Clin Pharmacol Ther*. 2019;106(1):211-218. doi:10.1002/cpt.1378

267. Dong YH, Choudhry NK, Krumme A, et al. Impact of hospitalization on medication adherence estimation in claims data. *J Clin Pharm Ther*. 2017;42(3):318-328. doi:10.1111/jcpt.12517

268. Wardle M, Spencer A. Implementation of SNOMED CT in an online clinical database. *Future Hosp J*. 2017;4(2):126-130. doi:10.7861/futurehosp.4-2-126

269. NHSBSA. Dictionary of medicines and devices (dm+d). Accessed October 28, 2019. https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/dictionary-medicines-and-devices-dmd

270. EBM DataLab, University of Oxford. BNF sections. OpenPrescribing.net. Accessed October 31, 2019. https://openprescribing.net/bnf/

271. NHSBSA. BNF SNOMED mapping. Published 2021. Accessed September 21, 2021. https://www.nhsbsa.nhs.uk/prescription-data/understanding-our-data/bnf-snomed-mapping

272. Boudreau DM, Doescher MP, Saver BG, Jackson JE, Fishman PA. Reliability of Group Health Cooperative automated pharmacy data by drug benefit status. *Pharmacoepidemiol Drug Saf*. 2005;14(12):877-884. doi:10.1002/pds.1119

273. Pit SW, Byles JE, Cockburn J. Accuracy of telephone self-report of drug use in older people and agreement with pharmaceutical claims data. *Drugs Aging*. 2008;25(1):71-80. doi:10.2165/00002512-200825010-00008

274. Haapea M, Miettunen J, Lindeman S, Joukamaa M, Koponen H. Agreement between self-reported and pharmacy data on medication use in the Northern Finland 1966 Birth Cohort. *Int J Methods Psychiatr Res*. 2010;19(2):88-96. doi:10.1002/mpr.304

275. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol*. 2010;69(1):4-14. doi:10.1111/j.1365-2125.2009.03537.x

276. Wang H, Wang B, Zhang X, Feng C. Relations among sensitivity, specificity and predictive values of medical tests based on biomarkers. *Gen Psychiatry*. 2021;34(2):e100453. doi:10.1136/gpsych-2020-100453

277. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation – A discussion and demonstration of basic features. Chiacchio F, ed. *PLOS ONE*. 2019;14(7):e0219854. doi:10.1371/journal.pone.0219854

278. RECOVERY Collaborative Group, Abbas F, Abbas M, et al. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *The Lancet*. 2021;397(10285):1637-1645. doi:10.1016/S0140-6736(21)00676-0

279. RECOVERY Collaborative Group, Horby PW, Pessoa-Amorim G, et al. Aspirin in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *MedRxiv Prepr Serv Health Sci*. Published online August 6, 2021. doi:10.1101/2021.06.08.21258132

280. RECOVERY Collaborative Group, Horby PW, Campbell M, et al. *Colchicine in Patients Admitted to Hospital with COVID-19 (RECOVERY): A Randomised, Controlled, Open-Label, Platform Trial*. Infectious Diseases (except HIV/AIDS); 2021. doi:10.1101/2021.05.18.21257267

281. RECOVERY Collaborative Group. Azithromycin in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *The Lancet*. 2021;397(10274):605-612. doi:10.1016/S0140-6736(21)00149-5

282. RECOVERY Collaborative Group. Casirivimab and imdevimab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *Lancet Lond Engl*. 2022;399(10325):665-676. doi:10.1016/S0140-6736(22)00163-5

283. Schwartz DN, Evans RS, Camins BC, et al. Deriving measures of intensive care unit antimicrobial use from computerized pharmacy data: methods, validation, and overcoming barriers. *Infect Control Hosp Epidemiol*. 2011;32(5):472-480. doi:10.1086/659760

284. Rikala M, Hartikainen S, Sulkava R, Korhonen MJ. Validity of the Finnish Prescription Register for measuring psychotropic drug exposures among elderly finns: a population-based intervention study. *Drugs Aging*. 2010;27(4):337-349. doi:10.2165/11315960-000000000-00000

285. Caskie GIL, Willis SL. Congruence of self-reported medications with pharmacy prescription records in low-income older adults. *The Gerontologist*. 2004;44(2):176-185. doi:10.1093/geront/44.2.176

286. Sutherland JJ, Morrison RD, McNaughton CD, et al. Assessment of Patient Medication Adherence, Medical Record Accuracy, and Medication Blood Concentrations for Prescription and Over-the-Counter Medications. *JAMA Netw Open*. 2018;1(7):e184196. doi:10.1001/jamanetworkopen.2018.4196

287. Siemieniuk RA, Bartoszko JJ, Zeraatkar D, et al. Drug treatments for covid-19: living systematic review and network meta-analysis. *BMJ*. 2020;370:m2980. doi:10.1136/bmj.m2980

288. The REMAP-CAP, ACTIV-4a, ATTACC Investigators, Zarychanski R. *Therapeutic Anticoagulation in Critically Ill Patients with Covid-19 – Preliminary Report*. Intensive Care and Critical Care Medicine; 2021. doi:10.1101/2021.03.10.21252749

289. Sholzberg M, Tang GH, Rahhal H, et al. Effectiveness of therapeutic heparin versus prophylactic heparin on death, mechanical ventilation, or intensive care unit admission in moderately ill patients with covid-19 admitted to hospital: RAPID randomised clinical trial. *BMJ*. Published online October 14, 2021:n2400. doi:10.1136/bmj.n2400

290. Rostami P, Heal C, Harrison A, Parry G, Ashcroft DM, Tully MP. Prevalence, nature and risk factors for medication administration omissions in English NHS hospital inpatients: a retrospective multicentre study using Medication Safety Thermometer data. *BMJ Open*. 2019;9(6):e028170. doi:10.1136/bmjopen-2018-028170

291. Walsh KE, Marsolo KA, Davis C, et al. Accuracy of the medication list in the electronic health record—implications for care, research, and improvement. *J Am Med Inform Assoc*. 2018;25(7):909-912. doi:10.1093/jamia/ocy027

292. Feder SL. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. *West J Nurs Res*. 2018;40(5):753-766. doi:10.1177/0193945916689084

293. Personal communication from NHS England. Published online February 2023.

294. Yin AL, Guo WL, Sholle ET, et al. Comparing automated vs. manual data collection for COVID-specific medications from electronic health records. *Int J Med Inf*. 2022;157:104622. doi:10.1016/j.ijmedinf.2021.104622

295. Jensen TB, Jimenez-Solem E, Cortes R, et al. Content and validation of the Electronic Patient Medication module (EPM)-the administrative in-hospital drug use database in the Capital Region of Denmark. *Scand J Public Health*. 2020;48(1):43-48. doi:10.1177/1403494818760050

296. Davenport RJ, Dennis MS, Warlow CP. The accuracy of Scottish Morbidity Record (SMR1) data for identifying hospitalised stroke patients. *Health Bull (Edinb)*. 1996;54(5):402-405.

297. The West of Scotland Coronary Prevention Study Group. Computerised record linkage: Compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. *J Clin Epidemiol*. 1995;48(12):1441-1452. doi:10.1016/0895-4356(95)00530-7

298. Barry SJE, Dinnett E, Kean S, Gaw A, Ford I. Are Routinely Collected NHS Administrative Records Suitable for Endpoint Identification in Clinical Trials? Evidence from the West of Scotland Coronary Prevention Study. Pajewski NM, ed. *PLoS ONE*. 2013;8(9):e75379. doi:10.1371/journal.pone.0075379

299. Public Health Scotland. Information Services Division. https://beta.isdscotland.org/

300. Office for Statistics Regulation. Quality Assurance and Audit Arrangements for Administrative Data. Published 2023. https://osr.statisticsauthority.gov.uk/guidance/administrative-data-and-official-statistics/quality-assurance-of-administrative-data/

301. University of Oxford. The LENS Trial. https://www.ctsu.ox.ac.uk/lens

302. Bell S, Davey P, Nathwani D, et al. Risk of AKI with gentamicin as surgical prophylaxis. *J Am Soc Nephrol JASN*. 2014;25(11):2625-2632. doi:10.1681/ASN.2014010035

303. Black M, Bhattacharya S, Philip S, Norman JE, McLernon DJ. Planned Cesarean Delivery at Term and Adverse Outcomes in Childhood Health. *JAMA*. 2015;314(21):2271-2279. doi:10.1001/jama.2015.16176

304. MacBride-Stewart S, Marwick C, Houston N, Watt I, Patton A, Guthrie B. Evaluation of a complex intervention to improve primary care prescribing: a phase IV segmented regression interrupted time series analysis. *Br J Gen Pract*. 2017;67(658):e352-e360. doi:10.3399/bjgp17X690437

305. Zhou L, Stearns SC, Thudium EM, Alburikan KA, Rodgers JE. Assessing Medicare Part D Claim Completeness Using Medication Self-Reports: The Role of Veteran Status and Generic Drug Discount Programs. *Med Care*. 2015;53(5):463-470. doi:10.1097/MLR.0000000000000341

306. Nielsen MW, Søndergaard B, Kjøller M, Hansen EH. Agreement between self-reported data on medicine use and prescription records vary according to method of analysis and therapeutic group. *J Clin Epidemiol*. 2008;61(9):919-924. doi:10.1016/j.jclinepi.2007.10.021

307. Klungel OH, de Boer A, Paes AH, Herings RM, Seidell JC, Bakker A. Influence of question structure on the recall of self-reported drug use. *J Clin Epidemiol*. 2000;53(3):273-277. doi:10.1016/s0895-4356(99)00167-5

308. Cotterchio M, Kreiger N, Darlington G, Steingart A. Comparison of self-reported and physician-reported antidepressant medication use. *Ann Epidemiol*. 1999;9(5):283-289. doi:10.1016/s1047-2797(98)00072-6

309. Taipale H, Tanskanen A, Koponen M, Tolppanen AM, Tiihonen J, Hartikainen S. Agreement between PRE2DUP register data modeling method and comprehensive drug use interview among older persons. *Clin Epidemiol*. 2016;8:363-371. doi:10.2147/CLEP.S116160

310. Noize P, Bazin F, Dufouil C, et al. Comparison of health insurance claims and patient interviews in assessing drug use: data from the Three-City (3C) Study. *Pharmacoepidemiol Drug Saf*. 2009;18(4):310-319. doi:10.1002/pds.1717

311. Wu CS, Lai MS, Gau SSF, Wang SC, Tsai HJ. Concordance between Patient Self-Reports and Claims Data on Clinical Diagnoses, Medication Use, and Health System Utilization in Taiwan. Multhoff G, ed. *PLoS ONE*. 2014;9(12):e112257. doi:10.1371/journal.pone.0112257

312. Gnjidic D, Du W, Pearson SA, Hilmer SN, Banks E. Ascertainment of self-reported prescription medication use compared with pharmaceutical claims data. *Public Health Res Pract*. 2017;27(4):27341702. doi:10.17061/phrp27341702

313. Pedersen E, Truong KNL, Garcia BH, et al. Self-reported medication use among coronary heart disease patients showed high validity compared with dispensing data. *J Clin Epidemiol*. 2021;135:115-124. doi:10.1016/j.jclinepi.2021.02.015

314. Fujita M, Sato Y, Nagashima K, Takahashi S, Hata A. Validity assessment of self-reported medication use by comparing to pharmacy insurance claims. *BMJ Open*. 2015;5(11):e009490. doi:10.1136/bmjopen-2015-009490

315. Glintborg B, Hillestrøm PR, Olsen LH, Dalhoff KP, Poulsen HE. Are patients reliable when self-reporting medication use? Validation of structured drug interviews and home visits by drug analysis and prescription data in acutely hospitalized patients. *J Clin Pharmacol*. 2007;47(11):1440-1449. doi:10.1177/0091270007307243

316. Pisa FE, Casetta A, Clagnan E, Michelesio E, Vecchi Brumatti L, Barbone F. Medication use during pregnancy, gestational age and date of delivery: agreement between maternal self-reports and health database information in a cohort. *BMC Pregnancy Childbirth*. 2015;15:310. doi:10.1186/s12884-015-0745-3

317. Monster TBM, Janssen WMT, de Jong PE, de Jong-van den Berg LTW, PREVEND Study Group Prevention of REnal and Vascular ENT Stage Disease. Pharmacy data in epidemiological studies: an easy to obtain and reliable tool. *Pharmacoepidemiol Drug Saf*. 2002;11(5):379-384. doi:10.1002/pds.722

318. Sediq R, van der Schans J, Dotinga A, et al. Concordance assessment of self-reported medication use in the Netherlands three-generation Lifelines Cohort study with the

pharmacy database iaDB.nl: The PharmLines initiative. *Clin Epidemiol*. 2018;10:981-989. doi:10.2147/CLEP.S163037

319. Boudreau DM, Daling JR, Malone KE, Gardner JS, Blough DK, Heckbert SR. A validation study of patient interview data and pharmacy records for antihypertensive, statin, and antidepressant medication use among older women. *Am J Epidemiol*. 2004;159(3):308-317. doi:10.1093/aje/kwh038

320. Drieling RL, LaCroix AZ, Beresford SAA, Boudreau DM, Kooperberg C, Heckbert SR. Validity of Self-Reported Medication Use Compared With Pharmacy Records in a Cohort of Older Women: Findings From the Women's Health Initiative. *Am J Epidemiol*. 2016;184(3):233-238. doi:10.1093/aje/kwv446

321. Allin S, Bayoumi AM, Law MR, Laporte A. Comparability of self-reported medication use and pharmacy claims data. *Health Rep*. 2013;24(1):3-9.

322. Reith C, Landray M, Devereaux PJ, et al. Randomized clinical trials--removing unnecessary obstacles. *N Engl J Med*. 2013;369(11):1061-1065. doi:10.1056/NEJMsb1300760

323. Califf RM, Cavazzoni P, Woodcock J. Benefits of Streamlined Point-of-Care Trial Designs: Lessons Learned From the UK RECOVERY Study. *JAMA Intern Med*. Published online October 24, 2022. doi:10.1001/jamainternmed.2022.4810

324. J O'Shaughnessy. Independent report. Commercial clinical trials in the UK: the Lord O'Shaughnessy review—final report. Published online 2023. Accessed November 20, 2023. https://www.gov.uk/government/publications/commercial-clinical-trials-in-the-uk-the-lord-oshaughnessy-review/commercial-clinical-trials-in-the-uk-the-lord-oshaughnessy-review-final-report

325. Wright RS, Ray KK, Raal FJ, et al. Pooled Patient-Level Analysis of Inclisiran Trials in Patients With Familial Hypercholesterolemia or Atherosclerosis. *J Am Coll Cardiol*. 2021;77(9):1182-1193. doi:10.1016/j.jacc.2020.12.058

326. Andrade SE, Kahler KH, Frech F, Chan KA. Methods for evaluation of medication adherence and persistence using automated databases. *Pharmacoepidemiol Drug Saf*. 2006;15(8):565-574. doi:10.1002/pds.1230

327. Lau HS, de Boer A, Beuning KS, Porsius A. Validation of pharmacy records in drug exposure assessment. *J Clin Epidemiol*. 1997;50(5):619-625. doi:10.1016/s0895-4356(97)00040-1

328. Christensen DB, Williams B, Goldberg HI, Martin DP, Engelberg R, LoGerfo JP. Assessing compliance to antihypertensive medications using computer-based pharmacy records. *Med Care*. 1997;35(11):1164-1170. doi:10.1097/00005650-199711000-00008

329. Rowan CG, Flory J, Gerhard T, et al. Agreement and validity of electronic health record prescribing data relative to pharmacy claims data: A validation study from a US electronic health record database. *Pharmacoepidemiol Drug Saf*. 2017;26(8):963-972. doi:10.1002/pds.4234

330. Hudson M, Rahme E, Richard H, Pilote L. Comparison of measures of medication persistency using a prescription drug database. *Am Heart J*. 2007;153(1):59-65. doi:10.1016/j.ahj.2006.10.018

331. Hess LM, Raebel MA, Conner DA, Malone DC. Measurement of Adherence in Pharmacy Administrative Databases: A Proposal for Standard Definitions and Preferred Measures. *Ann Pharmacother*. 2006;40(7-8):1280-1288. doi:10.1345/aph.1H018

332. Bharat C, Degenhardt L, Pearson SA, et al. A data-informed approach using individualised dispensing patterns to estimate medicine exposure periods and dose from pharmaceutical claims data. *Pharmacoepidemiol Drug Saf*. 2023;32(3):352-365. doi:10.1002/pds.5567

333. Stephenson JJ, Bae JP, Raval AD, Kern DM. Bridging the Gap Between Self-Reported and Claims-Derived Adherence Measures for Basal Insulin Among Patients with Type 2 Diabetes Mellitus. *Adv Ther*. 2019;36(1):118-130. doi:10.1007/s12325-018-0828-4

334. Pazzagli L, Brandt L, Linder M, et al. Methods for constructing treatment episodes and impact on exposure-outcome associations. *Eur J Clin Pharmacol*. 2020;76(2):267-275. doi:10.1007/s00228-019-02780-4

335. Min L, Ha JK, Aubert CE, et al. A Method to Quantify Mean Hypertension Treatment Daily Dose Intensity Using Health Care System Data. *JAMA Netw Open*. 2021;4(1):e2034059. doi:10.1001/jamanetworkopen.2020.34059

336. Meeker-O'Connell A, Glessner C, Behm M, et al. Enhancing clinical evidence by proactively building quality into clinical trials. *Clin Trials Lond Engl*. 2016;13(4):439-444. doi:10.1177/1740774516643491

337. NHS Digital. GPES data for pandemic planning and research (COVID-19). Published September 20, 2021. Accessed September 21, 2021. https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research

338. Godlee F. What can we salvage from care.data? *BMJ*. Published online July 14, 2016:i3907. doi:10.1136/bmj.i3907

339. Department of Health and Social Care. Data saves lives: reshaping health and social care with data. Published June 15, 2022. Accessed February 16, 2024. https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data

340. Ben Goldacre D of H and SC. Better, broader, safer: using health data for research and analysis. Published April 7, 2022. Accessed February 16, 2024. https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis

341. UK Government. Policy paper: UK Life Sciences Vision. Published June 7, 2021. https://www.gov.uk/government/publications/life-sciences-vision