Statistical methods for studying the frequency of monitoring chronic conditions



Jason Oke

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity term, 2014

The thesis is dedicated to Caroline, Charlie and Ellie.

Acknowledgements

I am extremely grateful to my supervisors Dr Richard Stevens and Professor David Matthews, who have provided consistent support and advice throughout. I am particularly indebted to Dr Stevens for his unwavering commitment to advancing my scientific career.

I would also like to acknowledge Professor Rafael Perera and Professor Andrew Farmer, as without their backing, I would not have been able to study for the DPhil.

I would like to thanks my colleagues Thomas Fanshawe, Emily McFadden, Sarah Stevens and Jennifer Hirst for their helpful comments on the thesis and their encouragement and support.

I would also like to acknowledge Dr Peter Scanlon and Irene Stratton who were willing to offer their time to answer the many questions I had about diabetic retinopathy and Professor David Dunger and Professor Neil Dalton for the ORPS data.

This thesis was funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR). The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health.

Statement of Originality

All the writing and statistical analysis was carried out by myself between September 2010 and July 2014. To my knowledge all of the work in this thesis is original, except where referenced and has not been submitted for a degree at any other university.

Abstract

Evidence for the timing and frequency of monitoring chronic conditions in primary care is often of low quality or absent. Recently, a methodology based upon repeated measures of monitoring tests has been developed to evaluate monitoring strategies which differ by frequency and the length of interval. In studies using this approach, it has been shown that there is potential for substantial over-detection and over-treatment through routine monitoring.

The aim of this thesis was to extend the existing methodology to outcomes with highly skewed distributions and tests that are interpreted as dichotomous or categorical states. These methods were applied to find evidence for the intervals of monitoring for microalbuminuria in people with type 1 diabetes and screening for diabetic retinopathy.

Statistical models were used to describe the progression of albuminuria and retinopathy and estimate the rates of false positive and false negative tests over time. Simulation was then used to create hypothetical cohorts of people being monitored and screening at different frequencies and for different baseline risk. The impact of changing the timing of tests within the programme was measured and evaluated.

The basic framework of the methodology can be usefully extended to tests that have categorical outcome or highly skewed distributions. More frequent monitoring of microalbuminuria in people with type 1 diabetes, and more frequent screening of diabetic retinopathy leads to disproportionate numbers of false positive results and potential over -treatment or referral to specialist services.

As with any other clinical intervention, monitoring and screening intervals should be based upon solid evidence. Monitoring for microalbuminuria and screening for diabetic retinopathy less frequently would reduce the rates of false positives tests but the implications of less surveillance need further investigation.

Abbreviations

ACE Angiotensin-converting enzyme.

ACR albumin creatinine ratio.

ADA American Diabetes Association.

AER Albumin excretion rate.

ARB Angiotensin receptor blocker.

CEA Carcinoembryonic antigen.

CKD Chronic kidney disease.

CV Coefficient of variation.

CVD Cardio-vascular disease.

DR Diabetic retinopathy.

eGFR Estimated glomerular filtration rate.

ENSPDR English National Screening Programme for Diabetic Retinopathy.

GP2DRS National General Practice to Diabetic Retinopathy Screening.

HbA1c Glycosylated Haemoglobin.

HTA Health Technology Assessment.

IDF International Diabetes Federation.

INR International normalised ratio.

IOP Intra-occular pressure.

MCMC Monte-carlo Markov chain.

NICE National Institute for Health and Care Excellence.

ORPS Oxford Regional Prospective Study.

PPV Positive predictive value.

PSA Prostate specific antigen.

QOF Quality outcomes framework.

RCT Randomised control trial.

SNR Signal to noise ratio.

STDR Sight-threatening diabetic retinopathy.

TIA Transient ischeamic attack.

TPM Transition probability matrix.

UAE Urinary albumin excretion.

WESDR Wisconsin Epidemiological Study of Diabetic Retinopathy.

Contents

List of Figures **vii**

List of Tables **xi**

1 Introduction **1**

1.1 Harms associated with screening and monitoring 3

1.2 The evidence base for monitoring 5

1.3 Historical perspective on evaluating monitoring and screening strategies 7

Screening 7

1.4 Factors affecting the optimal frequency of monitoring 9

Accuracy of the monitoring test 9

The probability that important changes will occur and the speed at which they occur 10

Responsiveness of the test 11

1.5 Modelling monitoring strategies 12

Models for disease progression 12

1.6 Clinical rationale for the evaluation of monitoring strategies . . 15

False positives and false negatives 16

1.7 Aims and objectives of the thesis 21

2 Models for highly-skewed continuous data: Albumin-creatinine ratio **23**

2.1 Introduction 24

2.2 Data . . . 29
Modelling or derivation data . . . 29
Validation data . . . 30
2.3 Exploratory analysis . . . 31
2.4 Methods . . . 35
Direct method . . . 35
Increasing variance model . . . 37
Flexible variance model . . . 39
T model . . . 41
Evaluation of monitoring prior to initiation of ACE-i therapy . . 45
Clinical Validation . . . 48
2.5 Model Estimation . . . 51
Direct Method . . . 51
Increasing variance model . . . 54
T model . . . 60
2.6 Evaluation of monitoring prior to initiation of ACEi therapy . . 64
Direct method . . . 65
Increasing variance model . . . 65
Flexible variance model . . . 65
T model . . . 66
2.7 External Validation . . . 68
Surrogate method . . . 68
Regression to the mean method . . . 69
Prevalence method . . . 70
2.8 Conclusion . . . 73
Statistical Validation . . . 73
Clinical validation . . . 75
Evaluating strategies for monitoring of diabetic nephropathy . . 77

3 Models for dichotomised data: Persistent micro-albuminuria 83

3.1 Introduction . . . 83

A monitoring model for dichotomous outcomes . . . 84

Accommodating tests with different short-term variability . . . 86

Aims . . . 86

3.2 Data for modelling persistant microalbuminuria . . . 88

Validation data . . . 90

3.3 Methods . . . 91

Fixed effects approach . . . 91

Random effects models (increasing and flexible variance versions) 93

3.4 Evaluating monitoring prior to initiation of ACE-i therapy . . . 94

Clinical validation of the statistical models . . . 96

3.5 Results . . . 98

Fixed effects model . . . 98

Random effects models . . . 99

3.6 Evaluation of monitoring prior to initiation of ACEi therapy . . 104

Random effects model . . . 105

Fixed effects model . . . 105

3.7 Clinical Validation . . . 107

Surrogate method . . . 107

Regression to the mean method . . . 108

Prevalence methods . . . 109

3.8 Conclusion . . . 111

Statistical validation . . . 111

Clinical validation . . . 112

4 Models for ordinal data: Diabetic retinopathy 121

4.1 Introduction . . . 121

Classification of diabetic retinopathy . . . 123

Sources of variation in screening for diabetic retinopathy . . . 123

Evaluating the effectiveness of screening . . . 126
Aims and objectives . . . 129
4.2 Data . . . 130
4.3 Methods . . . 132
Introduction . . . 132
The hierarchical ordinal model . . . 134
Evaluating screening strategies . . . 136
Validation of the models . . . 137
External validation . . . 138
4.4 Results . . . 139
Model estimation . . . 139
Prediction and internal validation . . . 141
4.5 Evaluating screening for DR . . . 146
Sensitivity and specificity . . . 146
Responsiveness of screening . . . 146
4.6 Clinical validation of models . . . 149
Comparison with estimates from diagnostic accuracy studies . . . 149
Comparison with estimates from screening studies . . . 149
4.7 Conclusion . . . 151
Restatement of findings . . . 151
Discussion . . . 154

5 Models for categorical data: Diabetic retinopathy 159
5.1 Introduction . . . 159
Markov models . . . 159
Estimating the true state and probability of misclassification: An example . . . 163
Aims and objectives . . . 166
5.2 Data . . . 167
Modelling data . . . 167

Validation data 167
5.3 Exploratory analysis 168
5.4 Methods 170
Introduction 170
Description of the model 172
Estimation of the model 174
Internal validation 174
Evaluation of different screening intervals 176
External validation 177
5.5 Results 180
Model estimation 180
Prediction and internal validation 184
State change plots 184
5.6 Screening for retinopathy and maculopathy 187
Sensitivity analysis 190
5.7 Validation 193
Comparison with estimates from diagnostic accuracy studies . . 193
Comparison with estimates from screening studies 193
5.8 Conclusion 199

6 Summary **207**
Monitoring diabetic nephropathy 208
Final remarks and indications for future research 210
In practice 213
The case against less monitoring 214
Final remarks 216

Appendices **218**

A Model estimation using hierarchical models **221**

A.1 Hierarchical models 221

Hierarchical Model for Cholesterol 222

B JAGS MCMC model code **225**

B.1 Flexible variance model 226

B.2 T model 227

B.3 Binary flexible variance model 228

B.4 Hierarchical ordinal model 229

C Variogram Method **231**

D Publications from the thesis **235**

Bibliography **237**

List of Figures

1.1 Keiding's Life history model 13
1.2 Representation of the tracking phenomenon 14
1.3 Recurrent Markov model 15
1.4 Illustration of advance time in monitoring 20
1.5 Illustration of delay time in monitoring 21

2.1 Confirmatory test algorithm for persistent microalbuminuria . . 25
2.2 Dotplots of mean and variance of ACR 32
2.3 Plot of predicted mean ACR overlaid on observed aggregated mean 33
2.4 Dotplots of mean and variance of log ACR 34
2.5 Density plots of the standard normal and t distributions 42
2.6 Caterpillar plots of residuals ordered by size of inter-quartile range 43
2.7 Observed variogram of log ACR 51
2.8 Direct method estimates of average log ACR stratified by gender with observed average overlaid 54
2.9 Direct method estimates of the variance of log ACR by duration of diabetes with observed variance overlaid 55
2.10 Comparison of the observed variance (size of the black blobs is proportional to the number of data points) with the fitted increasing variance model variance 56
2.11 Scatter plot of log ACR by duration of diabetes 57

2.12 Histograms of the estimated random effects for the linear model (a) the intercepts α_i and (b) the slopes β_i 58
2.13 Stacked boxplots of level 1 residuals from the fitted linear model (a) Females and (b) Males 59
2.14 The predicted total variation from the increasing variance (iv) model (dashed line) and the flexible variance (fv) model (solid line) against the observed data (black blobs) 59
2.15 Plots of observed proportions of assessments that would be positive with re-created outcome data from the flexible variance model 60
2.16 Histograms of the estimated random effects for the T model panel (a) the intercepts α_i and panel (b) the slopes β_i with estimated density overlaid 62
2.17 Plot of observed variance against the variance predicted from the T model 62
2.18 Plots of observed proportions of assessments that would be positive for microalbuminuria against the proportions of positive tests expected under the T model 63
2.19 Comparsion of modelled prevalence against prevalence estimates from published studies 72

3.1 Unidirectional model of CD4 count prior to clinical AIDS 84
3.2 Binned residual plot for the fixed effects model. 99
3.3 Histograms of the estimated random effects for the increasing variance logistic model 101
3.4 Scatter plot of the individual fitted intercept and slope terms from the flexible variance model 102
3.5 Observed and modelled rates of positive tests for persistent microalbuminuria 103

3.6 External validation of models for binary ACR. Symbols represent point estimates of prevalence from published studies with error bars representing 95% confidence intervals. 110

4.1 Graphical representation of the latent variable model 133
4.2 Representation of two latent models for biomarkers with small and large variability 134
4.3 Histograms of the estimated random effects for the ordinal model (a) the intercept α_i and (b) the slopes β_i 141
4.4 Observed and modelled survival from referable retinopathy . . . 142
4.5 Observed proportions of retinopathy grades by years since first screen with modelled predicted proportions overlaid. 143
4.6 Statechange plot of the observed and predicted forward transitions from the model 144
4.7 Statechange plot of the observed and predicted backward transitions from the model 145

5.1 Example of a simple multistate model 160
5.2 Directed graph of a hidden Markov model for a sequence of observed retinopathy grades for a single person 164
5.3 Illustration of the four state model for diabetic retinopathy screening 173
5.4 Modelled screening prevalence versus the observed screening prevalence for the first three states of retinopathy 185
5.5 Observed and modelled survival from referable maculopathy . . 185
5.6 Observed and modelled survival from referable retinopathy . . . 186
5.7 Statechange plot of the observed and predictions from the model 186
5.8 Comparison of Kaplan-Meier estimates of survival from referable retinopathy for the hidden Markov model, and the observed survival from the Gloucester and Lambeth data 197

5.9 Comparison of Kaplan-Meier estimates of survival from maculopathy for the hidden Markov model, and the observed survival from the Gloucester and Lambeth data 197

5.10 State change plots comparing time to event between those observed in the Lambeth data and those in the modelled data . . . 198

C.1 Variograms for a range of plausible models for longitudinal marker data. (a) a random intercept model, (b) random intercept with serial correlation, (c) random slope with small σ^2_β (d) random slope with larger s σ^2_β, (e) log variogram of model (d) 233

List of Tables

1.1 Estimated percentages of true and false positive total cholesterol tests 2

2.1 Classifications of abnormal urinary albumin excretion 25
2.2 Theoretical observed and simulated ACR data for one person . . 41
2.3 Direct method estimates for males and females separately 52
2.4 Direct method estimates of the total variance, variance of change and long-term variance component estimates for the first 8 years 52
2.5 Parameter estimates from the direct method and increasing variance model 56
2.6 Comparison of the parameter estimates for the increasing and flexible variance models 58
2.7 Comparison of the parameter estimates from the T model and flexible variance model 61
2.8 Resource and clinical implications of different intervals of monitoring under the IV model 65
2.9 Resource and clinical implications of different intervals of monitoring under the FV model 66
2.10 Resource and clinical implications of different intervals of monitoring under the T model 67
2.11 Results from the ACR study 68

2.12 Comparison of the model predictions of the rates of true and false positives, true and false negatives compared to the observed rates in the ACR Cohort study 69
2.13 Comparison of the model predictions of the amount of regression to the mean between two time points 69
2.14 Characteristics of the studies included in the external validation 71

3.1 Current NICE guidelines for monitoring people with occular hypertension or suspected chronic open angle glaucoma 84
3.2 Table of baseline characteristics stratified by persistent microalbuminuria (MA) status at the first assessment 89
3.3 A comparison of two different approaches analysing grouped data 91
3.4 Parameter estimates from the fixed-effects model 98
3.5 Estimated proportions of true negative, false negatives, false positives and true positives by duration of diabetes under the fixed effects model 100
3.6 Parameter estimates for in the increasing and flexible variance models 100
3.7 Resource and clinical implications of different intervals of monitoring prior to ACEi therapy under the random effects model . . 105
3.8 Resource and clinical implications of different intervals of monitoring prior to ACEi therapy under the fixed effect model 106
3.9 Cross-classified table of the surrogate true and persistent diagnosis using three tests from the ACR cohort study 107
3.10 Predicted true and false positives, true and false negative probabilities under the random effects and fixed effects model and the surrogate method estimate from the ACR cohort study 108
3.11 Predicted sensitivity and specificity under the random effects and fixed effects model and the surrogate method estimate from the ACR cohort study 108

3.12 Predicted and reported regression of microalbuminuria at two-year intervals 109

4.1 Classification of diabetic retinopathy 124
4.2 Frequency of agreement between first and second reading of 148 fundus photographs for presence and severity of haemorrhages and/or microaneurysms 125
4.3 Variation in grading between retina specialists 126
4.4 Baseline characteristics of the modelling cohort taken from the Gloucester screening data 131
4.5 Example formatting of the data required to fit ordinal model for DR 135
4.6 Parameter estimates from the increasing variance ordinal model for diabetic retinopathy 139
4.7 Predicted cumulative incidence (%) of true referable retinopathy after initial screened showed no DR 140
4.8 Transition probability matrix (1 year) for the hierarchical ordinal model 140
4.9 Error matrix for the ordinal model. 141
4.10 Sensitivity of screening for STDR after initial screen showed no retinopathy 146
4.11 Sensitivity of monitoring for proliferative retinopathy after detection of pre-proliferative retinopathy 147
4.12 Specificity of screening for STDR after initial screen showed no retinopathy 147
4.13 Specificity of monitoring for proliferative retinopathy after detection of pre-proliferative retinopathy 147
4.14 Responsiveness of screening for STDR from programmes with different length intervals. 148

4.15 Comparison of estimates of sensitivity and specificity for detection of retinopathy from four studies and the hierarchical ordinal model estimates. 149
4.16 Comparison of regraded images with initial screen result of no retinopathy with model predictions of screening errors. 150
4.17 Table comparing empirical estimates of referral errors with model predicted errors 150

5.1 Hypothetical transition probability matrix for a three state model 161
5.2 Hypothetical error probability matrix for a three-state model . . 163
5.3 Example of a likelihood computation for the four state model . . 165
5.4 Observed transitions between states of retinopathy and maculopathy in people undergoing diabetic retinopathy screening. . . 168
5.5 The four states representing retinopathy and maculopathy progression 171
5.6 Intensity matrix for the four state retinopathy model 173
5.7 Intensity matrix for the unadjusted four state retinopathy model 180
5.8 Average sojourn time for the first two states of retinopathy . . . 181
5.9 Transition probability matrix (1 year) for the four state model. . 181
5.10 Hazard ratios for the effect of standardised duration of diabetes 182
5.11 Hazard ratios for the effect of standardised HbA1c 182
5.12 Error matrix for the four state model. Estimates represent probability of observing state given the true state (95%) C.I. 183
5.13 Expected referrals for maculopathy and the true unobserved level of disease per 10,000 people screened, who had no retinopathy at baseline. 187
5.14 Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had no retinopathy at baseline. 188

5.15 Expected referrals for maculopathy and the true unobserved level of disease per 10,000 people screened, who had background retinopathy at baseline. 188
5.16 Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had background retinopathy at baseline. 189
5.17 Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had no retinopathy at baseline (Sensitivity Analysis). 190
5.18 Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had background retinopathy at baseline (Sensitivity Analysis). 191
5.19 Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had no retinopathy at baseline (Sensitivity Analysis). 191
5.20 Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had background retinopathy at baseline (Sensitivity Analysis). 192
5.21 Comparison of estimates of sensitivity and specificity for detection of retinopathy or maculopathy. 193
5.22 Comparison of regraded images with initial screen result of no retinopathy with model predictions of screening errors. 194
5.23 Table showing agreement between screening service grading and consensus grading for referable retinopathy 194
5.24 Table showing agreement between screening service grading and consensus grading for referable maculopathy 195
5.25 Table comparing empirical estimates of referral error rates with rates predicted by the model 195

5.26 Observed transitions between states of retinopathy and maculopathy in people undergoing diabetic retinopathy screening in Lambeth. 196

6.1 Bossuyt's staged evaluation for the development of monitoring. . 213

Chapter 1

Introduction

Monitoring or regular *surveillance* is so enshrined in clinical practice, it appears at first, absurd to suggest it should not happen at all or much less frequently than it currently does because it may be detrimental to health. There is however, a growing appreciation of the harms that results from too much testing and from over-detecting or diagnosing people. The origins of which include, but are not limited, to more people being tested because it is cheaper and easier than before, more people being treated because thresholds that define disease are lower than before [1], and more problems being found because technological advances mean we can look harder than before.[2].

Monitoring is defined as periodic assessment that guides the management of chronic disease [3]. Often, these assessments will be used to measure a patient's current state, or response to treatment but just checking symptoms or patient's wellbeing is a form of *monitoring*. It is helpful to break down the practice of monitoring into different phases [4]. The early phases of monitoring (pre-treatment phase) share many attributes with repeated screening programmes and I will draw this comparison throughout this thesis. Following identification of the clinical event, monitoring enters a titration phase assesses the response to therapy, and looks for possible side effects to treatment. Monitoring is then concerned with maintaining control, but may lapse back to the titration phase as control is lost and then back to control again.

In this thesis, I have concentrated on the early pre-treatment phases of monitoring. The aim of pre-treatment phase is to detect abnormalities long before they become symptomatic or cause a serious event. What are often referred to screening programmes have the same objective but cast their *net* (people or patients) far wider. Screening and monitoring (at least in the early) phases are both *early detection* programmes and are prone to the same problems.

In 2008, Glasziou and colleagues published a paper in which they stated that following treatment for high cholesterol (perhaps with dose-titration to reach a target value) subsequent monitoring could be much less frequent than is currently recommended and much of current testing will detect only false positives results [5].

Definition 1 *False positives in monitoring - Observed changes that are related to either short-term biological variation or analytical error [4]*

Glasziou *et al.* also suggested that retesting adherent patients every 3 to 5 years may be sufficient once an adequate response has been attained. This recommendation was based upon their calculations which showed that for intervals less than 5 years, the majority of positive tests are false positives (see table 1.1). At the time, the Medicare guidelines suggested monitoring three

Initial true level	True Level > 5 mmol/L %	True-positive results %	False-positive results %	False pos/true pos
4.5 mmol/L (176 mg/dL)				
Year 1	1.5	0.87	14	16
Year 3	13	8.9	14	1.6
Year 5	21	15	13	1
4.0 mmol/L (156 mg/dL)				
Year 1	0.0004	0.0006	1.7	> 1000
Year 3	0.7	0.43	4.3	10
Year 5	2.7	1.7	6.0	3

Table 1.1: *Estimated percentages of true and false positive total cholesterol tests*

times a year and other guidelines did similarly [5]. In the United Kingdom, the National Institute for Health and Care Excellence (NICE) made no recommendation for the frequency of monitoring in their lipid modification guideline (CG67) but did acknowledge the considerable uncertainty in a single cholesterol measurement [6]. An empirical study of the frequency of cholesterol measurement in Oxfordshire showed that for the period 2005-2007 the majority of lipid measurements were for monitoring purposes and they were taken annually or more frequently [7].

The evidence base for the intervals of monitoring cholesterol were unclear and the work of Glasziou *et al* arguably represented the first attempt at providing evidence for the frequency of monitoring cholesterol in a primary care setting. Following the cholesterol paper, Keenan *et al*[8] applied the method to monitoring blood pressure in people with a history of stroke or transient ischaemic attack (TIA) and came to similar conclusions. In 2012, Oke *et al.* applied the same model to control phase monitoring of glycosylated haemoglobin (HbA1c) in people with type 2 diabetes and also found that for people who had well controlled blood sugar levels, following the recommended intervals of monitoring led to many more false positive results than true positives tests [9].

1.1 Harms associated with screening and monitoring

Harms can be divided into three categories, those directly associated with the diagnosis itself, those associated with investigation and treatment and those associated with adverse events from treatments that stemmed from diagnosis [1]. All three are applicable to monitoring of chronic conditions. For example, regular monitoring of renal function in adolescents with type 1 may find persistent microalbuminuria when in fact the finding is due only to inaccuracy of the test or other factors which lead to temporarily high levels of albumin in the urine. This finding may trigger therapy, which is not necessary (because their

albumin is normal) and cannot benefit the patient but may cause side effects. Hence, a false positive in this case may cause physical harm. In addition, there are psychological harms. The false positive increases the burden to the patient as they may feel that this is evidence of their condition worsening and this may in turn have further detrimental effects. Short-term (day to day) variation in biochemical parameters can lead to long term over-treatment of healthy people. In a systematic review of trials comparing ambulatory, with clinic and home blood pressure monitoring, Hodgkinson *et al*[10] estimated that both clinic and home monitoring had poor specificity (clinic = 74.6%, home = 62.4%) and speculated that if used as a screening test for hypertension in people under 40 years of age, clinic measurement would lead to three out of every four positive diagnoses being false positives and potentially one-third of people who have been diagnosed with hypertension and are being treated have only white-coat hypertension.

Definition 2 *Overdiagnosis in monitoring - Over treatment in patients who would never otherwise require it or would improve without intervention*

In some cases monitoring cause unnecessary anxiety and inconvenience. In a programme on BBC Radio 4 (Inside Health) a women described how she had experienced a false alarm from a test she did not even know she'd had.

> " When I called back for the results I was told that an additional test had been done to check kidney function, that this had given rise to some concerns and there was nothing to worry about but they would like to investigate to see whether there was in fact a problem with my kidney function. This all took six months. I was extremely annoyed because I hadn't given informed consent to the test being taken place and it hadn't been properly explained to me why it was being done or what the actual expected outcome or the problems that they had diagnosed actually were and I was left in doubt for six months, only to be told that in fact there was no problem at all. They now knew my kidney function level and this was a benchmark against which they could measure the deterioration of my future life and if it deteriorates at the rate that they predict without treatment I would need kidney dialysis aged 90. I thought this was completely

> ridiculous because I believe that by that time I would probably have developed other medical conditions which may actually have prevented me from reaching 90". [11]

A goal of screening and pre-treatment monitoring is to detect consequential disease early and this drives enthusiasm for more frequent and detailed investigation. The experience gained from many decades of screening is that often the harder you look the more you find and often much of the excess diagnoses represents indolent disease which will not progress and threaten life [12]. There is certainly the potential for the similar problems to arise in monitoring chronic conditions.

1.2 The evidence base for monitoring

There have been many studies examining the efficacy of self-monitoring in a number of chronic conditions that are managed in primary care. Some examples include self- monitoring of blood sugar in people with diabetes [13], self-monitoring of the international normalised ratio (INR) for people on warfarin [14] and mobile phone based monitoring in asthma [15]. There are also many studies of monitoring that takes place in secondary care setting including intracranial pressure monitoring [16], monitoring women at risk for pre-term labour [17] and continuous electronic physiological monitoring [18] to cite but a few.

High quality randomised control trial's (RCT's) evidence for monitoring of any kind is the exception rather than the norm and very few studies address the issue of intervals or frequency of monitoring. As a consequence, guidelines rarely provide explicit recommendations for the interval of monitoring. Moschetti *et al* reviewed guidelines for the prevention and treatment of cardiovascular disease and found that 51% of them provided no information on when to monitor or what to do if the measurements were out of range [19]. When guidance is given it is often based upon consensus rather than trial evidence. For example, in a systematic review for the U.S Preventative Services

Task Force (USPSTF) of screening and monitoring the treatment of Chronic Kidney Disease (CKD) stages 1 - 3, the authors found 110 RCT's for drug effectiveness trials but no RCT's of CKD screening or monitoring, no trials that directly showed a benefit for screening or monitoring and the harms of monitoring poorly described [20]. However, the NICE guidelines for CKD (CG73) provide recommendations for the frequency of monitoring kidney function (with eGFR) for people with stages 1-5 CKD, although they acknowledge that the information provided is based on consensus of opinion and not on evidence [21]. In a position statement from the American Diabetes Association (ADA) on the standards of care in diabetes [22], recommendations are made for monitoring blood sugar, renal function and diabetic retinopathy. Recommendations for HbA1c monitoring are given but are not specific;

> "Measurement approximately every 3 months to determine whether a patient's glycemic targets have been reached and maintained"

Annual screening for nephropathy and retinopathy is recommended and evidence is rated as "supported" from well-conducted prospective cohort studies. In each case, high quality research for optimal strategies is lacking. Perhaps, this is not surprising as even for breast cancer screening with mammography, an area of research which has seen enormous investment and 8 large scale randomised control trials of screening in over 600,000 women [23] there remains a question over the optimal interval and timing of screening. For breast cancer screening, there is no direct evidence for the interval of screening, or when it should start and finish and it remains unlikely that further trials will be conducted to investigate the mortality benefits from extending beyond the current age limits or to test different test schedules [24]. Ultimately, decisions about monitoring should be grounded on solid evidence and specific monitoring strategies should not be used in practice unless it can be shown that it improves or maintains patients health and is cost effective [25].

1.3 Historical perspective on evaluating monitoring and screening strategies

Even though there are many similarities, research on the effectiveness of monitoring and the effectiveness of screening has largely developed independently. A considerable amount of research has gone on to evaluating screening, mainly for cancer, and this work often uses models to measure population outcomes. In contrast, what research there has been into evaluating the effectiveness of monitoring, has until quite recently, focused on the properties of tests in isolation and not considered effectiveness of monitoring programmes as applied in practice.

Screening

In the 1960's, researchers (particularly mathematicians) had turned their attention to the problem of evaluating alternative policies of repeated medical examination [26] [27], culminating in the development of a general theory of screening for chronic disease, starting with the work by Zelen and Feinlab [28] and later Prorok [29][30]. Research in this area diverged with considerable attention being paid to particular aspects of screening, such as the estimation of lead-time, detection rates of pre-clinical disease, tumour progression rates, over-diagnosis and expected mortality benefits. A recent appraisal of the benefits and harms of mammography screening for breast cancer [24], where the authors used Markov type simulation models to measure benefits in terms of mortality reductions and life years gained for a range of different intervals, provides a good example of the type of methods that are typical in research evaluating screening.

Monitoring

At around the same time, other researchers had started to investigate effects of between-person and within-person variability on repeated measures of clinic

measurements. Gardner and Heady [31] proposed a general model for the case in which repeated measurements on individuals give different results, and this provided a framework for others to quantify regression to the mean in epidemiological studies and controlled trials, and later to estimate misclassification in repeated examinations [32]. Gardner and Heady showed that misclassification will always occur as long as the within-subject variation is not zero (the test is not perfect) and proposed an equation to gauge the reliability of a measurement. Their *coefficient of reliability*, G, is the ratio of the between person variation and the sum of the between person-variation and within-person variation. When there is no within-person variability, which could (theoretically at least) occur with a perfect test and a completely stable condition, the coefficient of reliability is 1. As the within-person variation increases G goes to zero. Hence, it is a measurement of agreement or confidence you could have if you were able to re-test. Wilson *et al* [32] extended the basic model to account for dependence between measurements on an individual and Johnson and George [33] extended the method separating the within-person variation into subject effects and measurement error. However, both of these studies focussed on quantifying the degree of regression to the mean rather than misclassification rates.

Some years later, Thompson and Pocock [34] examined the variability of serum cholesterol and its effect on monitoring and screening interventions. They estimated rates of misclassification based upon knowing the true cholesterol and determining the within-person variation from published estimates of the coefficient of variation (CV). They stated that if the CV was 7.4% and a persons' true cholesterol was 7.0 mmol/L then 95% of measurements would fall between 6 mmol/L and 8 mmol/L. The concluded that monitoring (for the effect of interventions) may not be helpful and could potentially be harmful.

Glasziou *et al*[5] extended the general misclassification model to include population and subject-specific drift in the population being measured and used it to find evidence for intervals of monitoring cholesterol. With the exception

of the work by Kimmel *et al* on optimal frequencies for CD4 cell count [35] this represented a first for evaluating monitoring programmes in a primary care setting. This approach has since been applied to a number of chronic conditions including hypertension [8], diabetes [9] and glaucoma [36].

1.4 Factors affecting the optimal frequency of monitoring

The optimal interval for monitoring can be thought of one which maximises the benefits, minimises the risks and harms and is balanced economically in terms of the total health-care expenditure [37].

The optimal interval of monitoring will depend on the following factors:

- The accuracy of the monitoring test
- The probability that clinically important changes will develop, and the speed at which they develop, and
- The responsiveness of the test

Accuracy of the monitoring test

The accuracy of a monitoring test is often described in the context of the signal to noise ratio (SNR) or the coefficient of variation (CV). The concept of SNR originates from engineering and is has been used to describe issues around monitoring [38]. The SNR was developed to measure the strength of electrical signals in the presence of background interference. It can be thought of as the ratio of useful information to false or irrelevant information. The signal in the context of monitoring is often used to describe clinically important changes over time and the noise to mean the all other sources of background variability. When trying to establish a baseline before initiation of treatment, the signal is more likely to mean the true level at a particular point in time, whereas in pre-treatment or control phase monitoring the change may be more

informative. In establishing a baseline, the noise will comprise of both short-term variation attributable to the test procedure and biological variation and also the distribution of true levels in the population.

Accurate assessment of a baseline level depends on the amount of short-term variation (σ^2) which is likely to be a combination of both measurement error and biological variation. Biological variability can often be reduced by strategies such as measuring at the same time of day or taking measurements at rest and measurement error can be minimized by standardising the measurement process [39]. Taking multiple measurements can also reduce this variability. The average error, or degree of short-term variation, tells us something about the reliability of measurement but does not capture a phenomena known as regression to the mean. Regression to the mean is statistical term for the commonly observed phenomena of extreme measurements tend to more normal measurement on re-testing. Regression to the mean depends on both the measurement error and the how variable the population is (τ^2).

The probability that important changes will occur and the speed at which they occur

Monitoring intervals are in general dictated by the speed and the probability that clinically important changes will occur in the phase of monitoring being considered and the population under surveillance. In monitoring the response to therapy, changes will happen quickly and so the interval will be short, whereas in monitoring for complications of chronic diseases (pre-treatment phase), clinically important changes can take years to develop and affect only a minority, therefore the interval will be longer. When the probability of significant clinical change (signal) is low then monitoring or screening can be cost-ineffective and potentially harmful. In screening, repeat testing in asymptomatic or healthy people often results in large numbers of false-positive results and the discovery of non-progressive disease (over-diagnosis). Starting monitoring for the complications of chronic conditions too early in the course of disease,

or in stable and well-controlled people will also produce many false positive results and lead to unnecessary treatment changes and further testing. For example, monitoring for microalbuminuria in type 1 diabetes is recommended to start at age 11 (IDF guideline) and 12 (NICE guideline) years of age to reflect the fact that microalbuminuria is rarely detected before puberty [40]. In contrast, screening for diabetic retinopathy starts immediately following diagnosis of type 2 diabetes and this reflects the significant risk of having background retinopathy or worse by the time a clinical diagnosis of diabetes is made. In the Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR), 20% of people had some degree of retinopathy five years after diagnosis of type 2 diabetes [41].

Responsiveness of the test

Definition 3 *Responsiveness - How quickly does the test respond to changes in the treatment or condition [39]*

The responsiveness of the test will be an important consideration in determining intervals for some phases of monitoring. For example, HbA1c is used to monitor blood glucose levels in people with diabetes and is thought to reflect average glycaemia over the preceding 6-8 week period. This is attributed to the long lifespan of red blood cells (120 days) [42]. Hence short-term changes in blood glucose concentrations may not be reflected in HbA1c measurements taken closely in time. Therefore it seems sensible to set a lower boundary of three months for the intervals of monitoring in the initiation or titration phase as the full effect of any change to therapy may not be reflected in the test. However for long-term monitoring, where intervals will usually be longer (6 months - 15 months), the slow responsiveness of HbA1c is seen as a positive attribute compared to other measures of glucose control.

1.5 Modelling monitoring strategies

In the absence of direct randomised control trials of monitoring or screening, researchers have tended to use stochastic models. These are usually mathematical models of the natural history with monitoring strategies and treatment effects incorporated so that the costs, clinical benefits and cost effectiveness of different strategies for monitoring can be compared. Estimates are often derived through simulation, and models can be built so that many strategies can be compared and the sensitivity to parameter values can be assessed. Despite its attractions, modelling is not a replacement for well run RCT's, but it does allow identification of the set of approaches to monitoring that are most likely to be effective [4].

Models for disease progression

Monitoring a disease needs to take account of the underlying natural history of disease progression [4]. Many diseases exhibit a heterogeneous progression; in some people deterioration and severe illness can develop rapidly whereas in some progression can be very slow and never go on to cause serious problems in their lifetime. For example, in diabetic nephropathy, it is now recognised that microalbuminuria remains stable in some patients, whereas in others it abates transiently or even permanently [43]. Many conditions exhibit a rate of progression in the early stages that is positively correlated with the rate of further progression. For example, people with type 2 diabetes have an underlying pathology of different rates of failure of the pancreatic beta cells. In the Belfast diet study, a 10-year prospective cohort trial of 432 newly diagnosed diabetic patients aged $40-69$, increases in fasting plasma glucose varied from 0.2 mmol/L per year to 1.5 mmol/L [44]. Random effect or growth curve models are well suited to modelling the natural history of diseases such as these and are able to estimate the amount of variation due a number of sources simultaneously.

Markov models

For screening evaluation, Markov models have often been used to describe the underlying progression of disease, including progression of bronchiolitis obliterans syndrome in lung cancer patient [45], breast cancer screening [46] and CD4 count [47]. The progression of disease can often be described as a series of stages or conditions relating to surrogate outcomes for the underlying pathology. A simple example is shown in figure 1.1. A strong assumption

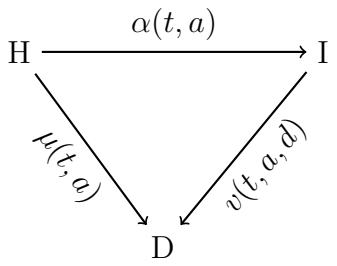


Figure 1.1: *Life history model from Keiding [48]; H - Healthy, I - Illness, D - Death, t - calendar time, a - age, d - duration*

implicit in these models is the Markov assumption, where the future course of the disease depends only on the current "true" state and not disease history. When a disease has a low number of states this assumption will usually hold, however when modelling a disease with a finer gradation of disease severity, the Markov assumption may be unrealistic.

Tracking models

Markov and random effects models imply different prognoses for patients in individual-patient simulations. For example, consider monitoring for diabetic nephropathy which uses a continuous measurement (mg/mmol of albumin to creatinine) but is often interpreted as categorical stages of disease; normo, persistent microalbuminuria and macroalbuminuria. In the random effects model, individuals who progress quickly from normo to micro will also progress quickly to macroalbuminuria and beyond (see figure 1.2). In contrast, in a multi-state

Markov model, the time to transition from a normo state to the micro state is independent of further progression to macroalbuminuria. That does not mean that in Markov models rapid progression cannot occur from both normo to micro and micro to macro for some individuals, it may, but only by the play of chance. This feature, known as *tracking* may be an accurate representation of disease progression. Tracking can be accommodated into Markov models but its use is limited [49]. Markov models are however very flexible

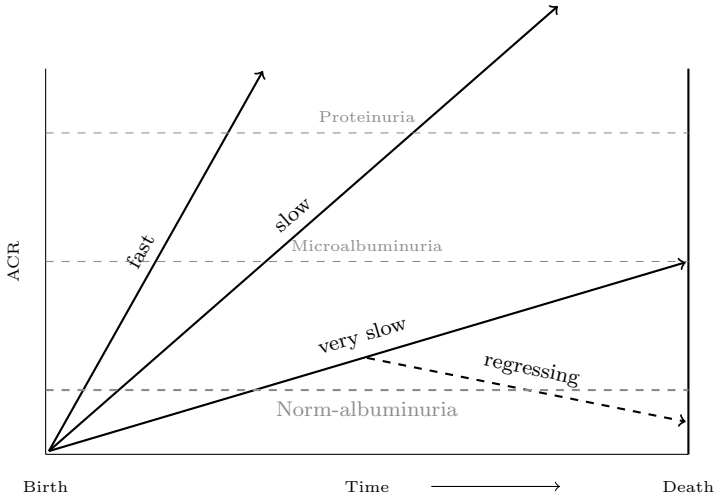


Figure 1.2: *Representation of the tracking phenomenon inherent in a linear model for progression and variation using stages of diabetic nephropathy as an example*

and accommodate complex disease patterns which are challenging using random effects models. For example, Saint-Pierre *et al.*[50] analysed longitudinal data of physician graded control scores in asthma patients over a four year period. Control scores were categorised as one of three states; *optimal control*, *sub-optimal control* and *unacceptable control*. The authors used a three-state recurrent Markov model. Figure 1.3 is a diagram of the model used. A model such as this would be very difficult to emulate using a random effects model structure.

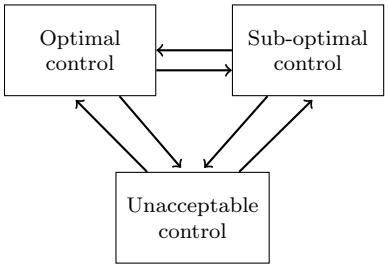


Figure 1.3: *Recurrent Markov model for asthma control (from Saint-Pierre et al) [50]*

1.6 Clinical rationale for the evaluation of monitoring strategies

> Screening has distorted public belief. In our desire for good population coverage we have said that screening is simple, effective, and inexpensive. In truth, it is complex, of limited effectiveness, and very expensive. *from "The Screening Muddle", Lancet, 1998*

Since the 1960's, the number of screening tests for cancer and other diseases has risen exponentially but evaluations of screening effectiveness have not kept pace[2]. The failure to evaluate screening strategies has led to some costly mistakes and similar errors are now being made with monitoring [4] and it is only recently that the issue of evaluating monitoring strategies has received the attention it deserves.

Monitoring is a complex intervention involving multiple components and adaptive decision-making on the part of the clinician and patient [4]. However, monitoring and screening programmes have often been rolled out unanimously, with very little effort to find optimal times and frequencies. It may not be necessary to monitor everyone immediately following diagnosis, and the frequency of monitoring could vary between groups at different levels of risk diagnosis. It is likely that the effectiveness of monitoring will be enhanced by a greater understanding of when and how often to monitor, who should monitor and what action should follow. Monitoring may not always be beneficial and should not be assumed too.

False positives and false negatives

A large proportion of the harms due to screening and monitoring are linked to false positive or false negative tests. They are also commonly major factors in cost-effectiveness analyses and strategies that minimise both are highly desirable.

False positives

Previous research on monitoring cholesterol, blood pressure and HbA1c concluded that frequent monitoring led to many more false-positive results than true positives. In populations where the risk or prevalence of the adverse event is low (such as pre-treatment monitoring), a false-positive is much more likely to occur than either a true positive or a false negative result. A nice explanation for this phenomenon is given by Shuster [51]. He uses the context of his area of expertise; malignant melanoma but the ideas can be applied to any similar type of disease seeking. He argued that if you have a test to detect malignant melanoma that in 10% of the time says that a benign lesion is malignant (false positive) and 10% of the time it says that malignant tumours are benign (false negative), then intuitively you would expect the errors to occur equally. However, if 100 people (for the sake of argument) appear in a year with lesions that have a borderline histological appearance but are all benign, then, with diagnostic error of $\pm 10\%$, 10 will be reported as having a grade 1 malignant melanoma, but, as none had true malignant melanoma, none will be missed (0 false-negatives). This "bias" towards more false-positives will always occur until the prevalence (the number attending the clinic) exceeds 0.5, where-after the bias will "flip".

There is an analogous result for tests that are continuous. If the distribution of cholesterol in people attending a clinic is 5 mmol/L with standard deviation 1 mmol/L and, as before, a test used in the clinic has an average error of $10\% = 0.5$ mmol/L, then using a cut-off of 6 mmol/L, 15 in every 100 people

attending will have true cholesterol above the threshold and 85 will not. Of the 85, 6 will be false-positive for the threshold of 6mmol/L but only 3 will be false-negative. In a "healthier" population with mean cholesterol 4.5mmol/L, an equivalent test produces 4 false-positive test and 1.5 false-negative tests per 100 people.

Another consequence of using imperfect tests and a low prevalence rate is that estimates of prevalence are biased [52]. For any imperfect test (sensitivity (se) and/or specificity (sp) less than 1), the standard estimate of prevalence (ϕ = no. of cases divided by the total number of tests) is biased for the true prevalence (θ). Positive cases will consist not only of true positives (se * θ) but also false positives ($(1 - sp)*(1 - \theta)$. The non-cases will be comprised of true negatives and false negatives but as the false positives will outnumber false negatives, the usual estimate of the prevalence will be positively biased. For example, a test with sensitivity of 0.9 and specificity of 0.8, the bias is positive until the true prevalence is greater than 2/3. As the objective of early-detection programmes is to minimise false negatives (high sensitivity), often at the expense of false positives (specificity), and usually for diseases which are uncommon, false positives will be the dominant error made by monitoring and screening.

False negatives

Definition 4 *False negatives in monitoring - Missed opportunities to detect clinically important changes due to biological or analytical variation.*

In early screening studies, the occurrence of false-negative results was often the focus of attention due to the importance attributed to early detection and the reasoning that any false positive result would later be resolved. It is in fact very difficult to estimate the rates of false negatives with any precision because unlike people with a positive screen result who may have independent verification of true disease state, individuals with negative findings are often

not followed up [53]. An exception to this is the example of cervical smears collected during 1990-1995 in Kent, England, when 91,000 cervical smears were rechecked and 333 women who had been told that their smears were normal were in fact at high risk of progressing to cervical cancer, eight of these women died of cervical cancer and 30 underwent hysterectomies [54]. This represents a small percentage (0.4%) of the total test but the potential costs are very high. Without an immediate confirmatory test it is very difficult to distinguish true false negatives (errors in detection) from interval cases (disease that develops in between examinations and were occult at the time of screening). In retrospective analyses of cancers that developed between screening intervals, the Northern Breast Screening Radiology Audit Group found that of 167 interval cancers, 46% were true interval cases, 26% were false negatives, 11% were occult and 16% were not classifiable [55].

False negatives can be equally important in pre-treatment monitoring and in some cases failure to diagnose problems could have devastating consequences. Failure to detect sight-threatening diabetic retinopathy can lead to severe vision loss and blindness at a young age. In addition, a false-negative may give false reassurance and prevent someone from seeking help when they would have done otherwise. False negatives are more important in conditions that have a long pre-clinical or asymptomatic phase and virtually no clinical or symptomatic phase before a major event, for example, an abnormal prothombin time (INR) in someone taking warfarin may be asymptomatic until a major bleed [4]. With regular monitoring, a false negative test on one occasion is likely to be corrected later and thus the patient incurs a delay in an intervention and treatment change. In cases where timely intervention is known to be important, such as photocoagulation therapy in diabetic retinopathy (this is because vision loss cannot be reversed), false negatives and subsequent delays will cause harm but in examples where false negatives mean only opportunities to adjust treatment are missed very little is known about the possible harms.

Advances or delay in treatment initiation or change

In evaluating monitoring strategies with different intervals it is important to consider not only the number of false positive and false negative tests but also how long someone is a false positive or false negative for. In situations when a disease is progressive, a false-positive test can just mean an advancement in the timing of treatment changes (advance time) and a person may begin as a false positive (test positive, but underlying disease negative) but later become a true positive when the disease catches up with the diagnosis. Knowledge of the length of time a patient is being *over-treated* will be important for evaluation of harms and benefits. Similarly, for conditions that regress transiently or permanently (such as microalbuminuria) a false negative status (underlying disease positive, but test negative) may be temporary as they later regress and become a true negative. In this case, they may benefit from the error, as they will not be treated (with a treatment that cannot benefit them as they are not unwell) and they will be spared from potential side effects of treatment and the anxiety from being diagnosed.

Longer intervals of monitoring will increase the length of time a person goes undetected and delay intervention, which if it is effective, will mean the delay will be harmful. Early mathematical modelling approaches investigating recurrent medical examinations used this as the sole criteria for determining the optimal interval[26], paying no regard to the probability of false detection.

In theory, any test used for monitoring that is neither 100% sensitive nor 100% specific there exists, for each person a time in which the "diagnosis" has been advanced by a false-positive test (advance-time) or has been delayed (delay-time) as a result of a false-negative test. In the case, in patients which do not go on to develop the later stages of disease and do to test false positive, both will be zero. The *advance* time is the time from a false-positive test to the time when the "true" value of the marker being monitored becomes *positive*. Figure 1.4 illustrates the concept of advance time which is in terms of a true

state which is either *below* a threshold or *above* a threshold, but is generalisable to disease states which are either positive or negative. The advance time is

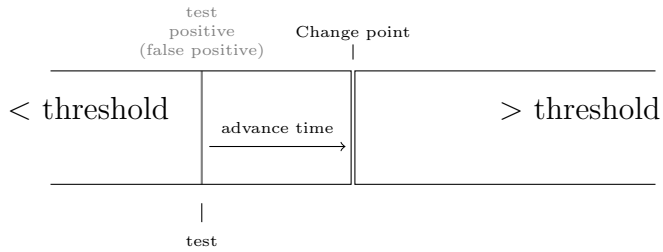


Figure 1.4: *Illustration of advance time in monitoring*

not directly affected by the length of the interval between assessments but is affected by the rate of progression and the short-term variability. The upper bound to the advance time could, in theory, be very long if a false positive result occurs early and treatment starts and carries on until a person's death. The delay time will be zero if the advance time is non-zero and also have an upper-bound of a person's remaining years. Note that the *advance* time is very similar in nature to lead-time used in cancer screening theory. However, lead-time is measured from the point at which screening detects an abnormality until the time it would have become symptomatic or detectable clinically. The *advance time* does not depend on the condition ever being symptomatic or detectable by other means, only that it has progressed to a certain point.

A delay in intervention will occur if a person becomes incident between monitoring assessments or has a false negative test. Unlike the advance time, the delay time will be a function of the interval between monitoring assessments and also the false-negative rate. Figure 1.5 is an illustration of the concept of delay time in monitoring.

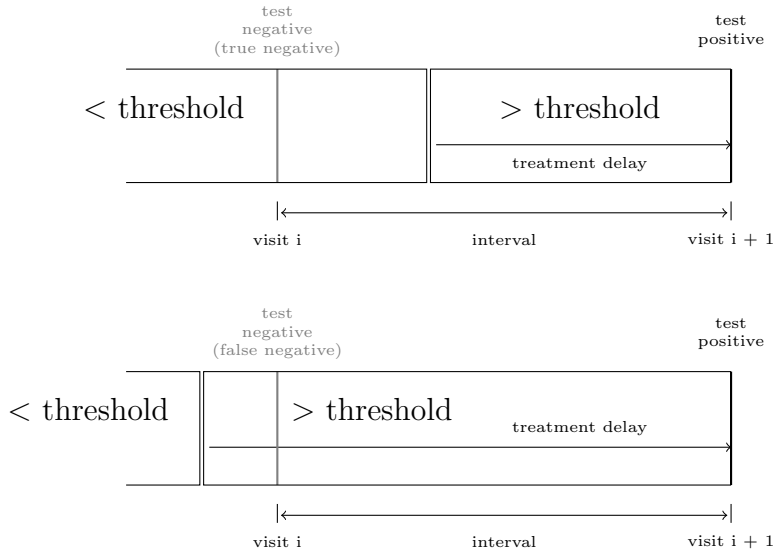


Figure 1.5: *Illustration of how detection/treatment delay can occur in monitoring, as interval cases (top panel) and one a patient has a false-negative test (bottom panel)*

1.7 Aims and objectives of the thesis

This thesis grew out of work on two separate Health Technology Assessment (HTA) funded projects, both of which were focused on monitoring for complications of diabetes. Kidney disease (diabetic nephropathy) and eye disease (diabetic retinopathy) are common and important complications of diabetes and annual monitoring is offered either immediately following or soon after diagnosis of diabetes. The aim of the first project was to find optimal intervals of monitoring diabetic nephropathy using albumin creatinine ratio (ACR) testing in primary care, and the aim of the second project was to find the optimal intervals for screening for diabetic retinopathy using digital fundus photography.

Previous work on cholesterol, blood pressure and HbA1c monitoring had shown that using just repeated measurements data from cohort studies, you could quantify the number of false positive and false negative tests you were likely to get from monitoring at different intervals. However, both of the outcomes for monitoring (ACR and retinopathy grade) represented challenges to the assumptions underlying the method, so an initial task was to assess whether

the original model could be usefully applied to these measures and others of a similar nature.

The aims and objectives of the thesis evolved to include exploring ways to quantify the uncertainty in the predictions (rates of false positives, false negatives etc) from the models, methods to check the goodness of fit and ways to validate predictions against new or existing published data. In addition, I wanted to find new metrics for evaluating monitoring strategies and changing frequency of monitoring.

The layout of the theses is presented in a format akin to four-case studies in evaluating monitoring strategies. There are two clinical areas and each is handled in two ways. Chapters two and three are based upon monitoring albumin-creatinine ratio in people with diabetes. The first approach (chapter 2) represents my early attempts at tackling the initial problem and focuses on comparing and assessing different models for continuous ACR and how the choice of model effects subsequent predictions. In chapter 3 (also monitoring ACR)I developed a model for progression albumin-creatinine ratio as a dichotomous outcome and used it evaluate monitoring strategies. Chapters four and five are based on the screening and monitoring for diabetic retinopathy. In the first approach (chapter four), my intention was to extend the linear method further, building on the model for dichotomous ACR to grades of retinopathy, and more generally, any ordinal outcome. The work on diabetic retinopathy represents my later work and includes more in-depth evaluation of the monitoring strategy as a whole, including discussions about how monitoring strategies may be appraised from a range of perspectives (responsiveness, test validity etc). The last case study on retinopathy, uses an established method for describing disease progression (Markov models) and my intention was to compare and contrast this method with the linear model based method and see whether monitoring strategies could be evaluated.

Chapter 2

Models for highly-skewed continuous data: Albumin-creatinine ratio

Monitoring the progress of diabetes and the complications arising from diabetes is a major health-care activity. Diabetic nephropathy is a common complication of diabetes and occurs in $20 - 40\%$ of all people with diabetes and is the single leading cause of end stage renal disease [22]. Microalbuminuria has been shown to be the earliest stage of diabetic nephropathy and early intervention is believed to prevent or slow the progression of the disease. Hence, monitoring or screening for the early stages of the disease are considered basic standards of care in diabetes management.

In this chapter, I describe different approaches to modelling the natural history of diabetic nephropathy using repeated measures of the albumin creatinine ratio from a prospective cohort study. In doing so, I extend the method proposed by Glasziou *et al* [5] for monitoring cholesterol levels.

Diabetic nephropathy is monitored by assessment of the albumin to creatinine ratio. The albumin-creatinine ratio (ACR) has a highly positively skewed distribution even after transformation by natural logarithms and an initial research question was whether the methods used to find evidence for the intervals of monitoring cholesterol could be *usefully* applied to ACR and similar *non-normal* measures.

To determine whether the methods could be usefully applied, I assessed the statistical and clinical validity of these approaches and developed a simulation model of pre-treatment monitoring with ACR in adolescents and young adults with type 1 diabetes. Using the simulation model I derived a number of statistics to capture the effects on changing the interval of monitoring.

2.1 Diabetic nephropathy

In the early 1980's, three landmark studies of patients with type 1 diabetes showed that microalbuminuria (elevated urinary albumin excretion) led to a 60 to 85 percent risk of developing overt proteinuria [43]. Although current estimates of risk are closer to half these initial estimates, with around 30% of type 1 patients progressing to overt nephropathy [40], microalbuminuria is rightly regarded as important health issue for people with diabetes. Persistent raised microalbuminuria has been shown to be not only a marker of the development of nephropathy, but is also a well established risk factor of cardiovascular disease (CVD). Following the detection of microalbuminuria, it is has been shown that intensive management, including control of glucose levels and blood pressure and restriction of dietary protein may delay the onset of increased urinary albumin excretion in people with type 1 [56] and type 2 diabetes [22].

Urinary albumin excretion can be measured by three methods: 1)albumin-to-creatinine ratio in a random spot collection; 2) 24-h collection with creatinine, allowing the simultaneous measurement of creatinine clearance; and 3) timed (e.g., 4-h or overnight) collection. In the clinical setting, monitoring of albumin excretion is done by assessing ACR from random spot urine samples as the others are impractical [57] and add little to the accuracy [58]. As there are many methods for measuring albumin excretion and different units to express them it can be quite difficult to compare results across the literature. Table 2.1 shows three of the more common ways to express albumin excretion and the thresholds for classifying degrees of nephropathy. Current guidelines from

Category	24-H Urine Albumin (mg/24h)	Overnight Urine Albumin (μg/min)	Spot Urine			
			Albumin (mg/L)	Albumin/Creatinine ratio		
				Gender	mg/mmol	mg/g
Normoalbuminuria	< 30	< 20	< 20	M	< 2.5	$10 - 20$
				F	< 3.5	$15 - 30$
Microalbuminuria	$30 - 300$	$20 - 200$	$20 - 200$	M	$2.5 - 25$	$20 - 200$
				F	$3.5 - 35$	$30 - 300$
Macroalbuminuria	> 300	> 200	> 200	M	> 25	> 200
				F	> 35	> 300

Table 2.1: *Classifications of abnormal urinary albumin excretion, after de Jong and Curnhan [59]*

the NICE, the American Diabetes Association (ADA) and the International Diabetes Federation (IDF) recommend annual testing to assess albumin excretion using random spot collection and simultaneous measurement of creatinine. Because of the large short-term variation in albumin excretion, all three guidelines suggest that persistent raised albumin is confirmed by two from three tests over a 3-6 month period (ADA), within 3-4 months (NICE) see figure 2.1. No guidance is given for the timeframe of remeasurement in the IDF guidelines.

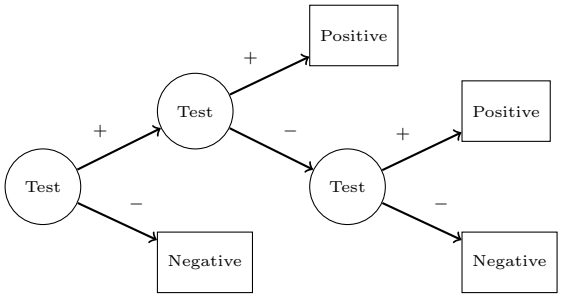


Figure 2.1: *Confirmatory testing algorithm for within-year assessment of persistent micro-albuminuria in people with type1 diabetes*

A prerequisite for monitoring is that the condition must have a sufficiently long asymptomatic phase (the phase prior to symptoms or events) so that action taken to slow or prevent consequential disease and confer an advantage over clinical presentation at a later stage with regards to mortality, morbidity or quality of life. Diseases with complications that have long asymptomatic phases are obviously good candidates for monitoring, whereas conditions with

short asymptomatic phases are often not. These conditions may seem to be satisfied in many cases under the assumption that problems can be detected quickly after they begin (monitoring is sensitive and responsive) and monitoring does not over-detect (monitoring is specific).

In a randomised control trial in people with type 2 diabetes, hypertension, and raised albuminuria, treatment with ACE inhibitor or ARB treatment reduced rates of progression to nephropathy [60] and a cost-effectiveness analysis based on these findings suggested that intervention at the stage of microalbuminuria (UAE $20 - 199$ μg/min) was favourable over treatment starting later when overt nephropathy (UAE > 200 μg/min) had developed [61]. There is evidence that nephropathy begins soon after and accelerates around the time of puberty in people with type 1 diabetes and monitoring usually begins around this time. However, there is currently no evidence of benefit from early intervention with ACEi inhibitors is adolescents and young adults and recommendations for treatment are inconsistent across guidelines.

As microalbuminuria may abate transiently, stabilise and even regress to normoalbuminuria permanently [43], many early diagnoses of microalbuminuria may represent overdiagnosis and overtreatment. Overdiagnosis has the potential to occur in two ways; firstly from false positive tests in people who never go on to develop true microalbuminuria within their lifetime and secondly, from people with true transient microalbuminuria who are true positive but would later regress back to normal levels of albuminuria irrespective of intervention. In both cases it is possible that this overtreatment will persist and not be detected because subsequent monitoring showing normal levels of ACR will be considered as evidence of compliance and effectiveness of the treatment. Given that ACR has such a low signal-to-noise ratio it is possible that much of the perceived regression of raised albuminuria could only be regression to the mean effects or chance variation and not real remission of disease.

As far I aware, there have been no formal assessments of the clinical benefit or costs for more frequent or less frequent monitoring of ACR, or the identifica-

tion of specific sub-groups for which monitoring frequency could be intensified or relaxed. Neither have there been studies on when monitoring should begin in relation to the duration of disease or the age of people with type 1 diabetes. Due to this lack of evidence, guideline recommendations for many aspects of monitoring for microalbuminuria in type 1 diabetes differ across the three major guideline bodies. The starting point for monitoring for microalbuminuria is for once duration of diabetes is ≥ 5 years in the ADA guidelines, from age 12 in the NICE guidance and the from 2 years duration and 11 years of age in the IDF guidelines. Treatment recommendations are also inconsistent across these three, with the IDF and the ADA recommending treatment ACE-inhibitors or ARB's in people with persistently raised albuminuria but no such recommendation exists in the current NICE guidelines. There also exists no specific recommendation in any of the guidelines for the monitoring of renal function after treatment has started, for either the initiation phase, control phase or cessation phase.

The work presented in this chapter has been developed from a model developed specifically for a Health Technology Assessment (HTA) project on monitoring for diabetic nephropathy in people with diabetes. I was asked to build a suitable statistical model for the progression and variation of albumin-creatinine over time so that different intervals of pre-treatment monitoring could be compared and built into a health economic model. This model would comprise of two parts; a model for the underlying progression of disease that could be extrapolated through a lifetime of diabetes and a model for the monitoring test strategy itself. As with many evaluations of screening and monitoring effectiveness that use simulation models the results and recommendations depend on the usefulness of the models for both the underlying disease and the monitoring or screening test. Poorly fitting models will often lead to poor predictions and produce misleading evidence. This chapter details the exploratory work, and the development of three different models for the underlying development of nephropathy and the test mechanism and their impact on evaluations of

different monitoring strategies. The work from this chapter has been published in the journal of Diabetes Research and Clinical Practice [62] (see appendix D, page 235 for all publications from this thesis)

2.2 Data

Modelling or derivation data

To model the progression and variation of albumin-creatinine ratio (ACR) over time I used data from the Oxford Regional Prospective Study (ORPS). ORPS was an incident cohort study of children and young adults with type 1 diabetes. Patients with newly diagnosed type 1 diabetes were recruited into the study between 1986 and 1996. Subjects were eligible for the study if they were diagnosed with type 1 diabetes before the age of 16 years and living within or moving into the Oxfordshire region within a year. Patients with diabetes secondary to another disease were excluded.

Subjects were assessed at the end of the first year from diagnosis and annually thereafter. Annual assessments included measurement of three consecutive early morning urine specimens for determination of albumin-creatinine ratio (ACR), blood pressure, HbA1c, lipids, height and weight and continued for up to twenty years. Assessments performed at 5 years from diagnosis included estimation of albumin excretion rate (AER) from three timed overnight urine samples and glomerular filtration rate from plasma clearance of inulin. Albumin was measured by an enzyme-linked immunosorbent assay, and creatinine using the modified Jaffe's method [63].

Subjects were followed up for a median (IQR) 9.0 $(5.3, 12.3)$ years, giving a total of 4098 person-years of annual assessments. 179 of the patients were diagnosed with type 1 diabetes between the period of 1985 and 1990 and 356 were diagnosed between 1990 and 1995.

Using a definition of microalbuminuria of $ACR > 3.5$ mg/mmol in males and $ACR > 4.0$ mg/mmol in females in 2/3 consecutive measurements., Schultz *et al.* [64] reported that 63 children (12.8%) developed microalbuminuria over a median follow-up of five years. In those which it could be ascertained, microalbuminuria regressed in $(21/43 = 49\%)$ as it was only evident in a single year and in 22 of 43 (51%) it persisted for two or more consecutive years. Overt

proteinuria developed in 8 people (1.6%) and persisted for at least 2 of 3 years, but regressed in one case [64].

Validation data

I obtained an additional data set to externally validate the models for albumin-creatinine ratio. The clinical trials unit based in the Nuffield Department of Primary Care Health Sciences, University of Oxford led by Professor Andrew Farmer conducted a 6-week cohort with participants from four general practices in Oxfordshire. Four first-pass urine samples were collected from 98 patients with type 2 diabetes at two weekly intervals. Urine samples were analysed at a central laboratory ratio using standard quality control procedures. Albumin-creatinine ratio was calculated from the urine albumin and creatinine measurements. This was study primarily designed to estimate the diagnostic accuracy of urine dipstick testing in screening for microalbuminuria and was published in 2013 [65]. Hereafter, this study will be referred to as the ACR cohort study.

2.3 Exploratory analysis

In ORPS, annual assessments (up to three repeat ACR tests) were scheduled but in practice intervals were irregular. The median interval between assessments was 1 year with inter-quartile range (IQR) of 0.7 to 1.3, and 17% of intervals were less than 6 months. To explore the data and to create a reference model, I used the direct method [5]. This method requires measurements to be taken at regular intervals (e.g. annually) so that the marginal statistics can be calculated. To create equal assessment times, I took the mean of all the ACR tests that fell within a given year. Baseline was defined as the point at which diabetes was diagnosed and subsequent years as years since diagnosis. When there were two assessments (of three tests each) that happened to fall within a year, all measurements were aggregated to an average. Reshaping the data in this way gave 3835 yearly mean ACR's from 535 people, with a maximum of 24 years follow-up. Although it should be noted that there were less than a 100 people contributing ACR measurements for more than 18 years. Figure 2.2 shows both the observed mean and variance for the aggregated ACR data by years since diagnosis. The trend in average albumin-creatinine ratio rises sharply and is indicative of a exponential increase over time. The variance also increases greatly in the same time period (but note the different scale in the variance graphic). I took the average change to be the average change in mean ACR over the course of the study;

$$E(\Delta_t) = E(Y_t) - E(Y_{t-1}) \quad t = 2, \ldots, n \tag{2.1}$$

I first estimated average change in mean ACR using all 20 years of data. However, this seemed to be excessively high (1.16 mg/mmol per year) and did not correspond well to the observed change over time. The estimated average change based on the first twelve years was much smaller 0.39 mg/mmol and fitted the observed change better. Figure 2.3 shows the same mean ACR plotted over time with the direct method estimated increase in mean ACR overlaid.

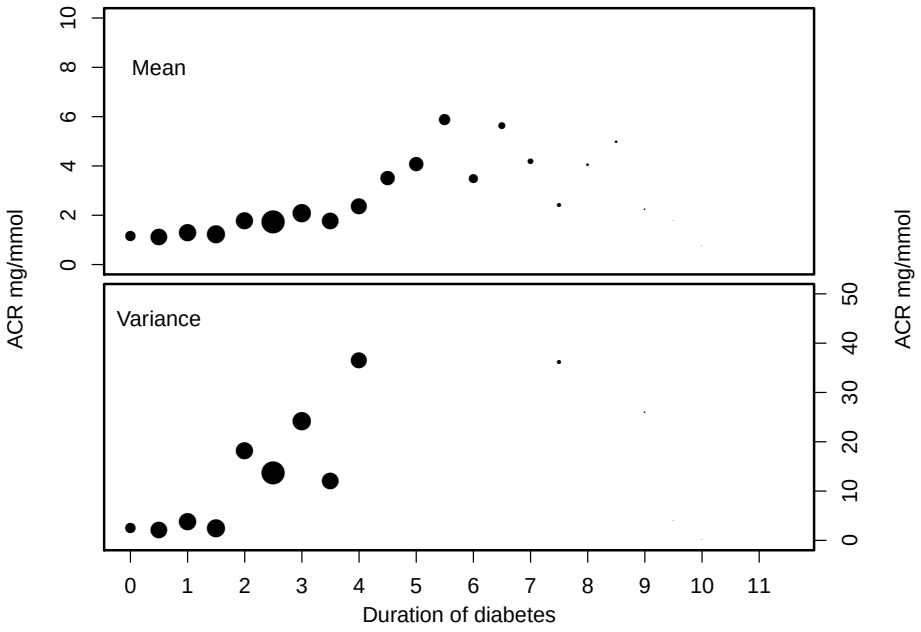


Figure 2.2: *Dotplots showing mean ACR (top panel) and (bottom panel) variance of ACR by duration of diabetes. Dot size is proportional to the number of data points at each time point. Note the change in scale of y axis*

The poor fit is due to the influence of the large positive changes from years 12 onwards (see individual change values in figure 2.3). This estimate is an improvement over the first estimates but fails to capture the curvature of the observed change in average ACR. The observed variance also rises rapidly in an exponential fashion and so I decided to try a natural logarithm transformation of ACR. The natural logarithm (log) has been used many times when analysing albumin-creatinine ratio, see for example [64]. By applying a log transformation, the log of albumin-creatinine ratio represents the difference between the level of albumin and the level of creatinine. An interesting property of the log-transformation is that for inferential purposes it does not matter which of the two variables divides the other as;

$$log(v_1/v_2) = -log(v_2/v_1)$$

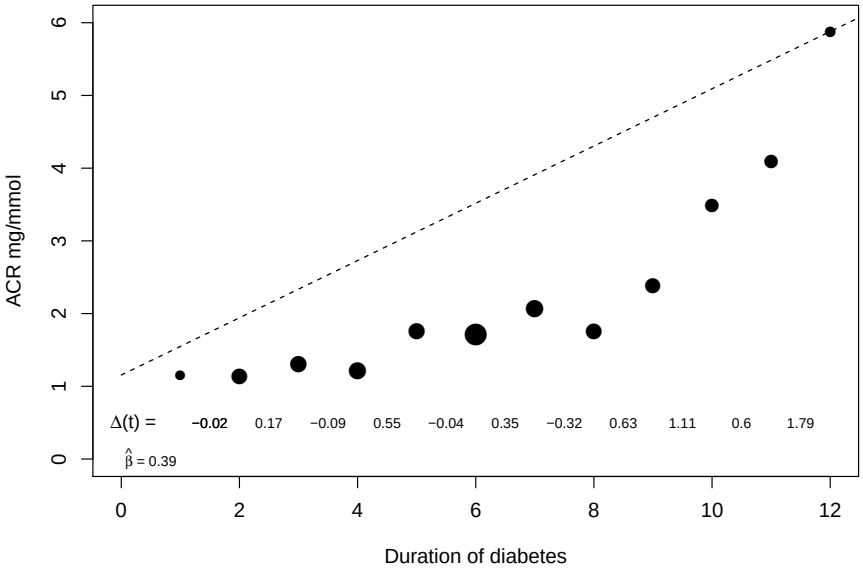


Figure 2.3: *Plot of predicted mean ACR overlaid on observed aggregated mean.Dot size is proportional to the number of data points at each time point.*

An important feature of the log transformation is that the back-transformed (antilog) averages become geometric means, which are always less than the arithmetic mean and less susceptible to one or a few aberrant larger values [66]. Additive effects become multiplicative effects and so a linear trend on the log-scale becomes an exponential trend when back-transformed. Moreover, when a standard deviation is back-transformed it is not a standard deviation on the original scale but a related quantity sometimes referred to as a geometric standard deviation [67]. Figure 2.4 shows the marginal means and variance of log ACR from the ORPS data. The log transformation appears to have linearised the increases in the mean and variance of ACR. Hereafter, all of the analysis will be using the log scale ACR.

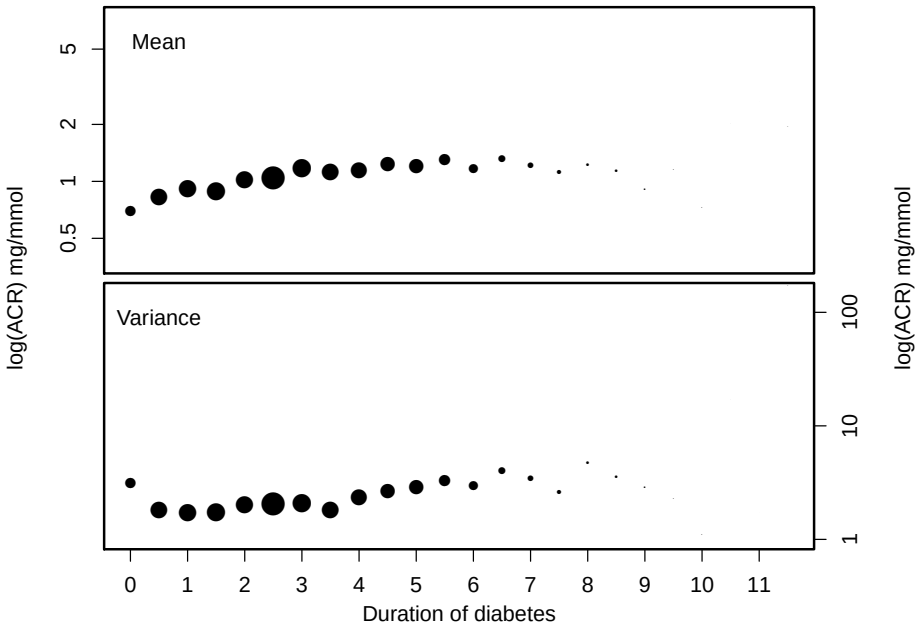


Figure 2.4: *Dotplots showing mean log ACR (top panel) and (bottom panel) variance of log ACR by duration of diabetes. Dot size is proportional to the number of data points at each time point. Note the change in scale of y axis*

2.4 Methods

I set out initially to create monitoring models using three approaches, two of which were existing methods; the *direct* method and the *increasing variance* model and a third model which is a special case of a more general method of fitting longitudinal data.

Direct method

Introduction

Glasziou [5] *et al* described two methods for estimating the proportions of false and true positives, false and true negatives over time from repeated measures data. The first was referred to as the *direct method*, whilst the second was referred to as the '*linear model with increasing variance*. The direct method can be thought of as an extension of what is referred to in longitudinal data modelling as a marginal approach [68]. In the marginal linear model, the average of the data is calculated for each time point t as if it is a series of cross-sectional studies. The average can be made to depend on factors of interest if required by splitting the data into levels of the factor. Obviously, a large data set is required to adjust for multiple factors or factors with many levels. One drawback with this method is that since repeated values are not likely to be independent, the marginal analysis must also include assumptions about the form of the correlation [68].

In the cholesterol model, the average or expected cholesterol $E(Y_t)$ was assumed to increase linearly as a function of time t;

$$E(Y_t) = \beta t \tag{2.2}$$

with β representing the change in the average cholesterol level from baseline or time $t = 0$. The increase over time (β) was estimated by first, calculating the sample average at each time point;

$$E(Y_t) = \frac{\sum_{i=1}^{n} y_{it}}{m} \tag{2.3}$$

then calculating first differences Δ of the averages, i.e.

$$E(\Delta_t) = E(Y_t) - E(Y_{t-1}) \;\; t = 2, \ldots, n \tag{2.4}$$

and then taking the average of these n differences;

$$\hat{\beta} = \frac{\sum_{t=2}^{t}(E(\Delta_t))}{n-1} \tag{2.5}$$

The method of calculating the variance proposed by Glasziou *et al* [5] is novel.

In the cholesterol paper, the authors considered that the fluctuations in cholesterol during the control phase comprise 3 components: 1) the population baseline variance σ_T^2, 2) the variation in trend σ_{1t}^2 and 3) short-term variability, which is a combination of analytic variability and week-to-week biological variation around a stable average σ_W^2. As these quantities are assumed independent of each, the total variance of cholesterol Y at time t is;

$$V_t = \sigma_T^2 + \sigma_{1t}^2 + \sigma_W^2$$

and the variance of the change from baseline is;

$$V(\Delta_t) = \sigma_{1t}^2 + 2\sigma_W^2$$

which is the twice the short-term variation (σ_W^2), plus the long-term variation (σ_{1t}^2). The long term variation (σ_{1t}^2) can then be calculated by subtracting twice the short-term variance (σ_W^2) from the variance of the change ($V(\Delta_t)$) and the short-term variance. This is called the "variogram" method, more details are given in appendix C.

Estimation

The *direct method* consisted of two steps; estimation of the average baseline and increase in albumin-creatinine ratio and estimation of the two components of variation (short-term σ_W^2 and long-term σ_{1t}^2) using the variogram estimator. As I had reason to believe that the prognosis of diabetic nephropathy differed by gender, I considered separate estimates of progression and variation for men

and women. In addition, as the data became sparse after 12 years of duration, the variogram was estimated using data from years 1 to 12 as was the average trend.

Diagnostics

Model diagnostics for the direct method included plots of the estimated total variance against the observed and estimated average ACR (for both men and women) against the observed.

Prediction

The direct method estimates were then used to calculate the proportion of tests that were true and false positive or true and false negative and the ratio of false positives to true positives for the case where an initial test was negative for micro-albuminuria and the interval between the next measurement was $1, 2, 3, \ldots, 10$ years.

Increasing variance model

The second model in [5] was a linear random effects model for the progression and variation of cholesterol over time. This was referred to as the model with *increasing* variance in the appendix of the paper. The term increasing is used because the model fixes the correlation between the random intercept and slope terms at zero, forcing the variance to increase over time. A random effects model is distinct from the *direct* or *marginal* method in that it models both the individual response and the marginal response simultaneously. These models also explicitly define a correlation structure which is not the case with the direct method.

The increasing variance and random effects models in general can be thought of a series of simple linear regression models for each person in turn, with the intercept and slope terms made to depend on each other. The term "random" from random effects is used because it is assumed that the *units* (people in

a monitoring context) are a sample drawn randomly (or can be considered as random for all intents and purpose) from some larger population.

Definition 5 ***Random effects*** *- A quantitative measure of how each person differs systematically from the average, representing the sum of all unmeasured characteristics that explain why people are inherently different from each other with respect to the outcome*

Throughout this thesis I have often assumed that the random effects are distributed normally and this is justified by the central limit theorem. The central limit theorem says that the sum of many independent variables (that constitute the random effects) will tend to a normal distribution as long as the individual variances are small compared to the total [69]. Unlike the direct method, the timing of visits does not need to be regular and arranged into a discrete time ordering. It is also easy to adjust for covariates without splitting the data. These models do make certain assumptions about the distribution of the random effects and residual variance which is not the case with the direct method.

Covariates Effects

To test the importance of gender on the progression of log ACR I first created a reference model (model 1) as a function of the duration of diabetes (centred) t_c without any adjustment for gender;

$$log(\text{ACR})_{it} = \alpha_i + \beta_i t_c + \epsilon_{it} \tag{2.6}$$

Equation 2.6 says that log (ACR) for the ith person, at time t is a linear function of a baseline or intercept α_i and a gradient or slope parameter β_i. A second model, adjusts the baseline for each person according to their sex S_i implying that sex affects the baseline level, and not the rate of progression;

$$log(\text{ACR})_{it} = \alpha_i + \beta_i t_c + \gamma S + \epsilon_{it} \tag{2.7}$$

A third model adjusts both the absolute level and the rate of progression by gender by incorporating an interaction term δ;

$$log(\text{ACR})_{it} = \alpha_i + \beta_i t_c + \gamma S + \delta t S + \epsilon_{it} \tag{2.8}$$

All three models have the same model for the population level or second level of the model. The population mean (μ_α) of the intercept terms α_i and mean (μ_β) of the slope terms β_i are assumed normally distributed;

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim \left[\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \phi = 0 \\ \phi = 0 & \sigma^2_\beta \end{pmatrix} \right] \tag{2.9}$$

and independent. In all three of these "random effects" models, the measurement errors or short-term variation are assumed normal with mean 0 and constant variance σ^2_W.

Flexible variance model

To assess the effect of constraining the variance to increase (as is the case with the increasing variance model), I fitted a model in which the covariance is estimated from the data. This changes the 2nd level of the model to

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim \left[\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma^2_\beta \end{pmatrix} \right] \tag{2.10}$$

This *flexible* variance model will be estimated using MLWin, a multi-level modelling software created by the centre for multilevel modelling at the University of Bristol [70]. I found this software allowed the user to change the covariance structure in contrast to *lme*[71] in R [72].

Model selection

From the four candidate models, I set out to select the best fitting model to use for the individual patient simulation model. To do this, I used the model deviance (a quantitative measure of the difference between the fitted model and observed data). Nested models (subsets of each other) with lower deviances indicate a better fit. However, as even if we add explanatory variables

that have no association with the outcome, the deviance will decrease, so we compare the reduction in deviance we observe against a reference distribution to produce a p value. Models with significantly lower deviance ($p < 0.05$) were considered better fitting models.

Diagnostics

Diagnostics or statistical validation for random effects model consist of checking two basic assumptions (1) within-person errors (short-term variation) are independent and identically distributed with mean zero and constant variance and are independent of the random effects, (2) The random effects are normally distributed, with mean zero and covariance matrix that do not depend on the group level covariates [73]. Histograms of the random effects were assessed and measurement error residuals checked for homogeneity across the predicted outcome and across predictors (e.g gender). The predicted variance was compared to the observed variance.

A more rigorous examination of the data involves comparing the predicted proportions of tests that are positive or negative over time. Although we cannot observe these quantities directly, the model must at least be able to accurately predict the total number of positive and negative tests. Therefore, this test of the model will be useful to rule out models (those that fail to predict the proportion of positive and negative tests) rather than definitely rule in because even if the model predicts the positive and negative tests it may fail completely on what proportion of these are false.

My preferred method to see assess model fit comes from Gelman [69]. This involves replicating the outcome (log ACR) in the original data by simulating each observation conditioning on the relevant covariates and the model parameters. For example, if a person in the data has four observations (see table 2.2) the method proceeds by simulating an intercept α_i and slope β_i term, and

ID	Gender	Duration (t)	Observed ACR (Y_t)	Fitted inter-cept	Fitted slope	Simulated ACR (Y_{sim})
1	Male	0	-0.5	α_1	β_1	$\alpha_1 + \beta_1 * 0 + \epsilon_1$
1	.	1.2	-1	.	.	$\alpha_1 + \beta_1 * 1.2 + \epsilon_2$
1	.	4.5	1.6	.	.	$\alpha_1 + \beta_1 * 4.5 + \epsilon_3$
1	.	6	2.9	.	.	$\alpha_1 + \beta_1 * 6 + \epsilon_4$

Table 2.2: *Theoretical observed and simulated ACR data for one person*

measurement error ϵ_t from their respective distribution;

$$\begin{aligned} \alpha_{new} &\sim N(\mu_\alpha, \sigma^2_\alpha) \\ \beta_{new} &\sim N(\mu_\beta, \sigma^2_\beta) \\ \epsilon_{new} &\sim N(0, \sigma^2_W) \end{aligned}$$

and then the simulated outcome data is calculated by doing;

$$Y_{sim} = \alpha_{new} + \beta_{new} * t_c + \epsilon_{new}$$

Any number of covariates can be included in the recreated outcome as long as it has been observed in the data. This technique can be used to compare any type of actual and simulated summary statistics. The most useful for a monitoring model is the observed and predicted number of tests that would be positive. To remain faithful to actual clinical practice this entails calculating the number of assessments (batches of three tests) that would be graded positive for persistent micro-albuminuria.

T model

Exploratory work on the ACR data suggested that the measurement error residuals (difference between the fitted value and the observed) although symmetric have substantially more outliers and more extreme values than the normal distribution predicts. The t distribution with low degrees of freedom has longer tails than the normal and can be used for modelling data with extreme values. The t distribution is used in place of the standard normal distribution for comparisons of two means from small samples (the *t test*) for this very reason. The t distribution has three parameters; the centre μ, the scale σ and a degrees

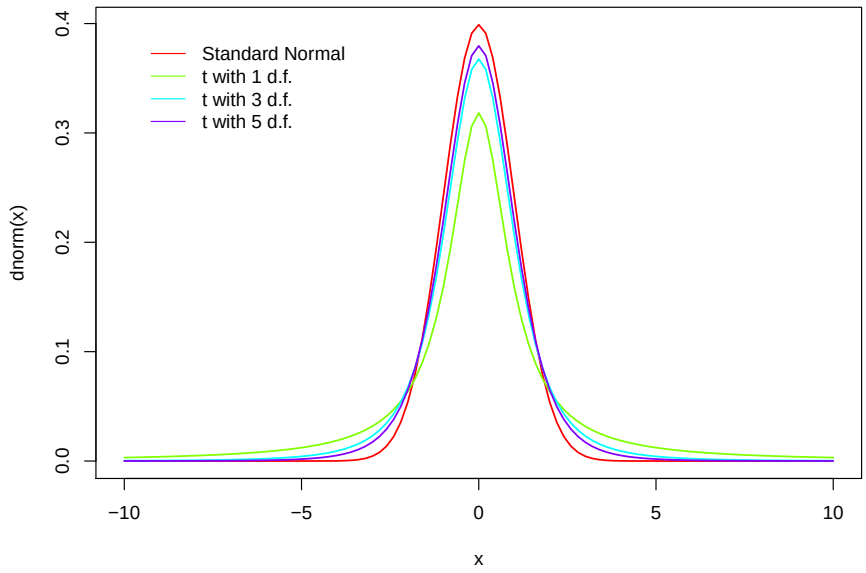


Figure 2.5: *Density plots of the standard normal and t distribution with varying degrees of freedom (d.f.)*

of freedom parameter ν that determines the shape of the distribution. When there is sufficient data it is generally appropriate to estimate the degrees of freedom, although when the t is intended to model extreme values it can be fixed at a small value [69]. As the degrees of freedom increases, the t distribution closely resembles the normal distribution (see figure 2.5). The t distribution can also be interpreted as a mixture of normal distributions with a common mean and variances distributed as a scaled inverse χ^2. For example, the model $y_i \sim t_\nu(\mu, \sigma^2)$ is equivalent to;

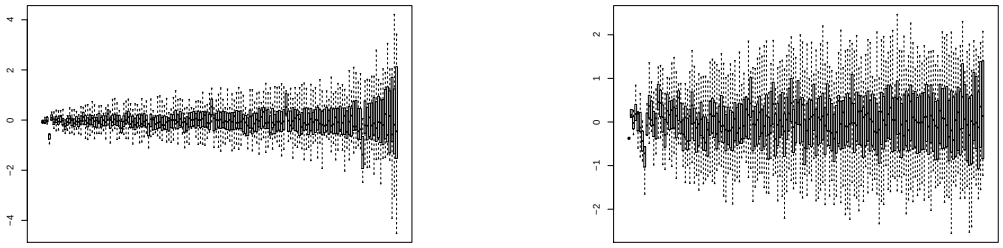
$$\begin{aligned} y_i &\sim N(\mu, V_i) \\ V_i &\sim Inv.\chi^2(\nu, \sigma^2) \end{aligned} \tag{2.11}$$

An inverse-χ^2 distribution for the variances is equivalent to a gamma distribution for the precisions (τ_i)

$$V_i \sim Inv - \chi^2(\nu, \sigma^2) \equiv \tau_i \sim gamma(\nu/2, \nu/2\sigma^2)$$

A T model therefore, can be interpreted as a model in which each person has their own measurement error term σ^2_ω. Evidence for this can be seen by ordering

boxplots of within-person residuals by size of the inter-quartile range. Figure 2.6 shows the observed residuals from the flexible variance model ordered by size of inter-quartile range and the same plot but for simulated normal residuals (which assumes everyone has the same standard deviation) to illustrate the difference.



(a) *observed residuals* **(b)** *simulated residuals*

Figure 2.6: *Caterpillar plots of residuals ordered by size of inter-quartile range for (a) residuals and (b)simulated residuals with common measurement error* σ^2_ω

Modelling the progression of ACR

The T model for log(ACR) can be written as in a similar way to the model with increasing and flexible variance. The first level of the model is;

$$log(\text{ACR})_{it} = \alpha_i + \beta_i t + \epsilon_{it} \tag{2.12}$$

where, as with the flexible variance model, the correlation ρ is estimated.

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim \left[\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \rho_{\alpha,\beta} \\ \rho_{\alpha,\beta} & \sigma^2_\beta \end{pmatrix} \right] \tag{2.13}$$

but instead of the measurement errors ϵ_{it} being distributed normally with constant variance; $\epsilon_{it} \sim N(0, \sigma^2_W)$, the T model fits a separate variance parameter for each individual and so;

$$\epsilon_{it} \sim N(0, \sigma^2_{Wi})$$

and the individual precision terms (1/variance) are given a gamma distribution, with parameters a, the shape and b the scale;

$$1/\sigma^2_{Wi} = \tau^2_{Wi} \sim \quad gamma(a, b) \tag{2.14}$$

Estimation using Monte Carlo Markov Chain (MCMC)

An introduction to estimating these models with MCMC techniques is given in Appendix A.

Because it is not possible to estimate the T model using standard random effects software e.g. *lme* in R, *xtmixed* in STATA or MLWIN, I used MCMC software, first WinBUGS [74] and then using JAGS [75] for estimation. The T model is an extension of the basic hierarchical linear model described in Appendix A. Following the notation of the appendix, the 1st level loops over the entire data from 1 = to N and is coded as;

```
y[i] ~ dnorm(mu[i], tau.omega[id[i]])
       mu[i] <- beta[id[i],1] + beta[id[i],2] * t[i]
```

where the intercept terms α_i and slope β_i notation from the previous models are now coupled together in a "vector" called beta[]. The id notation is used to signify that the intercept and slope parameters are held constant for each value of id and hence each person in the data. The second level of the model is wrapped in a second loop indexed from $j = 1$ up to J where J is equal to the number of people in the data,

```
       beta[j, 1:2] ~ dmnorm(mu.b[1:2], Tau.B[, ])
       tau.omega[j] ~ dgamma(a, b)
```

The precision (1/variance) parameter tau.omega is included in the loop as it also varies per individual. The precision matrix Tau.B is inverted and the individual standard deviations, sigma.a and sigma.b are given uniform priors;

```
Tau[1:2,1:2] <- inverse(Sigma[,])
 Sigma[1,1] <- pow(sig.a,2)
 Sigma[2,2] <- pow(sig.b,2)
 sigma.a ~ dunif(0,100)
 sigma.b ~ dunif(0,100)
```

whilst the correlation parameter ρ is given a uniform prior on the range $-1, 1$. The hyperparameters of the individual variances are transformed to scale *sigma* and degrees of freedom ν parameters of a t distribution by doing;

```
a <- nu/2
b <- (nu/2) * pow(sigma, 2)
nu <- 1/nu.inverse
```

with the scale parameter given a uniform prior distribution and the inverse of the degrees of freedom parameter a prior which ranges from 0 to 1;

```
sigma ~ dunif(0, 100)
nu.inverse ~ dunif(0, 1)
```

The model was estimated with multiple chains and sufficient iterations to obtain convergence. Convergence was assessed by equating the effective sample size to the actual sample size and the use of the Gelman Rubin statistic (Rhat). In addition, convergence was also assessed by inspection of density plots and traceplots.

Diagnostic checks of the model are the same for this model as for the random effects models. Distributional assumptions are checked using histograms of the fitted random effects and variance terms. The model was used to predict the variances over time. Simulated outcome data was generated under the model estimate and compared graphically with the observed data.

Evaluation of monitoring prior to initiation of ACE-i therapy

I developed a patient-level simulation model to evaluate how the different models for disease progression affect outcomes of different monitoring strategies. The individual patient simulations were then used to capture a number of relevant outcomes relating to monitoring treatment, operational costs, overdiagnosis and overtreatment.

I simulated a hypothetical patient cohort with similar characteristics to those in ORPS. In the simulation people entered the renal monitoring programme when they reached age 12 if they had already been diagnosed with type 1 diabetes and immediately after diagnosis if this occurred after their twelfth birthday. This follows the current NICE guidelines [21]. Persistence was diagnosed by two abnormal test results from a possible 3 tests. ACR was modelled by deriving for each person, 1) a baseline level and 2) annual change value by simulating from distributions with parameters estimated by the models. The course of true ACR over a lifetime was then extrapolated for each individual. The observed ACR (from which the assessments results are based on) is derived by adding random noise to the true ACR level. The cohort at the beginning of the simulation comprises of patients of age twelve and older based on the proportions found in the ORPS cohort. Threshold defining positive tests for microalbuminuria were 2.5 mg/mmol for males and 3.5 mg/mmol for females. If a patient has a positive diagnosis, they are no longer monitored, but remain accounted for in the simulation. This allows those whose positive diagnosis is premature to progress from a false positive state to a true positive state. If the diagnosis is negative, the patient continues to be monitored. To approximate life expectancy, I used a Gompertz function with shape and scale parameters configured to give an average length of life equal to those used in the full health economic analysis in [76]. The annual hazard with regard to age can be written as;

$$h(x) = 0.005e^{0.023x}$$

where x is the age of the patient.

The different monitoring strategies were evaluated using the following outcomes measures. The average number of tests/assays that a person undergoes prior to a positive diagnosis and initiation with ACE-inhibitor therapy, which is the sum of all the test undergone over a lifetime, up to and including a diagnosis of microalbuminuria if one occurs. The percentage of people who test positive

at any time and the percentage to these who at the time are false positive (true ACR < threshold). The average advance time is derived by calculating for each person the difference between the time of the first false positive test (if one occurred) and the time until their true ACR crossed the threshold. The percentage of false negatives is taken as the number of people who have false negative assessments to the total monitored. The total number of false negative assessments for each person is also calculated. The average delay time is calculated as the difference between the time at which the true ACR crossed the threshold and the time at which it was detected by monitoring. Finally, the percentage of people overdiagnosed by monitoring is calculated by determining the percentage of people who received a false positive test but whose true albumin creatinine ratio would never have crossed the threshold before the time of death. The individual simulation models are repeated for each of the three random effect models; the increasing variance model, the flexible variance model and the T model. The direct method cannot be used because it does not define individual progression.

Assumptions

I made the following assumptions in the simulation model. Firstly, there are no ACR assays which fail to give a quantifiable level of albumin excretion. Although the short-term variance parameters in the model capture assay variability it does not include variability in test results that may arise from different laboratories with different quality standards and procedures. I further assume that all the individuals participating in the monitoring programme attend the screening appointment at intervals defined by the monitoring programme and do not miss appointments or drop out of the programme. Lastly, I assume that there are no *intervals cases* ; people with clinical presentations of proteinuria/nephropathy that leads to a diagnosis either in primary or secondary care.

Clinical Validation

In this thesis, I have attempted to validate the statistical models using the definitions described by Altman and Royston [77]. They considered that models could be validated in two ways, statistically and clinically;

- A *statistically validated model* is one which passes all appropriate statistical checks, including goodness-of-fit on the original data set and unbiased prediction on a new data set.
- A *clinically validated model* is one which performs satisfactorily on a new data set according to context-dependent statistical criteria laid down for it.

Clinical validation (using the definition above) of these models is complicated (as it must be with any latent variable models) because the context-dependent outcomes (false positives etc) we are trying to predict cannot be observed directly or easily. This makes these models harder to validate and criticise. The following three methods represent my efforts to overcome this problem and find ways to provide independent validation of the monitoring models.

Surrogate method

This approach to validation was used originally in a paper by Gardner and Heady [31]. This method requires having a set of repeat measurements taken closely in time, some to form a surrogate *true* value, and at least another to act as the *observed* test. The observed and surrogate values can then be assessed either using quantiles of the data as in Gardner and Heady's paper or by two-by-two tables using an appropriate threshold. This method has limitations. It depends on having short-term repeat measures, which are unusual in trial data and highly unlikely in routine collected data. I managed to procure such a dataset, but it is relatively small in size (n = 100).

To derive the surrogate true level from the validation data set I averaged the first three measurements of ACR for each person. To assess sensitivity to the choice of the first three measurements I compared results obtained by choosing all other possible combinations of tests. The surrogate true level was dichotomised into a 0 if the measurement was less than the relevant threshold and 1 otherwise. The fourth measurement of ACR was also dichotomised in this way. Misclassification probabilities were then estimated by cross-tabulating the surrogate and the observed measure.

Regression to the mean method

If the between-person and within-person variation in the model reflects the truth, then the degree of regression to the mean and misclassification can be predicted. For example, consider the case of two measurements (y_{i1}, y_{i2} taken on n people one year apart. Each person has their own true constant average μ which varies around a population mean θ with variance σ^2 and the measurements taken are on average correct but vary by an amount δ^2. With this structure we would expect the average of the measurements at time 1 and 2; $E(Y_1)$ and $E(Y_2)$ to be the same, each close to the true value θ and the observed variance $Var(Y_1), Var(Y_2)$ also to be constant and close to the true variance $\sigma^2 + \delta^2$. The correct model for this data is a random effects model or one-way analysis of variance model [78]. A much simpler marginal model (model 2) estimates the mean of the data using the sample mean of the data Y_t and sample variance $Var(Y_t)$. Both models will correctly predict the marginal means $E(Y_t)$ and variance $Var(Y_t)$ for a new set of measurements taken on the same population up to a margin of error. However, only the correct model will predict the mean of the second observation conditional on the value of the first measurement. The first model has an inherent degree of regression to the mean which depends on the ratio of the between person and within-person variation. The second model does not, and predictions of the conditional mean and variance will be biased. I set out to test each of the models ability to predict

the observed regression to the mean as a method of validating the monitoring models.

Prevalence methods

The final method used for clinical validation is the simplest and consists of finding estimates of the prevalence from published cohort studies to form a prevalence curve over time. The model predicted prevalence can then be compared to the one compiled from the literature. A limitation of this method is that it can only check how well the model predicts rates of positive and negative tests and not how many of the positive tests are false.

2.5 Model Estimation

Direct Method

To establish whether the direct method estimates of progression and variation should be adjusted for gender, I split the data set into males and females and analysed them separately. There were 293 males and 242 females in the ORPS cohort and so splitting the data gives approximately equal size data sets (and therefore equal precision) for which to estimate the parameters. The variogram and average increase in log ACR (β) was estimated for males and females. The variogram was also used to estimate the short-term variance σ^2_W and the long-term variance σ^2_{1T} (see figure 2.7).

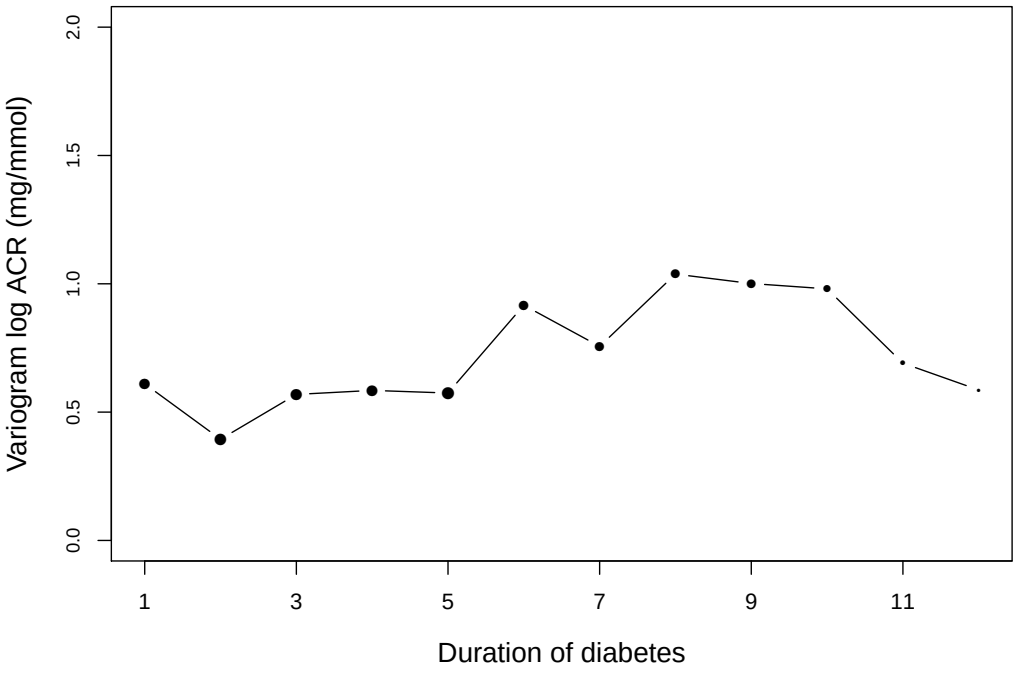


Figure 2.7: *Observed variogram of log ACR (mg/mmol)*

Parameter estimates

Table 2.3 shows the direct method parameter estimates for males and females. Table 2.3 shows that the average change in females is double that of the males

Parameter	Males	Females
Average log(ACR) mg/mmol at baseline	-0.345	- 0.389
Average increase of log(ACR) mg/mmol per year	0.031	0.076
Short-term variation in log(ACR)	0.57	0.51
Long-term variation 5 years	-0.02	0.129
Long-term variation 10 years	1.058	0.554

Table 2.3: *Direct method estimates for males and females separately*

by this method (0.031 vs 0.076). The short-term variation estimates are similar but the long-term variance estimates are irregular (sometimes negative) and may benefit from pooling the data together to improve estimation.

Diagnostics

The estimates of short and long-term variation and the variance of change since baseline for the pooled data is shown in table 2.4 Table 2.4 illustrates a

	Visit							
Variation	1	2	3	4	5	6	7	8
Variance of change $V(Y_t - Y_0)$	1.22	0.79	1.14	1.17	1.15	1.83	1.51	2.08
Short-term (σ_w^2)	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53
Long-term (σ_{1T}^2)	0.16	-0.27	0.08	0.11	0.09	0.77	0.45	1.02

Table 2.4: *Direct method estimates of the total variance, variance of change (2*variogram) and long-term variance component estimates for the first 8 years*

potential problem with the direct method. At the second time point, the long-term component (σ_{1T}^2) is negative and times 3,4 and 5 are less than the first time point. It is possible that these estimates are consistent with underlying

constant variance allowing for sampling variability. However, we expect the variance to be an increasing function of time and not for it to decrease. To assess sensitivity of these estimates to the linearity assumption in the regression model, I calculated a non-parametric regression (lowess) estimate. Although these models cannot formally extrapolate beyond the range of the data, informal inspection of the model fit suggested a value of around 0.5 to be reasonable for the variance of change at time 0. Substituting the lowess estimates for the short-term variation (σ_w^2) did not make substantive differences and the negative problem remained.

Prediction

I used the average change in log ACR for males and females and the observed average at baseline to predict the average ACR from years 1 to 12. I also calculated the observed variance for both males and females and predicted the total expected variance under the model by summing the estimated short-term variance and the long-term component. Figure 2.8 and 2.9 shows the model predictions against the observed means and variances.

Predictive Validation

Due to the nature of the direct method it is not possible to do this method of internal validation.

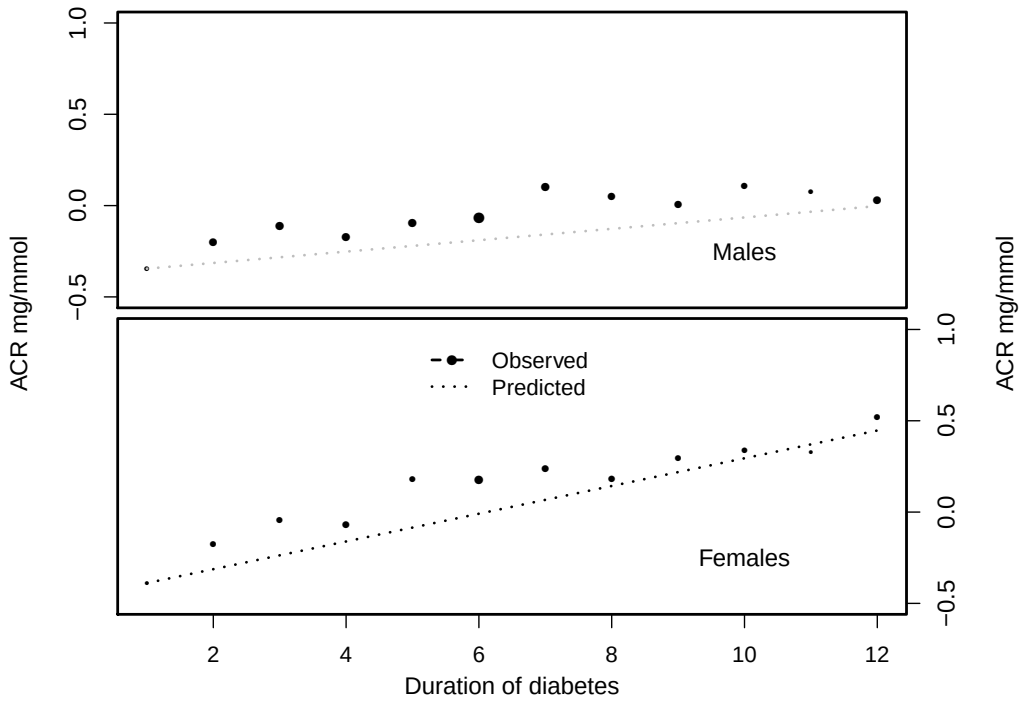


Figure 2.8: *Direct method estimates of average log ACR stratified by gender with observed average overlaid*

Increasing variance model

For the increasing variance model, the repeated measures were treated as separate measurements and not averaged within assessment or year. This gives 13004 measurements on 535 people. It is common practice in multi-level modelling to centre explanatory variables around their mean [79], often to ease computation of the parameter estimates. However, initial attempts to estimate the models using a centred version of duration of diabetes for the time variable failed to converge and hereafter all models use the time variable on its original scale.

The model unadjusted for gender (model 1) has deviance 30882.174 and the model with intercepts adjusted for gender (model 2) has deviance 30879.951, a difference of 2.223. This value can be compared with a chi-square distribution with 1 degree of freedom (p = 0.14), which says we would expect to see this

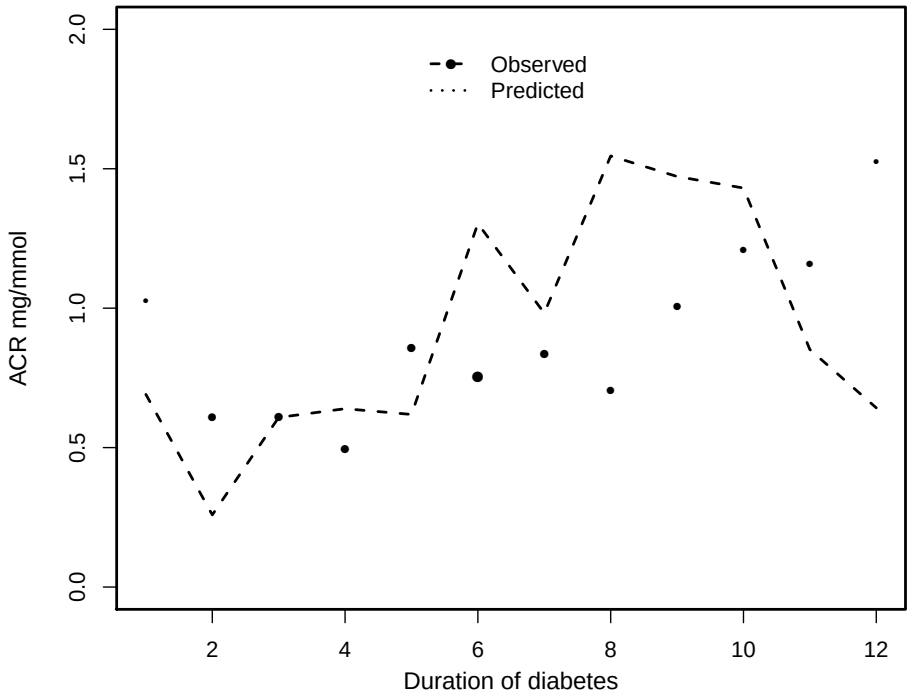


Figure 2.9: *Direct method estimates of the variance of log ACR by duration of diabetes with observed variance overlaid*

reduction 1 in 7 times if the variable has no association. Model 3 adjusts the intercept and slope terms for gender and has deviance 30877.96, a difference of 1.991, which is also not a significant improvement over the model 2, but is over model 1 (p = 0.16). Comparison of the variance parameters $(\sigma^2_\alpha, \sigma^2_\beta, \sigma^2_W)$ across the three models shows that they are robust to the changes to the patient level model. I decided to proceed with model 1 as it is the most parsimonious model. The direct model estimates from the previous section are compared with the increasing variance parameter estimates in table 2.5 Unlike the direct method, the increasing variance model imposes a variance which is quadratic in time (duration of diabetes);

$$Var(Y) = \sigma^2_\alpha + \sigma^2_\beta t^2 + \sigma^2_W$$

but as can be seen in figure 2.10 the observed variance looks linear rather than quadratic. The variance also appears to decrease in the first three years and

Parameter	Direct method	Increasing variance model
Average baseline at diagnosis	-0.36	-0.22
Average increase in log ACR	0.052	0.047
Short-term variance of log(ACR)	0.53	0.52

Table 2.5: *Parameter estimates from the direct method and increasing variance model. Units of log ACR are mg/mmol.*

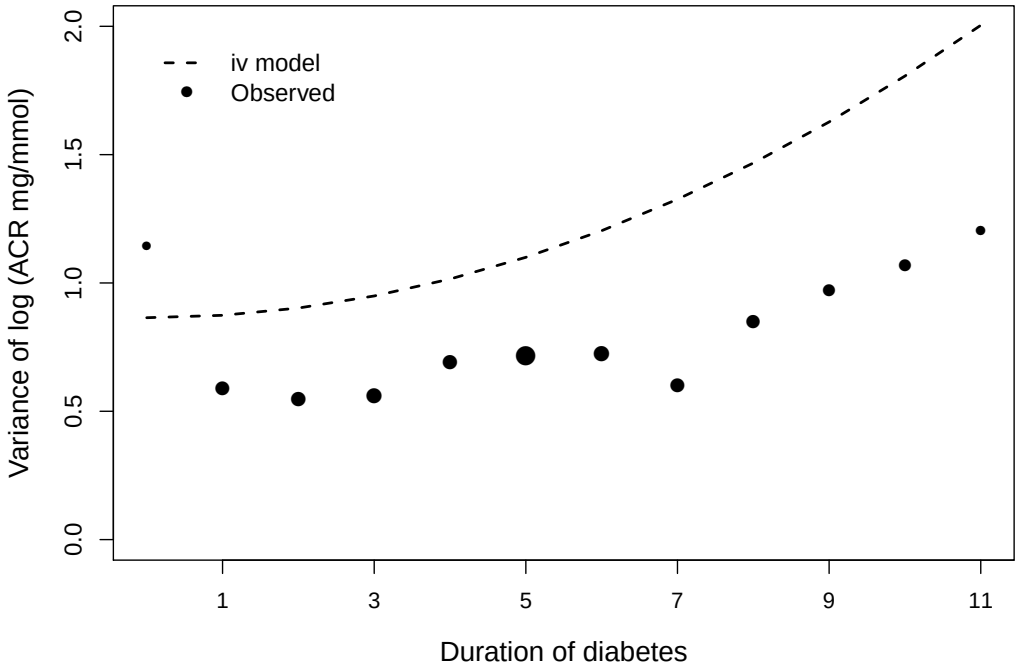


Figure 2.10: *Comparison of the observed variance (size of the black blobs is proportional to the number of data points) with the fitted increasing variance model variance*

this was also apparent from the direct method calculations of the long-term variance. Figure 2.10 compares the fitted variance from the increasing variance model against the observed variance estimated from the data.

Flexible Variance Model

Removing the restriction that the correlation between the intercept and slope terms is zero improves the model fit, reducing the deviance by 170 units, which

is far beyond what we would expect by chance. The correlation is estimated in this model as $\rho_{\alpha,\beta} = -0.64$ which implies that there is greater variation at both ends of the data and less in the middle (the average duration of diabetes). This pattern can be seen in a plot showing all the data points in a scatter point against duration of diabetes (see figure 2.11).

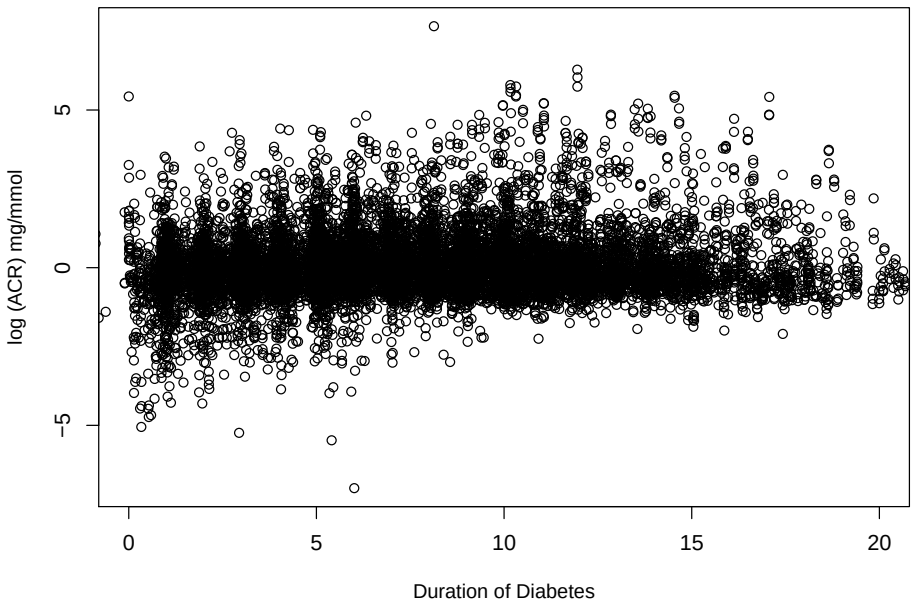


Figure 2.11: *Scatter plot of log ACR by duration of diabetes*

Parameter estimates

The estimates from both the increasing variance model and the flexible variance model are shown in table 2.6;

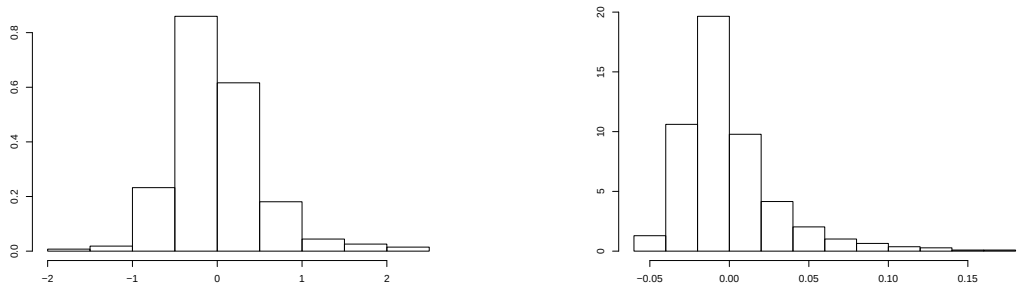
Diagnostics

The histogram plot of the random intercepts α_i's (figure 2.12(a)) is symmetric and plausible for the normal assumption. However, the distribution of the random slope β_i's estimates has a clear positive skew (long tail) and places doubt on the normality assumption. A multivariate test of normality, suggests

Parameter	Increasing variance model	Flexible variance model
Average baseline log(ACR) mg/mmol at diagnosis	-0.22	-0.22
Average increase in log ACR mg/mmol	0.047	0.047
Short-term variance of log(ACR) mg/mmol	0.52	0.52
Variance of baseline log(ACR)	0.395	0.541
Variance of gradients log(ACR)	0.010	0.013

Table 2.6: *Comparison of the parameter estimates for the increasing and flexible variance models*

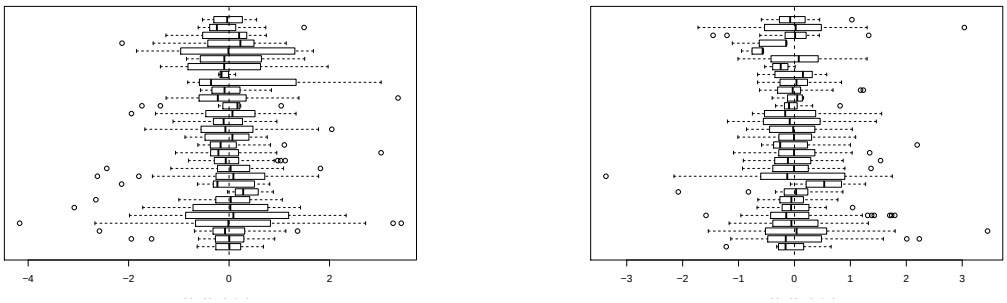
that this sample is highly unlikely if the true distribution is bivariate normal ($W = 0.8783$, p-value < 0.001).



(a) *Random intercepts* **(b)** *Random slopes*

Figure 2.12: *Histograms of the estimated random effects for the linear model (a) the intercepts α_i and (b) the slopes β_i*

We also need to check that the residuals are mean 0, have common variance and are independent of gender. Figure 2.13 shows the level 1 residuals (within-person errors) for thirty randomly selected males and females from the ORPS cohort. A boxplot of each set of residuals is drawn for each of the randomly selected thirty males and females. There does not appear to be a difference in the residuals across gender, and the assumption of mean 0 is plausible given that the boxplots are scatted either side of the dashed line in figure 2.13. There are however, a large number of outliers (circles) and large differences in the scale of the residuals across individuals, evident from the varying size of the boxes.



(a) *Females* **(b)** *Males*

Figure 2.13: *Stacked boxplots of level 1 residuals from the fitted linear model (a) Females and (b) Males*

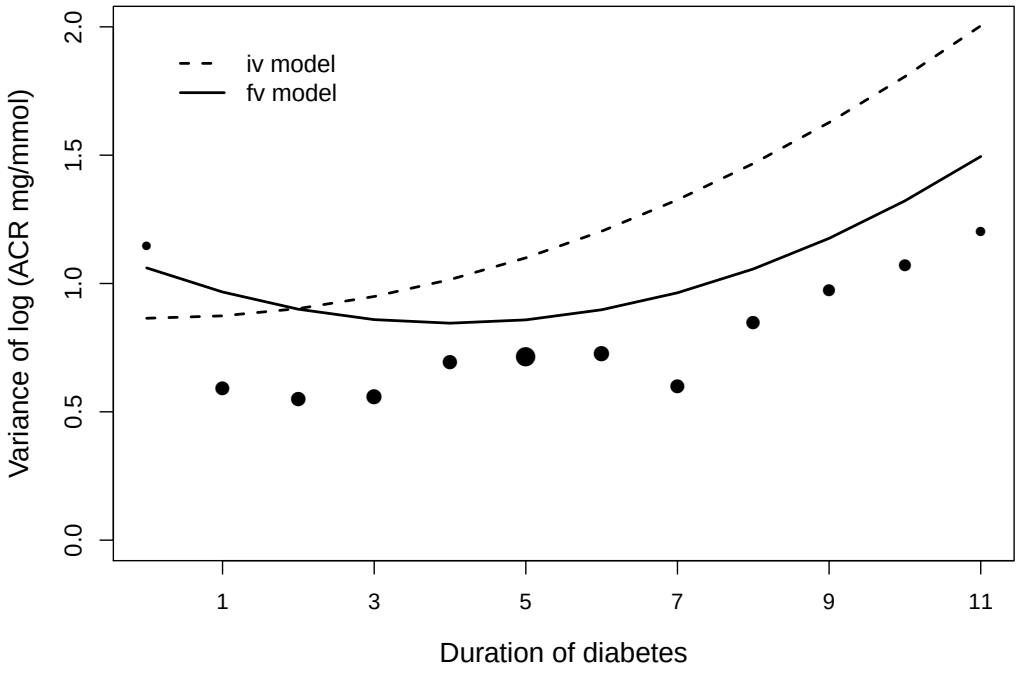


Figure 2.14: *The predicted total variation from the increasing variance (iv) model (dashed line) and the flexible variance (fv) model (solid line) against the observed data (black blobs)*

Prediction

Relaxing the assumption of independence has no effect on the fixed parameters (the average intercept and slope) but increases significantly the estimate of the baseline variance and to a lesser extent the slope variation.

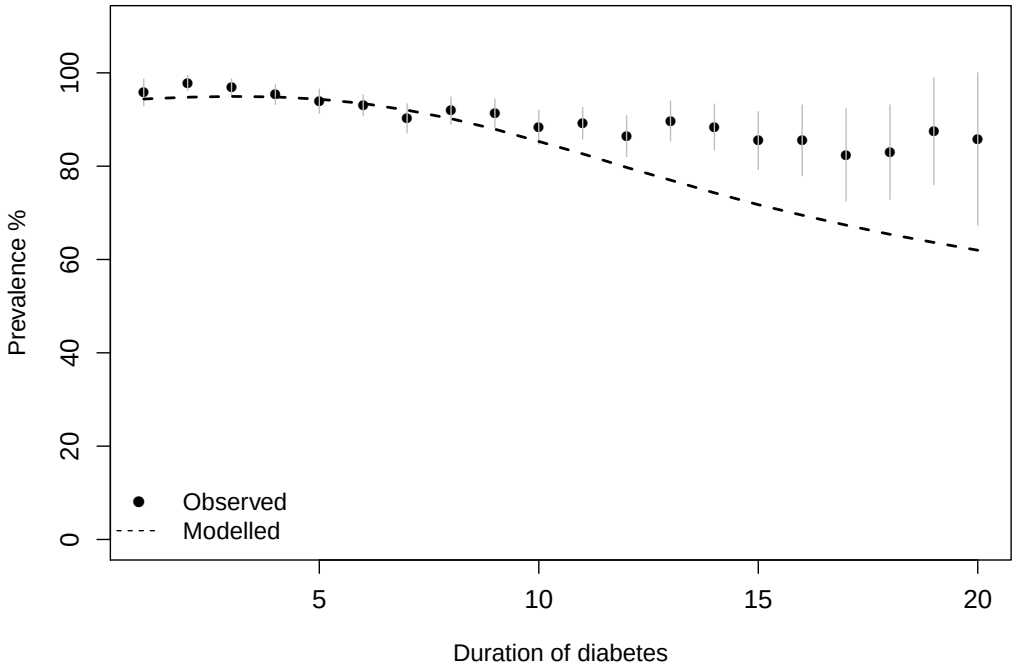


Figure 2.15: *Plots of observed proportions of assessments that would be positive with re-created outcome data from the flexible variance model*

Predictive Validation

The estimates from the flexible variance model were used to re-create the outcome (log ACR). The simulated outcome and the actual outcome were then categorised as positive for persistent microalbuminuria by the three test algorithm. Figure 2.15 plots the proportion of positive tests to the total for both the observed and simulated data.

T model

The increasing variance model and the flexible model were re-estimated using the JAGS to check consistency with the likelihood based method. Both showed close agreement indicating that the priors distributions added little discernible information to the existing data. The model was then adapted to the T model and was re-estimated with three parallel chains with 1000 burn-in iterations and

Parameter	Flexible variance model	T model
Average baseline log(ACR) mg/mmol at diagnosis	-0.22	-0.19
Average increase in log ACR mg/mmol	0.047	0.041
Std Dev of baseline log ACR (mmol/mmol)	0.74	0.69
Std Dev of increase in log ACR (mmol/mmol)	0.12	0.11
Correlation between baseline and increase	-0.635	-0.58
Short-term variance of log(ACR) mg/mmol	0.53	0.784

Table 2.7: *Comparison of the parameter estimates from the T model and flexible variance model*

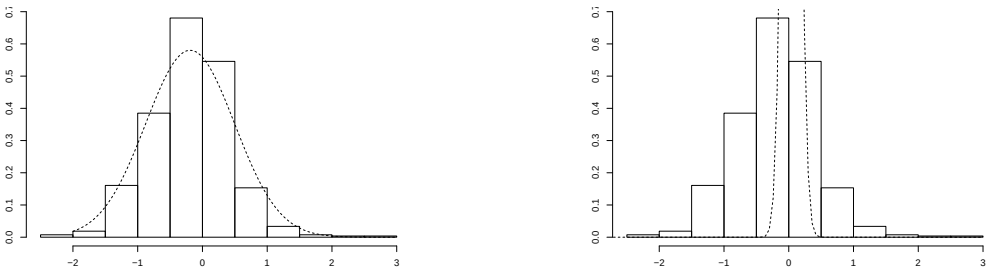
then an additional 5000 with every other draw of the Monte-carlo Markov chain (MCMC) removed (thinning). Convergence was assessed using the Gelman-Rubin (GR) statistic. The GR statistic approaches unity when the pooled within-chain variance dominates the between chain variance, indicating that the chains have reached their target distribution [80]. The model was run until the GR statistic was 1 to three decimal places for all parameters and the effective sample size was 2400 for all but the variation in slope terms σ_β.

Parameter estimates

Parameter estimates are presented in table 2.7 alongside the same or equivalent parameter estimates from the flexible variance model.

Diagnostics

The T model retains normal assumptions for the distributions of intercepts and slopes and hence these require checking. Figure 2.16 shows histograms of the random intercept (α's) and slopes (β's) terms from the T model with a normal density curve overlaid (mean and standard deviation taken from the estimated parameters).



(a) *Random intercepts* **(b)** *Random slopes*

Figure 2.16: *Histograms of the estimated random effects for the T model panel (a) the intercepts α_i and panel (b) the slopes β_i with estimated density overlaid*

Prediction

Figure 2.17 shows observed variance of log(ACR) against the variance predicted by the T model.

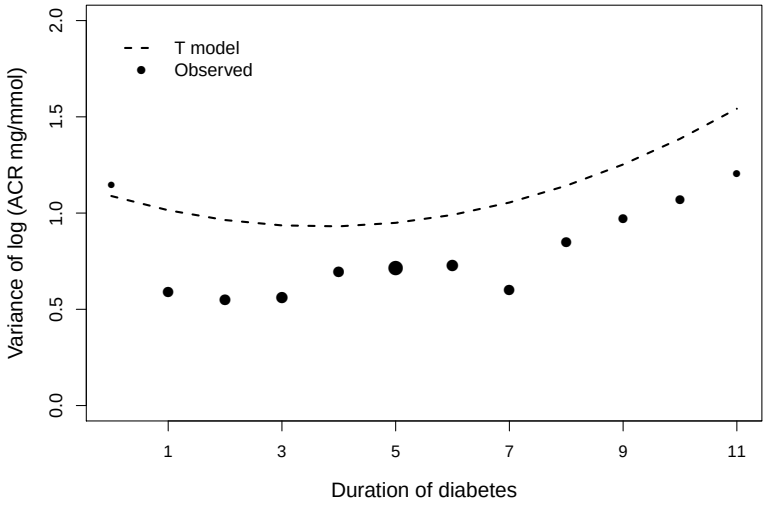


Figure 2.17: *Plot of observed variance against the variance predicted from the T model*

Predictive validation

The estimates from the T model were used to re-create the outcome (log ACR). The simulated outcome and the actual outcome were then categorised as positive for persistent microalbuminuria by the three test algorithm. Figure 2.18

plots the proportion of positive tests to the total for both the observed and simulated data.

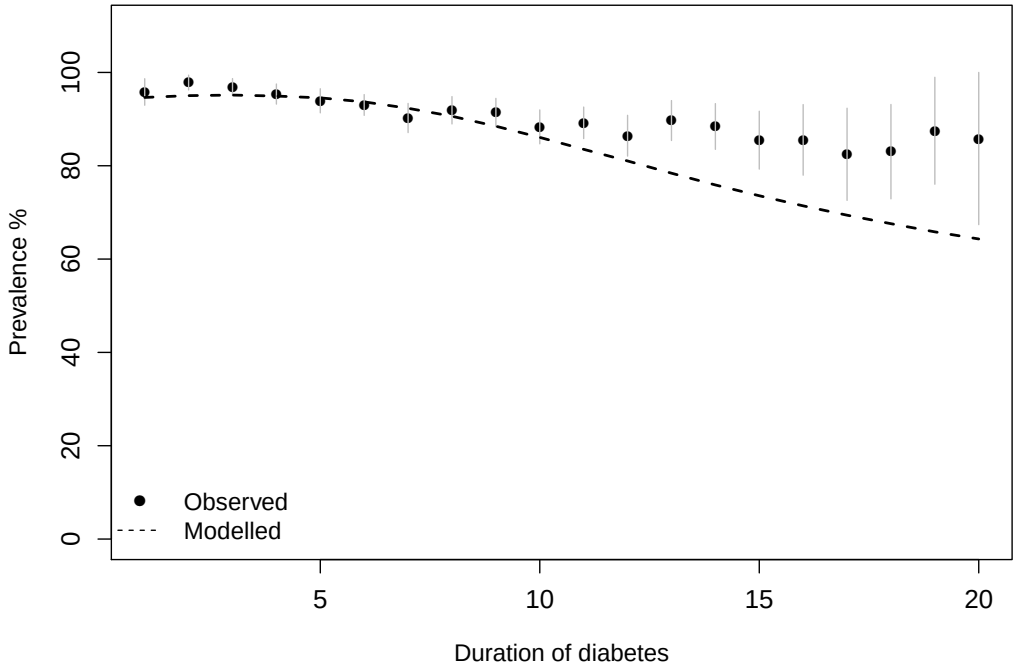


Figure 2.18: *Plots of observed proportions of assessments that would be positive for microalbuminuria against the proportions of positive tests expected under the T model*

2.6 Evaluation of monitoring prior to initiation of ACEi therapy

The frequency of monitoring albumin-creatinine was varied from the reference annual interval to intervals of 6 monthly, 2 years and 3 years. The analyses for different intervals of monitoring were repeated using different models for the progression of ACR.

Each of the table of results from the individual patient simulations (tables 2.8,2.9 and 2.10) can be interpreted as follows. The first row gives the average number of visits and equivalently ACR assays that a person undergoes before a clinical diagnosis of microalbuminuria under the four different monitoring strategies. The second row of each table gives the percentage of people who receive a positive diagnosis of persistent microalbuminuria during the course of the monitoring programme and the third row gives the average age of the first positive diagnosis and therefore starting ACEi therapy. Row four gives the percentage of false positive tests that were false at the time (a mixture of over diagnosed and premature diagnoses). Row five gives the average length of time from the false positive diagnosis and the time at which the underlying level of albumin creatinine ratio caught up with the diagnosis. Rows six gives the percentage if people who at some point in the monitoring programme who were incorrectly given a negative diagnosis by monitoring. Row seven (average delay time) is the average length between development of albuminuria and detection by monitoring, these figure include those who did not have a false negative result but developed microalbuminuria between assessments (interval cases). Finally, the last row gives the percentage of people who were overdiagnosed by monitoring. Overdiagnosis in this case is defined as a person who has a false positive diagnosis but never goes on to develop true microalbuminuria in their lifetime.

Outcome	Interval of monitoring ACR prior to ACEi therapy			
	6 months	1 Year	2 Year	3 Year
Average number of visits/assays prior to diagnosis	44	25	14	11
Percentage of people receiving positive diagnosis	62%	61%	59%	58%
Average age when starting ACEi therapy (years)	19	21	22	24
Percentage with false positive diagnosis	53%	43%	32%	26%
Mean (median) advance time in years	8.0(6.0)	6.7(5.0)	6.3(4.5)	6.1(4.2)
Percentage of people with false negative diagnosis	6%	10%	14%	16%
Mean number of false negative assessments	1.3	1.7	1.5	1.4
Mean (median) delay time in years	1.3(1.1)	2.2(2.0)	3.8(3.4)	5.3(4.8)
Percentage overdiagnosed	25%	20%	15%	13%

Table 2.8: *Resource and clinical implications of different intervals of monitoring under the IV (increasing variance) model*

Direct method

As stated in section 2.4 there is formal way to conduct individual patient simulation with the direct method.

Increasing variance model

Table 2.8 shows the implications of monitoring at different intervals in a cohort of people with type 1 diabetes with similar age and gender distributions as those of the ORPS cohort using the increasing variance model parameter estimates for progression of ACR.

Flexible variance model

Table 2.9 shows the implications of monitoring at different intervals in a cohort of people with type 1 diabetes with similar age and gender distributions as

those of the ORPS cohort using the flexible variance model for progression of ACR.

Outcome	Interval of monitoring ACR prior to ACEi therapy			
	6 months	1 Year	2 Year	3 Year
Average number of visits/assays prior to diagnosis	44	25	14	10.6
Percentage of people receiving positive diagnosis	64%	62%	60%	58%
Average age when starting ACEi therapy (years)	20	21	23	24
Percentage with false positive diagnosis	57%	45%	33%	26%
Mean (median) advance time in years	6.5(5.0)	5.7(4.0)	5.4(4.0)	5.4(4.0)
Percentage of people with false negative diagnosis	5.5%	10.3%	15%	16%
Mean number of false negative assessments	1.8	1.7	1.5	1.3
Mean (median) delay time in years	1.2(1.0)	2.1(1.9)	3.7(3.3)	5.1(4.6)
Percentage overdiagnosed	25%	19%	15%	13%

Table 2.9: *Resource and clinical implications of different intervals of monitoring under the FV (flexible variance) model*

T model

Table 2.10 shows the implications of monitoring at different intervals in a cohort of people with type 1 diabetes with similar age and gender distributions as those of the ORPS cohort using the T model for progression of ACR.

Outcome	Interval of monitoring ACR prior to ACEi therapy			
	6 months	1 Year	2 Year	3 Year
Average number of visits/assays prior to diagnosis	51	27	15	11
Percentage of people receiving positive diagnosis	58%	57%	55%	54%
Average age when starting ACEi therapy (years)	21	22	24	25
Percentage with false positive diagnosis	50%	38%	27%	21%
Mean (median) advance time years	5.2(4.0)	4.5(3.3)	4.2(3.0)	4.2(3.0)
Percentage of people with false negative diagnosis	6%	11%	14%	16%
Mean number of false negative assessments	1.8	1.6	1.4	1.3
Mean (median) delay time (years)	1.1(1.0)	2.0(1.8)	3.5(3.2)	4.9(4.5)
Percentage overdiagnosed	20%	16%	12%	10%

Table 2.10: *Resource and clinical implications of different intervals of monitoring under the T model*

2.7 External Validation

Surrogate method

The validation data set had twice-weekly measurements on 57 men and 42 women over a period of six weeks. Median (IQR) duration was 5 (2,10) years and mean (SD) body mass index was 33(6) Kg/m^2. There were 94 valid measurements of ACR at baseline, 94 at week two, 91 at week 4 and 93 measurements at 6 weeks. Eighty four people had all four measurements, 11 had three measurements, 1 had two and 1 one measurement of ACR and one person had no ACR taken in the six weeks.

To assess sensitivity to which three measurements from a total of four should form the surrogate I calculated the number of positive tests (ACR greater than gender specific threshold) for each set of three in turn. For the first two combinations, there were 13/93 = 14% positive tests in the surrogate outcome and the third method produced 15 from 93. Considering this difference to be negligible I used the first three measurements (week 0,2,4) as the surrogate and the fourth (week 6) as the observed measurement for the analysis. Dichotomising the observed measure produced 16/93 = 17% positive measurements at week 6. Table 2.11 shows a 2 by 2 table of the observed and surrogate true results. On the basis of just the fourth measurement, the prevalence of microalbumin-

		Observed Measurement (week 6)		
		$<$ threshold	$\geq$ threshold	Total
Surrogate true	$<$ threshold	76	4	80
	$\geq$ threshold	1	12	13
	Total	77	16	93

Table 2.11: *Results from the ACR study*

uria is 17%, whereas based upon the surrogate true estimate it is 14%. The estimated false-positives rate is therefore,3/93 = 3% and the true positive rate is 12/93 = 13%. The estimated false negative rate is 1/93 = 1% and true

negative rate (76/93 = 82%). This equates to a sensitivity of 12/13 = 92% and specificity of 95%. Table 2.12 compares the observed rates from the ACR cohort with the rates predicted by the direct method, increasing variance model (IV model) and T model.

Index	ACR Cohort	Direct method	iv model	T model
True −	82%	75%	80%	84%
False −	1%	3%	3%	2%
False +	4%	5%	11%	12%
True +	13%	17%	6%	3%

Table 2.12: *Comparison of the model predictions of the rates of true and false positives, true and false negatives compared to the observed rates in the ACR Cohort study*

Regression to the mean method

This method tests each models' ability to predict the degree of regression to the mean over a two measurements. I split the baseline measurements into tertiles of log ACR, and took the mean and standard deviation of the baseline and last measurement (6 weeks). The upper tertile group had mean (sd) log ACR 1.24(1.08) at baseline and mean (sd) log ACR 1.08(1.42) 6 weeks later, a reduction of 0.18 mg/mol.

Index	ACR Cohort	Direct method	IV model	T model
Mean (SD) mg/mmol upper tertile	1.24 (1.08)	1.4 (0.80)	1.27 (0.62)	1.06 (0.52)
Mean (SD) follow up	1.08 (1.42) mg/mmol	1.2 (0.99)	0.71 (1.04)	0.38 (0.93)

Table 2.13: *Comparison of the model predictions of the amount of regression to the mean between two time points*

Prevalence method

I searched Pubmed, Google scholar and Scopus databases using combinations of the terms *prevalence, microalbuminuria, type 1,cohort, database, nephropathy* and looked for prevalence estimates (cases/total) for microalbuminuria and or macroalbuminuria. Studies were included if estimates of prevalence were based on at least two positive sequential tests of urinary albumin excretion, were stratified by gender and reported either the duration of diabetes or the average duration. Estimates were taken text, baseline or result tables, or directly from graphs using the excel grabit macro if numbers were not reported in the text. Table 2.14 summarises the studies included in the validation and figure 2.19 shows the study estimates against the fitted prevalence curves from the models.

Study	Study type	Country	Age range	Definition of Microalbuminuria	Duration of follow-up (years)	N (total)	Estimate
DCCT (deBoer)	Prospective cohort	USA	13 - 39	2 UAE $>$ 30 mg/24h	10 20 30	632 310 48	14% 33% 38%
Cho *et al*	Cross-sectional	Australia	11 - 17	2 AER $>$ 20μg/min	4	729	3%
Orchard *et al*	Prospective cohort	USA	28	2 AER $>$ 20μg/min	19	658	21.3%
Salgado	Cross-sectional	Brazil	N.A	AER $>$ 20μg/min	11.3	205	11.2%
Andersen	Prospective cohort	Denmark	0 - 30	4 protein $>$ $0.5g/24$ h in $\geq$	5 15 25 35	1014 479 227 107	4.8% 15.7% 35.7% 43.9%
Hovind	Prospective cohort	Denmark	N.A	2 UAE $>$ 30mg/24h	5 10 15 20	190 131 90 62	3.1% 18% 24.6% 33.5%
Alleyn	prospective cohort	USA	12.9	2 AER $>$ 20μg/min	5.6	471	4.6%

Table 2.14: *Characteristics of the studies included in the external validation. References are deBoer [81],Cho [82], Orchard [83], Salgado [84], Andersen [85], and Alleyn [86]*

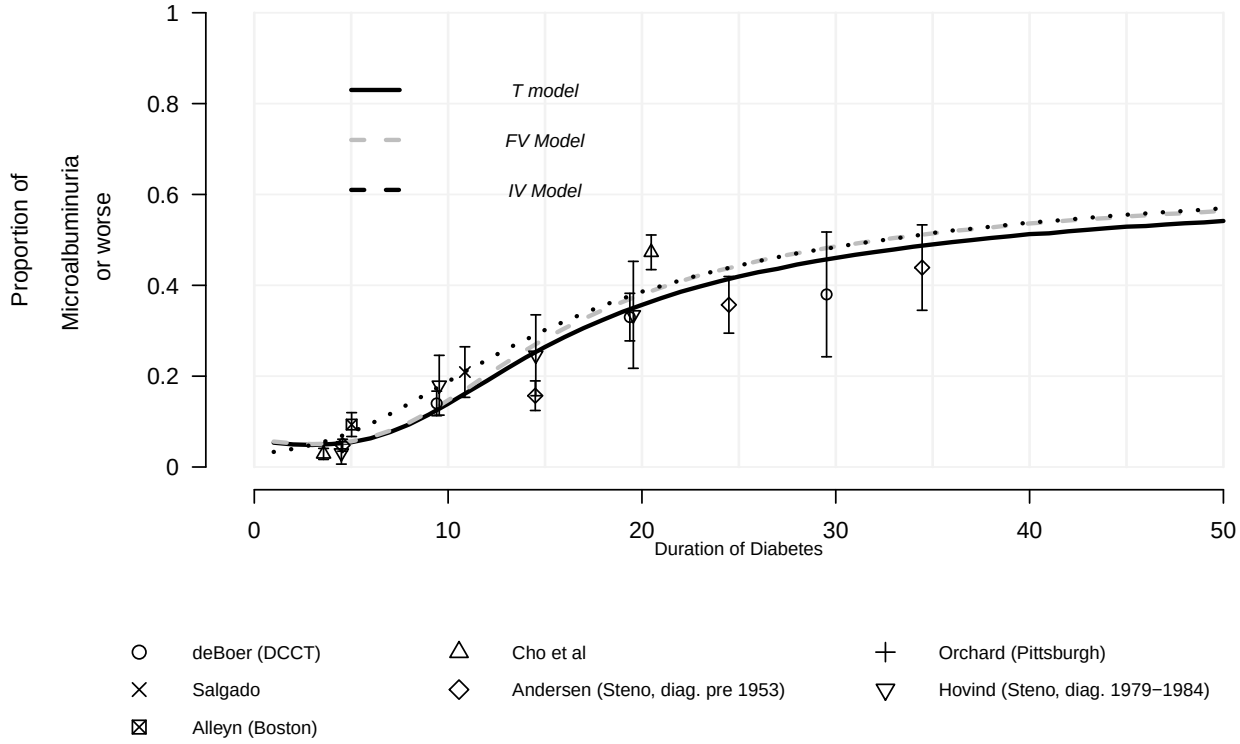


Figure 2.19: *External validation of the model for continuous log ACR. Symbols represent point estimates of prevalence from published studies with error bars representing 95% confidence intervals. Lines represent estimated prevalence rates of persistence micro-albuminuria using model parameters.*

2.8 Conclusion

In this chapter, I set out to establish whether the *Glasziou* monitoring model could be extended to a monitoring test with an irregular (non-normal) distribution. To test this, I used data from a longitudinal study in which ACR had been measured annually on people with type 1 diabetes. ACR is a good candidate measure as it is highly positively skewed even after log transformation. I attempted to fit a number of models and examined each of them for statistical validation, first by checking the assumptions of the model and then the goodness of fit for the observed and simulated data. I also examined the models for external or *clinical* validity by comparing predictions of true and false positives, true and false negatives from the models with outcomes observed in an independent data set with repeated measures of ACR. Lastly, I created an individual patient simulation to evaluate the impact of changing the interval of monitoring through statistics relating to the number of tests, rates of positive diagnoses, advance time, delay time and rates of overdiagnosis.

Statistical Validation

Direct method

The direct method produced negative estimates of the long-term variance component and lower estimates of the total variance at times $2, 3$ and 4 than at baseline. This was unexpected and at first appeared problematic because variances cannot be negative however, this problem also occurs with multi-level models with complex variance structures and advice is usually to not consider individual terms as variances in their own right, but as components of the total variance. This pattern will occur when there is a negative correlation between the baseline and progression terms and variation between slopes is small. This negative correlation implies that people with high baseline values progress more slowly than the those with lower values who seem to catch up. Hence, there is a coming together of values or as it often referred to a “fanning in” of the data.

Why this occurs in data is not clear, but it may be an artefact of the design of the study and its inclusion criteria, a *regression to the mean* effect. An alternative explanation is that there is true regression of raised albuminuria, and this is noticeable early on because of the low prevalence. The decreasing variance directly affects the ratio of false to true positives and contradicts the idea that false positive rates will always decrease with time. The direct method estimates that after one year the ratio of false to true positives is 0.34 for an interval of one year, then 0.9 for an interval of 2 years, and a ratio of 0.36 for three years. It is only after six years, that the ratio is lower than it was at an interval of one year. In previous examples (cholesterol, blood pressure, HbA1c), the ratio of false to true positives has been a monotonic decreasing function of the length of the interval. This has been used to justify increasing the monitoring intervals beyond current guideline recommendations. This finding is interesting if real because it may suggest that optimal monitoring intervals for some conditions need not be regular but should change as function of age or duration of disease. To assess whether this finding is due to the method, more rigorous models were applied to the data.

Increasing and flexible variance models

Random-effects models (the increasing and flexible variance versions) were fit to the ACR data. Assumptions of normality for the random effects were assessed using histograms and showed that the distribution of the slope random effects terms had a distinct asymmetric shape and the assumption of normality looked untenable. The flexible variance model did explain more of the variation in the data than the increasing variance model (had better measures of fit) and also tracked the total variation over time better than the increasing variance version did. This was most likely a result of not restricting the correlation between the intercepts and slopes to be zero (as it is in the increasing variance model). The reasoning for this is the following. The variance over time is due to two components, one relating to the variation in slopes ($\sigma_\beta t^2$) and another to do

with the covariance between the slopes and intercepts $2\phi t$. The variogram is a function of the slope term but not the covariance term and so this statistic can be useful to assess which components of the total variation fit or do not fit the data. Because, if anything, the variogram fit for the flexible variance model is a poorer fit (figure not shown) than the increasing variance model I assumed that the difference is due to allowing the covariance to be estimated rather than constrained to be zero. The assumptions of homogeneity of measurement error also appeared untenable in both increasing and flexible variance models and this motivated the development of the T model. Statistical validation against simulated data set revealed that short term predictions of the prevalence of microalbuminuria were good for the first 10 years, but systematically over-predicted after that.

T model

The T model was devised as a way to model the heterogeneous variation in measurement error or short-term variation across people. I thought that this may in turn improve the normality of the random effects and resolve this issue as well. Histograms of the random effects were an improvement over the same plots for the increasing and flexible variance models but there still appears to be substantial lack of fit in the way the slope terms vary. In addition, the model over-predicts the total amount of variation and over-predicts the prevalence as did the flexible variance model.

Clinical validation

In an attempt to clinically validate these models I used three methods; the prevalence method, regression to the mean test and the surrogate true method. A first and less stringent test of the model is to gauge how well they predict the prevalence over time (prevalence method). I found estimates of the prevalence of microalbuminuria from published studies and created a pooled prevalence curve over a period of 2 to 35 years duration of diabetes. I compared

estimates of prevalence from the T and flexible variance model to the pooled curve. Both models, although seem to over predict the prevalence in the ORPS cohort, compare well with external estimates of prevalence. The T and flexible variance model estimates appear to asymptote (level off) at a prevalence of microalbuminuria of around 50%, which is reflected in the literature.

The regression to the mean test is supposed to differentiate between models that can predict the cross-sectional averages and those which capture the true variance and covariance structure of the data. I used an independent data source to compare the degree of regression to the mean predicted by the three candidate models. I found the people whose ACR at baseline fell into the fourth quartile and calculated the mean and standard deviation, I then calculated the mean and standard deviation of the follow-up measurement. The proportion of regression to the mean to the baseline mean was $((1.24 - 1.08)/1.24 = 0.13$ in the ACR cohort study. The proportion predicted by the direct method was $.2/1.4 = 0.14$, the flexible variance model 0.44 and T model 0.64. The method which comes closest to predicting the correct amount of regression to the mean is the direct method, with the random effects models over predicting by a considerable margin. The expected amount of regression to the mean is related to both the distance of the threshold to the mean and the ratio of the between person variation and the total variation. Hence we can think of the regression to the mean test as being discriminatory for models with the proportionate ratio of the between-person variation to total variation.

The final validation was the surrogate method which compared estimates of true and false positives, true and false negatives with those estimated from the models. Although it is difficult to assess each of the models across all four domains, as for example, the T model predicts the level of true and false negative proportions well but does badly on predicting the proportion of false and true positive. As the proportions of positive tests that are false is arguably the most important quantity then our focus should be there. The ratio of false to true positives for the ACR cohort data was $(4/13 = 0.31)$, for the direct

method it was (0.29), the flexible variance model ($11/6 = 1.8$) and T model ($12/3 = 4$). It is clear that the direct method comes very close the observed ratio of false to true positive test and that the T and flexible variance model predict a much higher rate.

Evaluating strategies for monitoring of diabetic nephropathy

The results from the individual patient simulation analysis shed light on how monitoring frequency affects health outcomes and factors affecting the resources of health providers. Intuitively, monitoring more often increases the number of tests and this will add significantly to the cost of monitoring. This may seem obvious but under the confirmatory testing strategy there is potential for either $1, 2$ or 3 tests at each assessment and the exact number will depend on the underlying true prevalence of microalbuminuria and chance variation. I found that with 6 monthly monitoring there are on average three times as many tests as biennial monitoring and nearly twice as many as annual monitoring. The patient simulation modelling shows that annual monitoring will diagnose between 57% (T model) and 62% (IV model) of people as having persistently raised albuminuria over a lifetime of monitoring and these diagnoses will be a mixture of correct diagnoses, initially false positives who go on to develop microalbuminuria and overdiagnosed people (those who test false positive but never go on to develop microalbuminuria in their lifetime). A prevalence of around $55 - 60\%$ over thirty to forty years duration is in good agreement with estimates of prevalence from cross sectional studies, for example Parving *et al* [87] who reported the prevalence of micro and macroalbuminuria using 24 hour urine collection (analysis of albumin excretion by radio immunoassay) of 45% at 40 years duration.

More frequent monitoring corresponded to starting ACEi therapy earlier. For three yearly monitoring, the average age starting ACEi was around 24 years, for two yearly monitoring the average age was 23 years, annual monitoring,

21 years, and 6 monthly monitoring 20 years. The estimates of the average age starting ACEi from the T model were two years later than equivalent estimates from the increasing variance model for all intervals of monitoring reflecting the lower cumulative incidence. Whether there is any benefit from starting therapy earlier by this number of years is unknown and the first RCT of ACEi therapy in adolescent type 1's is currently underway [88]. There is little evidence at present that treatment has any effect on albumin-excretion in type 1's with normoalbuminuria. In a recent meta-analysis of the impact of renin-angiotensin-aldosterone system inhibitors (ACEi and ARB's) on type 1 and type 2 diabetic patients, there was no evidence that treatment reduced the ratio of mean urinary albumin excretion in patients with type 1 diabetes and normal albumin excretion [89].

This modelling suggests that monitoring for microalbuminuria leads to a lot of false positive test results. Some of these false positives will represent early diagnoses because although at the time of the positive test their true value was lower than the threshold, at some point in the future their condition catches up with their diagnosis. Others test positive by chance possibly because of confounding factors (consumption of protein, exercise, urinary tract infections, etc) and will never develop microalbuminuria in their lifetime, and others will test false positive and later develop microalbuminuria but progress so slowly that they never develop clinical proteinuria. In both of these cases, these people are overdiagnosed.

Some people will develop true microalbuminuria but then regress. There is some evidence of considerable degrees of regression from microalbuminuria to normoalbuminuria, either transiently or permanently. Perkins *et al*[43] reported that 58 percent of people in their study with microalbuminuria regressed over a six year period and this was not associated with the use ACEi treatment. Schultz also *et al* [64] reported that regression to normoalbuminuria was evident in 21 of the 63 (1/3) individuals who developed microalbuminuria in a median follow-up of five years. This is important as currently, detection of

raised albuminuria (outside of the UK) leads to therapy with ACEi or ARB's and true regression cannot be observed over and above reductions in urine-albumin brought about by treatment.

A consistent pattern across all three models is one of increasing numbers of false negatives as the interval of monitoring increases. The rate of false negatives with a three year interval is between two and three times the rate when monitoring takes place every 6 months. The false negative represent people whose assessment has failed to detect the early stages of diabetic nephropathy. Current guidelines for monitoring microalbuminuria state that an initial normal result means that no further investigation takes place until the next appointment. Therefore, it is easier in theory for false negatives to occur, than false positives, because no confirmatory testing (two positive tests from three, see figure 2.1) is done.

False negatives incur a delay to the detection and potential treatment which is exacerbated by increasing the interval of monitoring. If disease becomes incident shortly after an assessment and is missed at the next assessment then delay would be considerable. For monitoring intervals of three years this could mean a delay of up to 6 years. By definition, any false negative incurs a delay in detection at least as long as the interval of monitoring. Hence, for an interval of 6 months, the minimum time a false negative can have a detection delay is 6 months. See figure 1.5, section 1.6, page 19. For an interval of half a year between assessments, 40% of people who at any time are false negatives have a delay less than a year, but 45% experience a delay between 1 and 2 years, and 15% a delay longer than 2 years. When the interval is extended to 1 year, 60% of people who had a false negative result experience a delay between 1 and 2 years, and 25% a delay between 2 and 3 years.

Discussion

The conclusion of the cost-effectiveness analysis in our HTA monograph on optimal strategies for monitoring kidney disease was that current policies of

annual screening were supported even with substantial overdiagnosis as harms from overtreatment were minimal. Less frequent testing would save relatively small amounts of money at the expense of failing to identify small numbers of individuals at risk [76]. The report also stated that changing the values of progression of ACR had a large impact on the results for both diabetes models (type 1 and type 2). Lower rates of progression with less people requiring treatment and fewer progressing to end stage disease saw a reduction not only in costs but also QALY's and taking the average progression to be the lower end of the 95% C.I. meant that annual screening was no longer cost effective [76]. The importance of obtaining accurate estimates of the incidence of progression was also noted by Kiberd and Jindal [90] in an early cost-effectiveness analysis on screening to prevent renal failure. The authors compared two strategies for monitoring microalbuminuria, strategy A which consisted of screening for microalbuminuria in patients with five years of duration and treatment with ACEi if two tests were positive and a second strategy (B) in which patients were not screened but could be treated on the basis of hypertension or macroproteinuria. They concluded that strategy (B) would be favourable if the cumulative lifetime incidence was less than 20% or equivalently if the prevalence of hypertension was greater than 65% at the onset of microalbuminuria.

The strength of the simulation model as a tool for evaluating monitoring strategies is that I have been able to explore how changing the interval of monitoring impacts on a number of critical outcomes relating to patient health. Whilst the proportions of false positives and false negative are important it is also crucial that we understand how long a person's diagnosis has been advanced or delayed as a result of monitoring. Once these statistics are coupled with costs and benefits a monitoring strategy can be evaluated from a cost and utility perspective.

The limitations of the modelling approach are primarily in the assumptions that are required to derive results. The assumptions are numerous and often difficult to justify empirically. With all types of mathematical models, it can

be the detail that is omitted that lead to the biggest weaknesses. I could not incorporate transition probabilities to advanced proteinuria and end-stage renal disease, or model the increased risk of death that is evident from advanced diabetic nephropathy. The poor fit of the models for the progression of albumin excretion were likely to be the cause of over prediction of the prevalence over time. Attempts to transform the outcome and model the variance were largely unproductive. In the next chapter, I detail the results of transforming the continuous measure of ACR to a dichotomous outcome, equivalent to two states; persistent microalbuminuria and normoalbuminuria. This transformation was considered first and foremost as a means to reduce the impact of extreme ACR values on the progression model but has also led to a potentially useful extension of the normal model to monitoring tests with binary or dichotomous outcomes.

Chapter 3

Models for dichotomised data: Persistent micro-albuminuria

3.1 Introduction

Monitoring chronic diseases often involves interpreting a number of test results and clinical judgement to inform decisions about treatment. In the previous chapter, we saw that current practice to detect persistent microalbuminuria involved an algorithm based upon (up to) three separate ACR test results. This confirmatory testing algorithm led to a binary outcome of persistent raised or normal albumin excretion. Another example of this type of interpretation is described in the current NICE guidelines for monitoring glaucoma. Guidelines for monitoring suggest continuous measures of intra-occular pressure (IOP) be interpreted as *at target* or not. If IOP is not adequately controlled medically then the target should be reviewed or treatment plan changed [91]. Similarly, the risk of conversion to COAG, which is to be judged clinically in terms of age, IOP, central corneal thickness and the appearance of the optic nerve head is combined to form a simple low or high risk dichotomy. Table 3.1 shows the current recommendations as they appear in the NICE guidelines.

Many mathematical models for the progression of illness exploit the natural interpretation of stages of disease rather than continuous measurements used to define them. Even when there is no obvious interpretation in terms of stages, continuous measures are sometimes categorised into stages. One example of

Clinical Assessment			Monitoring Intervals (months)	
IOP at target	Risk of conversion to COAG	Outcome	IOP alone	IOP, optic nerve head and visual field
Yes	Low	No change in treatment plan	Not applicable	12 to 24
Yes	High	No change in treatment plan	Not applicable	6 to 12
No	Low	Review target IOP or change treatment plan	1 to 4	6 to 12
No	High	Review target IOP or change treatment plan	1 to 4	4 to 6

Table 3.1: *Recommended monitoring intervals for people with occular hypertension (OHT) or suspected chronic open angle glaucoma (COAG) who are recommended to receive medication. IOP - intra-occular pressure. Table adapted from NICE guideline CG85 [91]*

this is in an analysis by Satten and Longini [92], where they modelled decline in CD4 cell count before the onset of clinical AIDS. Their model categorised CD4 count into six levels with an additional further absorbing state to signify the development of AIDS. Figure 3.1 shows pictorially one version of their multistage model for CD4 count and the development of AIDS.

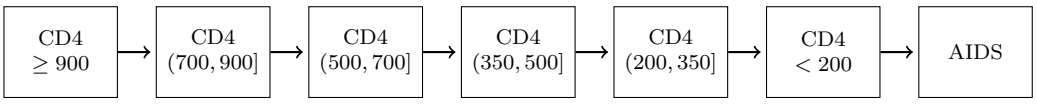


Figure 3.1: *Unidirectional model of CD4 count (cells/mm^3) prior to clinical AIDS , after Satten and Longini [92]*

A monitoring model for dichotomous outcomes

Tests that have dichotomous outcomes are subject to same sources of short-term variation as continuous markers (laboratory/assay variability and biological variability) and thus frequent monitoring in people at low risk of an event or change in the underlying condition may only detect spurious changes or false positive results. Therefore, a model that can estimate both true underlying

disease and short-term variation in dichotomous outcomes could be useful for evaluating strategies for monitoring involving dichotomous tests.

A logistic regression is commonly used to model dichotomous data. A logistic regression models the probability p that an event will occur as a function of independent or explanatory variables. As p cannot be less than 0 or greater than 1, it is customary to transform p using the logit function;

$$logit(p) = log\left(\frac{p}{1-p}\right) \tag{3.1}$$

and $logit(p)$ is assumed linear in terms of the independent variables;

$$logit(p) = \alpha + \beta x$$

A random effects logistic regression model can be used to fit separate coefficients to individuals and be used to model repeated measures data;

$$\begin{array}{ll} logit(p) = & \alpha_i + \beta_i x \\ \alpha_i \sim & N(\mu_\alpha, \sigma_\alpha^2) \\ \beta_i \sim & N(\mu_\beta, \sigma_\beta^2) \end{array} \tag{3.2}$$

A model with independent random effects (as in equation 3.2) is example of an "increasing variance" model for the latent variable.

An important distinction between the continuous and logistic models, is that the logistic model does not have an independent parameter (σ) for the short-term variation. This at first seems to be a major obstacle to using these models to describe monitoring tests. A solution is provided if we interpret the logistic regression model in the latent data formulation [69]. This says that the outcome of the dichotomous test y is associated with an unobserved or latent outcome z_i defined as follows;

$$y = \begin{cases} 0 \text{ if } z_i < 0 \\ 1 \text{ if } z_i \geq 0 \end{cases}$$

and the latent variable is given a regression type model;

$$z_i = \alpha + \beta x + \epsilon_i \tag{3.3}$$

and where the errors ϵ_i are given a logistic distribution with parameters ($m = 0$ (location) and $s = 1$(scale)). In the linear model, the degree of variability was captured by a single parameter σ which was estimated using the data and could be tailored to match the short-term variation in the monitoring test. In the logistic model (with latent interpretation) the parameters controlling the short-term variation are fixed and not estimated, but controlled by another mechanism.

Accommodating tests with different short-term variability

The short-term variability is arguably one of the most important components of the monitoring model. In the logistic version of the model, the equivalent measurement error term cannot be estimated and has to be fixed as it cannot be identified simultaneously with the progression parameters. This at first appears to be a significant limitation, because there does not appear to be a means to distinguish biomarkers with large short-term variation (such as blood pressure) from biomarkers with small short-term variation such as HbA1c. The model can, in fact, accommodate different levels of variability by adjusting the value and scale of the other parameters; intercept (α) and slope (β terms.

Aims

Before starting the work for this chapter, I set out to show that the Glasziou method can be extended to model binary monitoring tests and then be used to evaluate monitoring and screening strategies. To do this, I dichotomised continuous ACR data to form a binary (0/1) outcome. Whilst this transformation would not necessarily done in practice, as dichotomising often leads to losses of information, the purpose was to illustrate the potential for the monitoring model to be applied to a binary outcome.

This chapter establishes the mathematical framework for the dichotomous outcome which is further extended to models for ordinal outcomes in chapter

4. As with the previous chapter, a statistical model is fitted to cohort data and then used to evaluate monitoring strategies and the impact of changing the interval of monitoring. The work in this chapter contributed to the development of the statistical model building used in the HTA funded project, *Optimal strategies for identifying kidney disease in diabetes: properties of screening tests, progression of renal dysfunction and impact of treatment* [76].

3.2 Data for modelling persistant microalbuminuria

The logistic model for persistent microalbuminuria is estimated using the same data (ORPS) as chapter 2, a full description of the data is given in section 2.2 on page 29.

Some additional data management was required to transform each within-year assessment into a binary outcome representing abnormal or normal albumin excretion. For each cluster of measurements, I applied the three test algorithm and considered an individual test to be positive if it was equal to or greater than the thresholds currently recommended by NICE (2.5 mg/mmol for males, 3.5 mg/mmol for females). In cases where there were missing test results, I excluded assessments if a) the first test was missing, even if the second two were present, and b) where the second and third were missing and the first test was positive and c) if the third test was missing and it was needed to confirm persistence.

In the analysis using the continuous version of ACR, there were 535 people with usable data contributing 4749 assessments and 4756 person-years of follow-up. There were no cases in which the first test was missing, 30 cases in which the first test was positive and the second and third were missing, and 43 assessments where the third test was missing and was required for confirmation of persistence. Therefore, I excluded 73 assessments in total from the data. No-one was excluded from the data entirely (still 535 people). Table 3.2 shows the baseline characteristics split by those with persistent microalbuminuria at baseline and those with normal albumin excretion.

Variable	Persistent MA	
	Negative n = 513	Positive n = 22
Gender - Male	232 (45%)	10 (45%)
Gender - Female	281 (55%)	12 (55%)
HbA1c (%)	9.9 (2.1)	10.6 (2.8)
Age at diagnosis of diabetes	9.0 (4.0)	8.6 (4.7)
Weight (Kg)	40.2 (15.3)	39.4 (17.4)
Height (cm)	144 (18.8)	146 (20.6)
Systolic BP (mm/Hg)	100.5 (11.2)	101.5 (14.1)
Diastolic BP	64.5 (11.7)	64.5 (10.5)
Mean ACR (mg/mmol)	1.64 (10.1)	8.1 (6.0)

Table 3.2: *Table of baseline characteristics stratified by persistent albuminuria status at the first assessment. Numbers are means (SD) or counts (%)*

Validation data

To clinically validate the persistant microalbuminuria model, I compared predictions of true and false positives to outcomes from the ACR cohort study. More details of this study are given in chapter 2, section 2.2.

3.3 Methods

In this section I describe two approaches to modelling dichotomous monitoring or screening data. The first is a fixed effects approach to modelling repeated measures which relaxes the assumptions of normality in the random effects models which often prove difficult to justify. The second approach is a logistic random effects model for binary data fitted in a bayesian framework. To assess how useful these models are, I assessed both models using principles described by Altman and Royston [77] for statistical and clinical validation of prognostic models.

Fixed effects approach

In the previous chapter, I showed that when the normality assumptions are not valid, random effects models led to poor predictions of the rates of true and false positives. An alternative approach to random effects for repeated measures data is a "fixed effect" approach. Fixed and random effects models can be thought of as reflecting different assumptions about the heterogeneity across the groups, or subjects in this case. Table 3.3 compares random and fixed effects models and their assumptions about the population heterogeneity.

Model	Population Heterogeneity (measured by τ)	Repeated Measures Model Interpretation
Random Effects	$0 < \tau < \infty$	People have different but similar α and β parameters
Fixed Effects	$\tau = \infty$	People have different and unrelated α and β parameters

Table 3.3: *A comparison of two different approaches analysing grouped data*

Estimation of the fixed effects model

The fixed effects model was estimated using a standard logistic regression routine in R (*glm*). I fitted this model by treating the patient or *id* labels in the data as factors and fitted individual intercept terms and created an interaction variable (subject * time) to obtain individual progression estimates. There were 535 people in the ORPS data and the model fitted $535 * 2 = 1070$ parameters and 535 regression equations, one for each person. I used the regression equations to fit "true values" for the time points observed in the data and classified these into negative or positive using a threshold of 0. I then used the observed and fitted or true values and classified each data point as either a true or false positive, true or false negative. I then tabulated proportions by "binning" or grouping observations into years since diagnosis of diabetes.

Diagnostics

As residuals plots are not particularly useful for logistic regression because of the discrete nature of the data, I used an alternative method proposed by Gelman and Hill [69]. This method groups or "bins" residuals into categories and compares the mean residual over the values of the categories with lines to show 95% limits. Residuals from the no-pooling model were binned into years since diagnosis of diabetes and a mean was derived for each 20 years of duration. If the model is correct, the residuals will have mean zero and asymptotic standard deviation ($ASD = \sqrt{p(1-p)/n}$), where p is the estimated proportion. Approximate 95% confidence limits are found by drawing lines at $-2 * ASD$ and $+2 * ASD$. An excess of points (more than 5%) outside these limits indicates a lack of fit.

Prediction

The proportions of true and false positives, true and false negatives were used to predict the annual rates of true and false positive, true and false negative

tests we would expect if the interval of monitoring was 1 year (reference) and increasing in one year increments thereafter.

Random effects models (increasing and flexible variance versions)

The logistic random effects model was fitted in a bayesian framework. I think it is helpful to think of bayesian models as consisting of separate probability models for each level of the data. In repeated measures models, the first level represents the observed test results. The observed test results are nested within individuals (second level) which are in turn nested within population averages (third level).

The first level of the logistic monitoring model represents the probability that an assessment Y (up to three ACR tests) on the ith person at time t is positive for persistent microalbuminuria. The series of tests on an individual is modelled as a series of Bernoulli trials (0/1) with probability p_{it} hence,

$$Y_{it} \sim Bern(p_{it})$$

The probability for a positive test is logit transformed and modelled as a linear regression.

$$\begin{aligned} logit(p_{it}) = &\quad z_{it} \\ z_{it} = &\quad \alpha_i + \beta_i t \end{aligned}$$

The baseline and progression parameters (α, β) represent the second subject level of the model. As stated in the introduction (section 3.1) the measurement error term ϵ_{it} is fixed and does not need to be estimated. Both the baseline (α_i) and slope (β_i) parameters are given distributions as they were for the continuous model, hence;

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim \left[\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \rho_{\alpha,\beta} \\ \rho_{\alpha,\beta} & \sigma^2_\beta \end{pmatrix} \right] \tag{3.4}$$

where ρ can be fixed at zero for the increasing variance model or estimated under the flexible variance model. The parameters μ_α and μ_β, the population average baseline and progression represent the third level of the model.

These models were estimated using the JAGS [75] program in R. The code used for the linear random effects models needs only to be modified so that the first level of the model accounts for the binary nature of the data. The first level of the model is now;

```
for (i in 1:n) {
    y[i] ~ dbern(p[i])
    logit(p[i]) <- b[id[i],1] + b[id[i],2] * x[i]
}
```

and the second level of the model and the priors are identical.

Diagnostics

Diagnostic checks of the models are similar to those of the other random effects models. The distributional assumptions of the random effects need to be assessed graphically. The model will also be used to recreate the outcome variable and compared to the observed rates of persistence microalbuminuria.

Prediction

The random effects model parameters were used to predict the expected number of positive tests for persistent microalbuminuria over twenty years duration of diabetes and these were compared to the proportions observed in the data.

3.4 Evaluating monitoring prior to initiation of ACE-i therapy

I adapted the patient-level simulation model used for evaluating continuous ACR from the previous chapter (section 2.4) to evaluate the impact of monitoring strategies using the binary outcome for microalbuminuria. The core of the simulation model was mostly the same and only differed with respect to how the assessment result and true value was derived. In the continuous model, three measurements of ACR were simulated from distributions and a positive

test was defined using the three test algorithm. In the logistic version, the latent variable z is simulated and transformed to either a positive or negative (0/1) outcome using the relevant threshold.

The latent variable z was modelled by simulating, a baseline value of α and a slope parameter β for each person extrapolating $z_t = \alpha + \beta t$ over the course of the simulation. The true status was positive if $z_t \geq 0$ and negative otherwise. The observed test result was derived by adding random noise ϵ to z_t and also considered positive if $z_t + \epsilon \geq 0$ and negative otherwise.

The different monitoring strategies were evaluated using the same outcomes measures as in section 2.4. (1) The average number of tests/assays that a person undergoes prior to a positive diagnosis and initiation with ACEi (2) The percentage of people who test positive and (3) the percentage to the whole who test false positive. The average advance time (4) which is derived by calculating for each person the time between the first false positive test (if one occurred) and the time until the true ACR crossed the threshold. (5) The percentage of false negatives taken as the number of false negative assessments to the total and (6) the average delay time calculated as the length of time between the point at which the true albumin-creatinine ratio crossed the threshold to the time it was detected with monitoring. Finally, the percentage of people overdiagnosed by monitoring is calculated by determining those who received a false positive monitoring test but whose true albumin creatinine ratio would never have crossed the threshold.

Assumptions

I made the following assumptions in the simulation model. Firstly, there are no missed appointments and all tests give a definitive result (no undetermined test results). I further assume that all the individuals participating in the monitoring programme attend the screening appointment at intervals defined by the monitoring programme and do not miss appointments or drop out of the programme. Lastly, I assume that there are no *interval cases* outside of

the monitoring programme, i.e. people who present with symptoms of proteinuria/nephropathy that leads to a diagnosis either in primary or secondary care and not by the monitoring itself. The models was run $75,000$ times in order to minimise Monte-Carlo error for intervals of monitoring of 6 months, 1 year, 2 years and 3 years.

Clinical validation of the statistical models

The logistic models were assessed for clinical validity by comparing predicted proportions of true and false positives and negative tests against observations from other data.

Surrogate true method

The validation of the two logistic models follows the same principles of those laid out in section 2.4. As before, the ACR cohort data was used as a validation data set by deriving true and observed albuminuria status. To define the observed test result, I used the three test algorithm and gender specific thresholds. The true status was estimated as the average of all four measurements. As the purpose of this validation is to gauge the models ability to predict proportions of false positive and false negative rates the conditional probabilities (sensitivity and specificity) were also compared to the model predictions. This was done to adjust for different prevalence's we might expect between people with type 1 diabetes and type 2 diabetes at the same duration.

Regression to the mean

The regression to mean method of validation cannot be done for the logistic model using the ACR cohort data as there are not enough repeat measurements. Ideally, $6-9$ measurements would be required to provide 3 results for the first assessment and another three for the second and preferably another three to obtain an independent measure of the true status, although an average of the first 6 may be sufficient. Instead, I tested each model by comparing the

amount of regression to normoalbuminuria we could expect in two-year intervals over a six-year follow-up period and compared them to the rates reported in Perkins *et al*[43]. For the flexible variance model, I simulated individuals to have durations of diabetes to match those of the test data. I then simulated trajectories using the model estimates and simulated assessments results of either microalbuminuria or normoalbuminuria at an initial evaluation and then 2, 4 and 6 years later. I then "follow-up" all those who had microalbuminuria at baseline and counted up for each additional time point the numbers and percentage of those with microalbuminuria and normoalbuminuria. As neither the flexible variance model nor the fixed effect model can predict transition from microalbuminuria to proteinuria, I selected out a proportion equivalent to the reference data at each time. Predicted numbers from both models falling into each category were scaled to match the numbers reported in the original data but the total number of simulated patients was $100,000$.

Prevalence method

Estimates of the prevalence of microalbuminuria predicted by the model and those obtained from the literature will also be compared graphically as in chapter 2.

3.5 Results

Fixed effects model

From a total of 4770 assessments on 535 people, 3785 of these assessments had three consecutive measurements of ACR, 720 had two measurements and 265 had just one measurement. Positive or negative persistence could be determined for 4617 assessments, whereas the remaining 153 persistence could not be confirmed or refuted. One person was lost because they did not have any complete assessments, leaving 534 people contributing data with up to 20 years follow up.

Parameter estimates

A logistic regression model was used to fit the fixed effect log ACR model. I used the *lme* function in R with a logistic link function. A warning message

```
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

was given but this did not cause the model to fail. Further investigation showed that this was a problem of separation and these probabilities occurred when the individual traces had either all negative (zero's) or all positive (1's) results. Fitted probabilities were obtained by calling the *fitted values* from the model object. Table 3.4 shows the mean and standard deviation of the intercept and slope terms.

Parameter	Estimate
Mean of logit^{-1} intercept terms	0.087
Standard deviation of logit^{-1} intercepts	0.24
Mean of logit^{-1} slopes	0.84
Standard deviation of logit^{-1} slopes	0.12

Table 3.4: *Parameter estimates from the fixed-effects model*

Diagnostics

The residuals from the logistic regression model were divided into categories of duration of diabetes and assessed for outliers or distinctive patterns. Figure 3.2 shows the mean residual for each year of duration with lines representing the theoretical 95% limits.

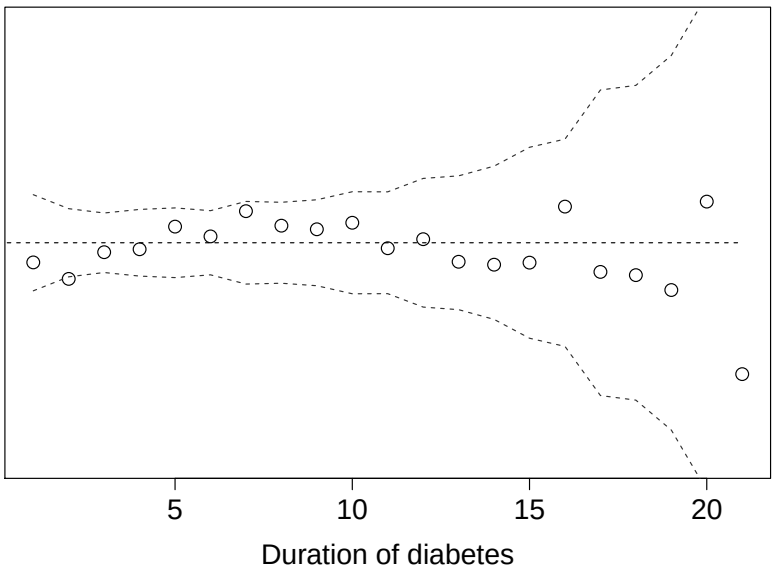


Figure 3.2: *Binned residual plot for the fixed effects model.*

Predictions

The observed and predicted true values were used to create a state vector classifying observations into one of four indices; false or true negative, false or true positive. States were then tabulated by yearly intervals of duration. Table 3.5 shows the proportions in each state from the first year to year 10.

Random effects models

Parameter estimates

Both the increasing variance ($\rho = 0$) and flexible variance ($\rho = \hat{\rho}$) versions of the model were fit using the JAGS function in R. Initial attempts to fit

Duration of Diabetes	True -	False -	False +	True +
1	0.95	0.00	0.02	0.02
2	0.98	0.00	0.01	0.01
3	0.97	0.00	0.03	0.01
4	0.95	0.00	0.02	0.02
5	0.94	N.E	0.04	0.02
6	0.93	N.E	0.03	0.04
7	0.90	N.E	0.05	0.05
8	0.91	0.01	0.04	0.04
9	0.90	0.02	0.04	0.04
10	0.87	0.01	0.05	0.07

Table 3.5: *Estimated proportions of true negative, false negatives, false positives and true positives by duration of diabetes under the fixed effects model. N.E - not estimated*

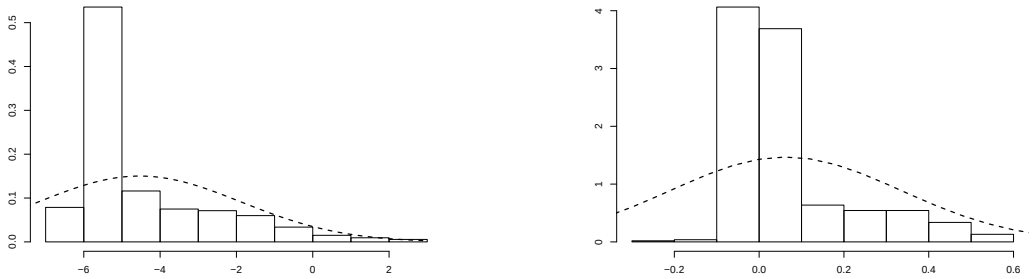
the model resulted in considerable auto-correlated simulations and I decided to centre the time variable by subtracting 7 from each original value to reduce the correlations and aid convergence of the model. This sped up the time it took for the models to converge to a stationary distribution. Parameter estimates for both models are shown in table 3.6.

	Increasing variance		Flexible variance	
Parameter	Mean	SD	Mean	SD
Mean baseline (7 years) logit PMA	-4.71	0.29	-4.52	0.28
Mean increase in logit PMA	0.20	0.03	0.05	0.04
Standard deviation of baseline logit PMA	2.72	0.25	2.65	0.24
Standard deviation of increase logit PMA	0.22	0.04	0.28	0.04
Correlation between baseline and slope	NA	NA	0.66	0.11
Deviance	1330.52	35.16	1330.34	36.19

Table 3.6: *Parameter estimates for in the increasing and flexible variance models. PMA - persistent microalbuminuria*

Diagnostics

The residuals or *random effects* from the flexible variance model were examined for the normality assumption. Figure 3.3 shows histograms of the intercept random effects and the slope term random effects, with the normal distribution curve based upon the point estimate of the mean and standard deviation. The bivariate normal assumption is also assessed by plotting the corresponding



(a) *Random intercepts* **(b)** *Random slopes*

Figure 3.3: *Histograms of the estimated random effects for the increasing variance logistic model (a) the intercepts α_i and (b) the slopes β_i*

intercept slope random effect (figure 3.4). This figure illustrates the positive correlation (i.e. higher baseline values (α_i's) are associated with higher progression values (β_i's).

Predictive validation

The flexible variance model was used to recreate the outcome data. Two hundred replicate data sets were constructed and in each one the number of positive tests for persistent microalbuminuria were counted. The mean of the two hundred replications was then used as the expected rate of persistent microalbuminuria over 20 years of duration. Figure 3.5 shows the observed rates of positive tests and the rate predicted by the model.

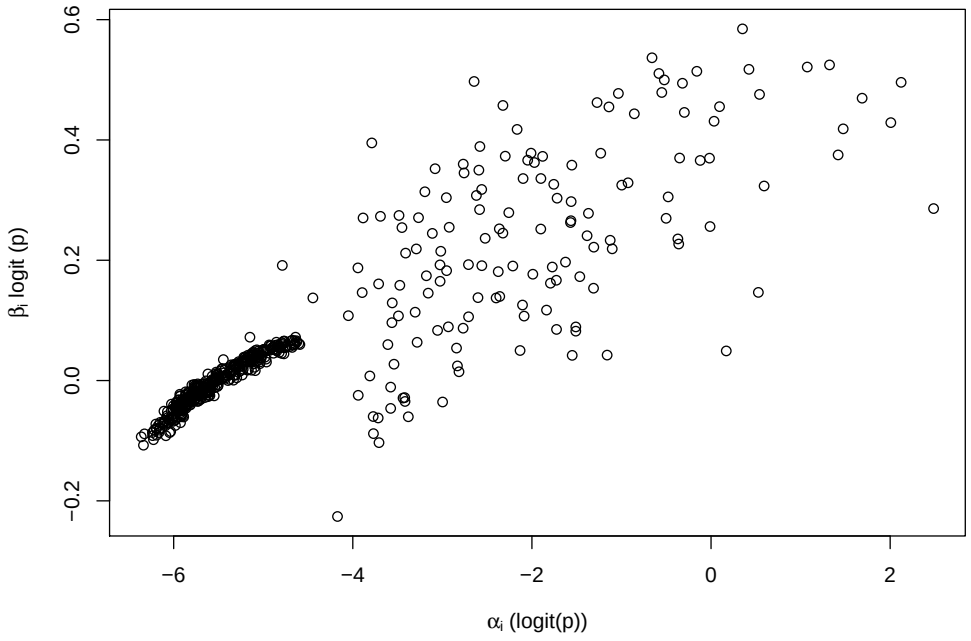


Figure 3.4: *Scatter plot of the individual fitted intercept and slope terms from the flexible variance model*

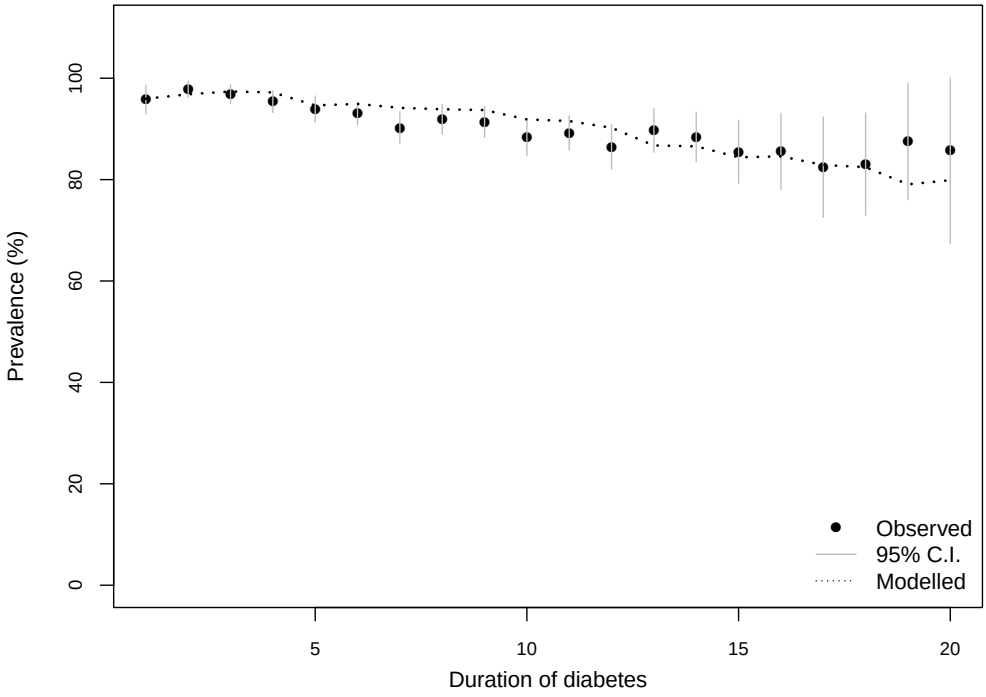


Figure 3.5: *Observed and modelled rates of positive tests for persistent microalbuminuria*

3.6 Evaluation of monitoring prior to initiation of ACEi therapy

The frequency of monitoring ACR was varied from the reference annual interval to intervals of 6 months, and to intervals of 2 and 3 years. The simulation models were run twice, allowing for different models for the progression of persistant microalbuminuria over time. In the first analysis, the estimates from the random effects logistic model were used for the progression model and in the second analysis, the fixed effects estimates were used.

The results of the evaluation of monitoring using the individual patient simulation models (tables 3.7 and 3.8) can be interpreted as follows. The first row gives the average number of visits (but not necessaarily the number of ACR assays) that a person undergoes before a clinical diagnosis of persistent microalbuminuria for four different monitoring strategies. The second row of each table gives the percentage of people who receive a positive diagnosis of persistent microalbuminuria during the course of the monitoring programme and the third row gives the average age at the first positive diagnosis and hence begins ACEi therapy. Row four gives the percentage of false positive tests that were false at the time (a mixture of over diagnosed and premature diagnoses). Row five gives the average length of time between the false positive diagnosis and the time at which the the underlying level of albumin creatinine ratio catches up with the diagnosis. Row six gives the percentage of people who, at some point in the monitoring programme were incorrectly given a negative diagnosis by monitoring. Row seven (average delay time) is the average length between development of albuminuria and detection by monitoring, these figures include those who did not have a false negative result but developed microalbuminuria between assessments (interval cases). Finally, the last row gives the percentage of people who were overdiagnosed by monitoring. Overdiagnosis in this case is defined as a person who has a false positive diagnosis but never goes on to develop true microalbuminuria in their lifetime.

Random effects model

Table 3.7 shows the consequences of monitoring at different intervals in a cohort of people with type 1 diabetes with similar age and gender distributions as those in the ORPS cohort using the random effects model for the progression of persistant microalbuminuria over time.

Outcome	Interval of monitoring ACR prior to ACEi therapy			
	6 months	1 Year	2 Year	3 Year
Average number of visits/assays prior to diagnosis	55	30	16	11
Percentage of people receiving positive diagnosis	50%	49%	47%	45%
Average age when starting ACEi therapy (years)	22	24	26	27
Percentage with false positive diagnosis	50%	43%	35%	29%
Mean (median) advance time in years	10.0(7.0)	8.8(6.0)	7.8(5.1)	7.4(5.0)
Percentage of people with false negative diagnosis	1.3%	2.8%	5.2%	6.3%
Mean number of false negative assessments	1.6	1.4	1.3	1.2
Mean (median) delay time in years	1.2(1.0)	1.9(1.7)	3.5(3.1)	4.8(4.5)
Percentage overdiagnosed	17%	14%	11%	10%

Table 3.7: *Resource and clinical implications of different intervals of monitoring prior to ACEi therapy under the random effects model*

Fixed effects model

Table 3.8 shows the implications of monitoring at different intervals in a cohort of people with type 1 diabetes with similar age and gender distributions as those of the ORPS cohort using the fixed effects model for the progression of persistant microalbuminuria.

Outcome	Interval of monitoring ACR prior to ACEi therapy			
	6 months	1 Year	2 Year	3 Year
Average number of visits/assays prior to diagnosis	78	39	20	14
Percentage of people receiving positive diagnosis	26%	26%	25%	24%
Average age when starting ACEi therapy (years)	19	20	21	22
Percentage with false positive diagnosis	15%	13%	11%	9%
Mean (median) advance time in years	8.7(5.6)	8.1(5.1)	7.5(4.6)	7.5(4.5)
Percentage of people with false negative diagnosis	1.8%	1.8%	2.1%	2.1%
Mean number of False negative assessments	1.2	1.3	1.2	1.2
Mean (median) delay time in years	0.6(0.6)	1.5(1.2)	3.1(2.8)	4.5(4.0)
Percentage overdiagnosed	5%	5%	4%	4%

Table 3.8: *Resource and clinical implications of different intervals of monitoring prior to ACEi therapy under the fixed effect model*

3.7 Clinical Validation

Surrogate method

The first three of the four measurements in the ACR cohort study were used as the "test" measurement and the average of all four formed the surrogate. This differed from the method used for the continuous models because three tests were needed to define persistence. Ninety-four of the 98 had albumin-creatinine ratio at baseline. Using the three test algorithm and gender specific thresholds, the prevalence of persistent microalbuminuria was $10/94 = 10.6\%$. As a sensitivity analysis, I used the second, third and fourth measurements when the first measurement was missing (n = 4). Three of these four had usable data and the fourth had no measurements of ACR at any visit. Using all the data, yielded an estimate of the prevalence as $10/97 = 10.3\%$. The average of all possible measurements was calculated and used as the surrogate true measurement. Estimated prevalence was $13/97 = 13.4\%$. A table cross-classifying the surrogate true and observed results is given in table 3.9. I also

		Three test result		
		$<$ threshold	$\geq$ threshold	Total
Surrogate true	$<$ threshold	84	0	84
	$\geq$ threshold	3	10	13
	Total	87	10	97

Table 3.9: *Cross-classified table of the surrogate true and persistent diagnosis using three tests from the ACR cohort study*

calculated the rates of true and false positives, true and false negatives from the table and compared them to predictions from each of the models (see table 3.10). To adjust for the difference in prevalence between the validation data set (ACR cohort) and derivation data (ORPS) I calculated the conditional probabilities and compared models to the observed rates, see table 3.11.

Index	ACR Cohort	Random Effects	Fixed Effects model
True −	86.6%	94.2%	95%
False −	3.1%	0.7%	N.E
False +	0%	3.5%	3%
True +	10.3%	1.6%	2%

Table 3.10: *Predicted true and false positives, true and false negative probabilities under the random effects and fixed effects model and the surrogate method estimate from the ACR cohort study*

Index	ACR Cohort	Random Effects	Fixed Effects
Sensitivity	76.9%	68%	77.4%
1 - Sensitivity	23.1%	31.7%	22.7%
Specificity	100%	96.4%	96.5%
1 - Specificity	0%	3.6%	3.5%

Table 3.11: *Predicted sensitivity and specificity under the random effects and fixed effects model and the surrogate method estimate from the ACR cohort study*

Regression to the mean method

Table 3.12 shows the results of the regression to the mean method of validation. This was intended to see how the models compared to empirical data on regression/progression over time. The table can be interpreted as follows. In the Perkins paper [43], there were 386 people with micro-albuminuria at the initial evaluation period, and these people were remeasured three more times and categorised as having regressed to normo-albuminuria, remained in the micro-albuminuric range or developed proteinuria. At each time point, the percentages in each albuminuria category for each method should add to 100. Progression to proteinuria was not included for either statistical model as the definition in the original paper was not based upon ACR measurements alone. Table 3.12 shows that the fixed effects model predicts less regression from microalbuminuria than the random effects model. Even though neither model fit well, the fixed effect model might have been expected to make bet-

Status	Model	Initial Evaluation Period	1st Follow-up	2nd Follow-up	3rd Follow-up
Proteinuria	Reference		24 (7)	37 (13)	33 (15)
Micro-albuminuria	Reference	386 (100)	191 (54)	141 (49)	99 (45)
	RE Model		271 (83)	214 (85)	165 (88)
	FE Model		276 (84)	229 (85)	157 (86)
Normo-albuminuria	Reference		136 (39)	110 (38)	88 (40)
	RE Model		56 (17)	37 (15)	22 (12)
	FE Model		51 (16)	39 (15)	30 (14)

Table 3.12: *Predicted regression of microalbuminuria at two year intervals for three follow up periods under the random effects (RE), fixed effects (FE) model and numbers reported by Perkins et al [43]*

ter predictions than the random effects model given the previous validation. However, because neither model permits true regression from microalbuminuria, the only apparent regression arises from measurement error, and as the random effects model has larger estimated measurement error there is more apparent regression of disease.

Prevalence methods

Estimates of prevalence over time were calculated using normal distribution functions. Sensitivity to the assumption of normality was checked by comparing estimates derived from simulation. Prevalence estimates from the fixed effects(FE) model were calculated directly from the ORPS data.

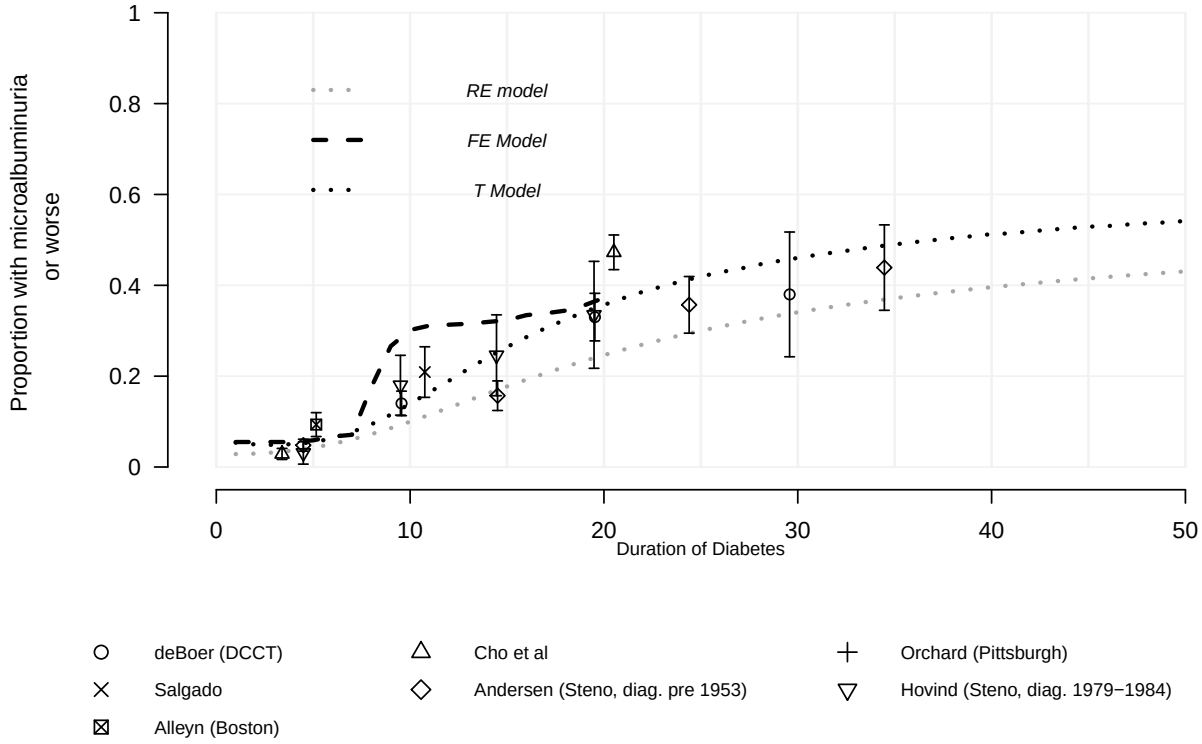


Figure 3.6: *External validation of models for binary ACR. Symbols represent point estimates of prevalence from published studies with error bars representing 95% confidence intervals. Lines represent estimated prevalence rates of persistence micro-albuminuria using model parameters.*

3.8 Conclusion

In this chapter, I set out to test whether the monitoring model could be extended to monitoring tests or assessments that are interpreted as a dichotomy. Monitoring diabetic nephropathy currently involves a strategy of confirmatory testing in order to establish whether a person has persistently raised albumin levels or not. I found that a logistic regression model, if interpreted in the latent format [69] can be used derive estimates of true and false positive and negative tests. By interpreting the outcome in the latent form as the underlying true value the existing framework needed only to be modified slightly in order to evaluate different monitoring strategies.

Statistical validation

Fixed effects models

The fixed effects model can be estimated with standard software and has the least assumptions but is computationally unstable. For people whose sequence of test results is consistently negative (0) or positive (1), the model fitting software will return an error stating that fitted probabilities are numerically 0 or 1. This does not represent a substantial problem if there are only a handful of people with this type of profile, but if they constitute the majority of responses then the model would be rendered useless. I also found that making predictions from this model had two drawbacks, firstly, the false negative rate could not be estimated for durations of 5,6 or 7 years (see table 3.5). Secondly, the estimates of true and false positives and negatives were erratic which I attributed to the relatively small data set. In other words, the method is inefficient. This method may only be useful for data from large RCT's in which the examination times are regular and patients are followed up for many years.

Random effects models

The logistic random effects model is equivalent to the continuous model of chapter 2, with a logit transformation of the outcome. Identifying this, allowed me to utilize all of the methods and techniques already in place. Even though these models do essentially the same job as the fixed effects model (they estimate regression lines for individuals) they do it far more efficiently and robustly. This comes at a price of having to justify assumptions of normality as we did with the continuous versions. Histograms of the fitted random effects (see figure 3.3) placed doubt on the suitability of normality assumption and a scatter plot of the intercept and slope effects, figure 3.4 did not look bivariate normal (ellipsoid) but resembled a mixture of two distinct forms. Even though the assumptions of the model looked untenable, the model recreated the prevalence of microalbuminuria over the first twenty years of duration successfully. This is to be expected given that this model estimates this directly in contrast to the linear model which estimates prevalence indirectly. In conclusion, the random effects model does not meet all of the criteria to be a statistically validated model, but it could still be clinically validated and hence useful in practice [77]

Clinical validation

Surrogate method

I used the ACR cohort data to obtain independent estimates of true and false positives and true and false negative rates in order to compare with the model predictions. An obvious limitation to this analysis is that the validation data comes from an older type 2 population whereas the model is based upon adolescents with type 1 diabetes. This was not ideal and a validation dataset with type 1 diabetes would have been preferred. However, given that I was unable to obtain such a dataset, a validation based on a cohort of people with type 2 diabetes was considered preferable to no validation at all. In addition, although the prevalence of kidney disease may differ between types of diabetes there is

little or no evidence to suggest that short-term and laboratory variation of ACR and hence misclassification error (adjusted for prevalence) should differ between these two cohorts.

In contrast to the continuous models of chapter 2, the total number of positives predicted by both models was half the prevalence observed in the validation cohort(e.g. 5% vs 10%). This is probably due to the fact that the case mix in the validation data is different from the cohort used to the fit the models. The ORPS data consists of adolescents with type 1 diabetes whereas the ACR cohort is older people with type 2 diabetes. Therefore, it is reasonable to expect the prevalence of microalbuminuria to be different.

A fairer comparison, involves comparing the sensitivity and specificity of the model predictions as these statistics adjust for the different prevalence's. Table 3.10 shows that when the prevalence has been taken into account the models are closer to the observed sensitivity and specificity. In particular, the fixed effects model's predictions were very good after adjusting for the prevalence. A limitation to this method of validation is that there was not a truly independent means to estimate the true level. When the same method was applied to the continuous model validation, one ACR measurement was assumed to be the test result and the average of the other three formed the surrogate true. In the dichotomous version, it was not possible to have separate estimates and the true comprised of the mean of all four and the observed comprised of three of the other test results.

Regression to the mean

I also compared predicted regression to the mean and to a study specifically designed to estimate this phenomenon in a large cohort of people with type 1 diabetes. In comparison to the published data, both models under-predicted the proportion regressing to normoalbuminuria at all follow-up visits. The amount of regression to normoalbuminuria observed by Perkins *et al* was close to 50% whereas the models derived from ORPS predicted regression in the

order of $15 - 18\%$. Estimates of regression reported by Schultz *et al* using the same ORPS cohort (but a different model) were in the region 33%, suggesting that the Perkins data may represent an upper limit of the amount of regression to normoalbuminuria.

However, the fact that the predicted amount of regression to the mean was significantly lower than expected suggests a lack of fit in the model for the natural course of disease. The reason for this is that the random effects models do not permit spontaneous regression and assume that people either progress or regress and do not change direction. The short-term variation in the outcome will give the impression of transient regression but the significantly higher rates observed in other studies hints that there may be more regression than that caused by variation alone.

The models used in this chapter suggest that about half to a quarter of what is thought to be regression is nothing more than a product of short-term variability. A model which explicitly takes this regression into account may provide a closer match to the rates observed by Schultz and Perkins studies. However, there are a number of studies that have modelled the progression of nephropathy and have not factored in true regression back from microalbuminuria in their Markov models. For example, neither Borch-Johnsen *et al* [93], Adler *et al* [94] nor Palmer *et al* [61] permitted transitions back from early stages of diabetic nephropathy. However, this can be easily accommodated into a Markov model and tested statistically by comparing models with and without true regression from microalbuminuria.

Prevalence method

I compared estimates from each of the models with the estimates of prevalence from published cohort and cross-sectional studies. These estimates were the same ones used in chapter 2 and thus allows us to compare the logistic model predictions with the continuous model predictions as well as the external data. Figure 3.6 shows the predicted prevalence from the logistic random

effects model as well as the estimates from the continuous T model. The logistic random effects model has a much lower estimate of the prevalence for duration of thirty years onwards than the T model. The estimates of the long-term prevalence from the fixed effects model extend only as far as the data from which they are derived from and closely follow the published estimates of prevalence early on but then climb dramatically thereafter.

Comparing outcomes from simulation models across different models for progression of ACR

I applied the same individual patient simulation model that was used in chapter 2 using models for the binary outcome of persistent microalbuminuria. Outcomes of this simulation using the logistic model are broadly comparable to outcomes from simulation using the models for continuous data. The logistic version of the model predicts that fewer (around 10%) people will be picked up as positive for persistent microalbuminuria when compared to the simulation using the continuous ACR models. The simulation using the logistic model predicts that 50% of everyone monitored will test positive at some point for intervals of 6 months or 1 year. The simulation also predicts that less of these observed positives will be false positives if monitored every 6 months (50% vs 57%) compared to the same simulation using the continuous models. The reduction in the false positive rate as the interval of monitoring increases is less marked with the logistic model than the continuous model and estimates are higher. At a 3 year interval the logistic model produces a false positive rate of 29% in the continuous flexible variance model it is 26%. As the logistic model has a slower increase in the prevalence rate for the first decade of diabetes duration (see figure 3.6) this is reflected in the average number of tests, average age starting ACEi therapy and advance time under the logistic model. People are false positive earlier and their progression is slower under this model. Two noticeable differences between the two models are the rates of overdiagnosis and false negatives. The logistic model simulation predicts a much lower rate

of false negatives than the continuous version although the average delay time is similar. The percentage overdiagnosed is 14% with annual monitoring under the logistic model and 19% with the continuous flexible variance model.

Discussion

Interest in monitoring for microalbuminuria began after early studies reported that the risk of overt proteinuria was between $60 - 85\%$ within 6- 14 years [43] and the discovery of reno-protective action of drugs in people with type 2 diabetes [60]. Annual monitoring of renal function is recommended by many guideline bodies but evidence for its effectiveness is limited. The work presented here suggests that because of the high-levels of short-term variability in albumin-creatinine ratio, even when taking into account the three test strategy, frequent monitoring is likely to be cost-ineffective. Frequent monitoring (6 monthly-annual) incurs more operational costs; more assays, more people starting pharmacotherapy and earlier on in life. Frequent monitoring leads to many false positives, some of which represent early detection of slowly deteriorating renal function, but others represent over-diagnosis. As there are actual physical harms associated with taking ACE inhibitors or ARB's, strategies that minimise false positive tests and over-diagnosis yet still detect those at risk of proteinuria would be highly desirable.

Comparison with the literature

The rate of progression of ACR has a large impact on the cost-effectiveness of annual and biennial screening. In the HTA monograph in which the continuous T model was used to evaluate monitoring ACR in a full health economic analysis, we reported that when the average rate of progression was set to the lower bound of the 95% C.I. (2% per year) fewer people progressed to end-stage renal disease (ESRD) and this reduced costs and QALYs [76]. As a consequence, it was questioned whether annual or biennial screening would be cost-effective if

the progression of nephropathy was at this rate. The simulation using the logistic model estimates a consistently lower prevalence of microalbuminuria than the model used in their HTA analysis and so a relevant question that warrants further investigation is, how robust is the cost-effectiveness of annual monitoring to changes to the statistical model underpinning the health economics model

Even if annual monitoring is cost-effective there may be alternative monitoring strategies that are more effective. Previous studies have suggested that the frequency of monitoring could be governed by the initial level of albumin excretion [95], with people who exhibit low levels of albumin excretion (ACR < 2 mg/mmol) in a single overnight collection not needing to be rescreened more frequently than every 18 months or 2 years. People perceived at much higher risk (ACR > 10 mg/mmol) could be called back every 2 months with a view to expediting treatment. Decisions about monitoring may also be enhanced by incorporating additional information. In renal monitoring, glycaemic control is associated with progression of nephropathy and this could help to identify patients at higher risk. A prediction model based on progression of ACR and HbA1c developed from the ORPS cohort and validated in an independent cohort suggested that combining serial measurements of albumin excretion and HbA1c can identify those at higher risk of developing clinical proteinuria prior to the development of microalbuminuria [40]. A monitoring strategy based on phases of duration of diabetes may also reduce false positive tests and overdiagnosis. This strategy could comprise of annual monitoring for the first three years since diagnosis, then a longer interval (2- 3 years) from 3 to 15 years duration and finally much less frequent monitoring thereafter (every three to five years) or complete cessation. If the prognosis is different for people who are incident early then for people incident later then this strategy could be highly effective. When Warram and colleagues attempted to recreate the prevalence data from a prospective cohort study, they noted that patients who developed microalbuminuria during the first 10 years of post pubertal insulin dependent

diabetes (IDD) must, in all cases, progress to overt proteinuria and without this characteristic they could not generate the observed prevalence of proteinuria observed at 20 years of duration [57]. They also noted that patients who develop microalbuminuria after the first decade of diabetes, on the other hand, have a different prognosis. A similar conclusion was drawn from a prospective cohort study in people with type 1 of long duration of diabetes (> 15 years) by Forsblom *et al*[96]. The authors reported that among their initially normoalbuminuric patients, those who developed micro or macroalbuminuria had a lower average duration of diabetes suggesting that as duration increases the risk decreases. The people who remained in the normoalbuminuric range had mean duration 5 years higher than those who progressed (23 vs 18, $p = 0.02$). This would suggest that there may be benefits to be had from adaptive monitoring strategies in which, those who develop microalbuminuria in the first decade since diagnosis of diabetes are watched and monitored annually, whereas people with consistently low levels of albumin excretion are moved to longer intervals of monitoring with potential for cessation of monitoring under certain conditions.

Limitations

The cohort used to model progression of ACR continued for 20 years, but had very few people who went on to develop clinical proteinuria and none who developed end stage renal disease. In this thesis, I have shown that monitoring frequently produces many false positives and increases operational costs but long-term implications from different monitoring strategies could not be assessed and remains unclear. The model for progression did not account for true regression of microalbuminuria, an accepted part of the pathophysiology of kidney disease. Assuming that progression is uni-directional will most likely effect evaluations of monitoring but cannot be addressed in this analysis. Finally, I have only addressed strategies that vary by interval and have not explored

strategies that change the interval of monitoring for different profiles of systemic risk (HbA1c, blood pressure) or chronicity of measurement.

Summary

There is a need to continue research into the effectiveness of monitoring for microalbuminuria, particularly in people with type 1 diabetes. The models described here will be useful for modelling monitoring strategies that use confirmatory testing schemes to determine persistent microalbuminuria taking into account short and long-term variability.

Chapter 4

Models for ordinal data: Diabetic retinopathy

Many chronic diseases are interpreted in terms of levels of severity and advancing disease and hence outcomes from monitoring tests are often ordinal in nature. To describe the progression of such diseases over time, multistate Markov models are often used. Markov models assume that future prognoses depends only on the current state and not on the history and this may unrealistic for some diseases. In this chapter I describe an extension to the logistic model developed for describing the risk of persistent microalbuminuria and misclassification in chapter 3. The ordinal version of the model is applied to repeated screening of diabetic eye disease (retinopathy). As the progression through the stages of disease in this model is not independent, this represents an alternative to the Markov based methods. The model is then used to assess the properties of the monitoring test and evaluate the impact of changing the interval of screening for different baseline risk profiles.

4.1 Introduction

Diabetic retinopathy (DR) is a long-term complication of diabetes affecting around a third of all people with diabetes [97]. Retinopathy refers to the pathology of the capillaries, arterioles and venules in the retina and the subsequent effects of leakage from or occlusion of the small vessels [98]. Severe visual

loss is due to DR in 86% of all patients with type 1 diabetes but only 33% of type 2 diabetes patients because of the competing risk of vision loss due to age related macular degeneration, cataract and glaucoma [99]. The majority of blindness (severe to total loss of vision) in people with diabetes results from either proliferative DR or diabetic macular oedema [100]. Treatment with laser therapy (photocoagulation) has been shown to prevent severe visual loss in eyes with proliferative retinopathy and reduce the risk of visual loss in eyes with clinically significant macular oedema [101]. Photocoagulation has also been shown to reduce the risk of progression to high-risk proliferative retinopathy by 50% [97] in people with non-proliferative disease. As it is uncommon for vision to be restored using photocoagulation, early intervention is important.

The aim of screening is to reduce the risk of vision loss with prompt identification and effective treatment of sight-threatening diabetic retinopathy (STDR) [101]. Screening for eye disease is routinely done by inspecting the retina for physiological abnormalities (lesions) of which the type, location and proliferation indicate a heightened risk of future vision impairment and the need for treatment. There is widespread agreement that this is best done using digital fundus photography and although alternatives have been trialled including direct ophthalmoscopy or optometrists slit-lamp biomicroscopy they both lack sensitivity for diabetic retinopathy.

Other retinopathy outcomes, as opposed to the physiological changes to the retina assessed by photography have been investigated but are rare. Monitoring of visual acuity by fundoscopy was assessed by Moore and colleagues and they concluded that although it was not specific for STDR but it may be useful for ruling out maculopathy [102]. Other outcomes have been cited as possible markers. Klein *et al.* investigated the relationship of venal vessel caliber with the incidence of diabetic retinopathy and found that it was independently predictive of proliferative retinopathy but not the incidence of less advanced retinopathy [41].

Classification of diabetic retinopathy

Systematic evaluation of the interior surface of the eye (fundus) is based on the class, number, location and severity of the lesions. This information is converted into scales or grades that range from absence of any disease to more advanced forms of diabetic retinopathy. The best validated grading system for diabetic retinopathy in terms of its outcome-related evidence base, is the system developed for the Early Treatment of Diabetic Retinopathy Study (ETDRS). The ETDRS scale has eleven main levels of severity and further subdivision within levels 43 - 53. However, seven-field stereoscopic photographs each covering 30 degrees are required for this gold standard grading system. An alternative simpler system was developed for the EURODIAB complications study and was based on the principle of the ETDRS. Using two-field (macular and disc centred) 45^0 photography the system is able to classify retinopathy severity into 6 ordinal categories. The simplified two-field system was considered to be in good agreement with the gold standard (ETDRS) and it was suggested that it would be useful for large epidemiological studies or multi-centre clinical trials [103]. As grading systems and photographic protocols used for monitoring and screening need to be simplified, two-field is now recommended by English National Screening Programme for Diabetic Retinopathy (ENSPDR). The ENSPDR uses a grading system similar to the EURODIAB system but with 4 levels in place of the 5 used in the EURODIAB. Table 4.1 compares the ETDRS and ENSPDR systems, the probability of progression and current screening intervals.

Sources of variation in screening for diabetic retinopathy

A single fundus photograph is subject to numerous sources of random and systematic technical variation. I have identified a number of potential sources of measurement variation, both intra- and inter-grader and variation due to the delivery of screening. These include, but are not limited to who is grading

Lesions	EDTRS Severity Scale	ENSP levels	One-year risk of progression to proliferative DR	Monitoring intervals
DR absent or questionable	No apparent retinopathy (10-15)	No detectable DR (R0)	-	Annually
Microaneurysms, retinal haemorrhages, exudates or venous loops	Mild NPDR (20-35e)	Background DR (R1)	Level 30 = 6% Level 41 = 11.3%	ENSP - Annually, ETDRS: 4-6 months
Venous beading or reduplication, IRMA multiple deep, round or blot haemorrhages	Moderate-Severe NPDR (43a - 53d)	Pre-proliferative (R2)	Level 45 = 20.7%	ENSP - Refer to ophthalmologist, ETDRS - 3 or 4 months
NVD or NVE, preretinal or vitreous haemorrhage	Mild PDR (61a-b)	Proliferative DR (R3)	-	-

Table 4.1: *Classification of diabetic retinopathy according to the Early Treatment of Diabetic Retinopathy Study (ETDRS) classification and the English national screening programme for diabetic retinopathy (ENSPDR), CWS, Cotton wool spots; HE, hard exudates; IRMA, intra retinal microvascular abnormalities; MA, microaneurysms; NVD, new vessels on disc; NVE, new vessels elsewhere, RH, retinal haemorrhage; VB, venous beading; VH, vitreous haemorrhage, modified from [99]*

the photographs [104], the number of fields photographed [105] and the use of mydriasis (dilation)[98]. Evidence of the intra-reader variation in grading photographs is limited but may be important. In one study, an ophthalmologist with some experience in grading fundus photographs was given 143 photographs and asked to rate them as absent, present, moderate or severe twice in a randomised order. On the first pass, the specialist rated 55 as absent for microaneurysms and rated 53 of these absent in round two and the remaining two as minimal but present. Table 4.2 shows the results for the agreement for the other grades of the photographs. There is notable disagreement be-

		1st reading			
		absent	minimal	moderate	severe
2nd reading	absent	53	1	0	0
	minimal	2	18	11	0
	moderate	0	4	44	0
	severe	0	0	9	61

Table 4.2: *Frequency of agreement between first and second reading of 148 fundus photographs for presence and severity of haemorrhages and/or microaneurysms*

tween the proximal grades of retinopathy but no misclassification with more distal grades. This is of course only one reader being assessed and gives us no indication as to whether this reader is average, worse than average or better.

Inter-reader variation

We might expect the inter-reader variability to be larger than the intra-variability. In one example of such a study, three retina specialists and three general ophthalmologists graded 400 digital fundus images from diabetic patients in a community hospital. They assessed the images independently and then together two months later to form a consensus grade. Table 4.3 shows the degree of variation between the retina specialists and ophthalmologists. This study also examined the variation between the consensus grade and three general ophthalmologists, three ophthalmic photographers and two nurses. The variation between the other specialists and the degree that anyone individual differed from

	Consensus of retinal specialists	Retina specialists			Std. deviation
		1	2	3	
Severity					
No retinopathy	291 (73)	252 (63)	285 (71)	311 (78)	**7.4**
Mild NPDR	38 (10)	30 (10)	30 (8)	15 (4)	**2.9**
Moderate NPDR	48 (12)	82 (21)	54 (14)	59 (15)	**3.7**
Severe NPDR	12 (1.5)	7 (2)	11 (3)	3 (1)	**1.3**
PDR	5 (3)	10 (3)	11 (3)	10 (3)	**0.2**
Ungradeable	4 (1)	11 (3)	7 (2)	2 (1)	**1.2**

Table 4.3: *Categorisation of 400 single field images by three retina specialists, adapted from Ruamviboonsuk et al, 2005*

the consensus grade was markedly larger than the retinal specialists [106]. In a study detailing the reasons for disagreements between a screening service and hospital biomicroscopy the authors reported moderate agreement ($\kappa = 0.60$) for diabetic retinopathy and $\kappa = 0.42$ for maculopathy [107]. Doctors in the hospital eye service tended to under-grade the severity of photographs and the majority of errors made by the diabetic retinopathy screening service were over-gradings.

Evaluating the effectiveness of screening

In *Principles and practice of screening*, Wilson and Junger [37] established the criteria still used widely today to appraise the viability and effectiveness of screening for disease. Scanlon [101] used this criteria to evaluate screening for diabetic retinopathy and concluded that there was good evidence to fulfil the Wilson and Junger criteria. Similar criteria was described by Mant [39] for monitoring tests. The consist of four items;

- Validity - How well does the test measure or predict clinically significant events?

- Responsiveness - How quickly does the test respond to changes in the condition or treatment?
- High signal-to-noise ratio (SNR) - Is a single test result easy to interpret or is subject to substantial random variation that makes interpretation difficult
- Practicality - Is the test simple, non-invasive and affordable?

Although screening and monitoring may meet each of the criteria by finding evidence for each element in isolation, putting screening and monitoring into practice can be very different. In particular, the timing and frequency of monitoring can have profound effects on each aspect in different ways. For example, valid screening tests will identify mostly those with the condition (they are specific) and not fail to detect many with the condition (they are sensitive). However, the timing and frequency of screening can alter both the sensitivity and specificity and so we should consider not only the validity of the test but the test in conjunction with the strategy. Similarly, a good screening or monitoring test may be able to detect pathological changes early enough to be beneficial and meet the *responsiveness* criteria but unless it is used at the appropriate times, the responsiveness will be diminished. Lastly, although it is well known that SNR can be directly affected by taking repeated measures (for example, with blood pressure) it is less well appreciated that timing can change the relationship between noise and signal with longer intervals allowing for signal to develop and be observed over the noise [4].

In the monitoring model, the test errors (equivalent to sensitivity and specificity) are not assumed constant but change as a function of the underlying progression of disease (as measured by either a surrogate marker or more proximal test). In other words, the test is more likely to be wrong when the true value is close to the threshold and less likely if the true value is far from the threshold [34]. Just as a tumour is more likely to be missed when it is barely visible

to the naked eye but unlikely to be missed once it is sufficiently large to be palpated, tests are unlikely to be false positive when the true value is far from the threshold. As the monitoring model assumes that the short-term variation is symmetric around the true value, the potential for incorrect classification is largest when the true value is on or close to the threshold.

In the previous two chapters, longer intervals of monitoring and were associated with better sensitivity and better specificity. However, longer intervals will inevitably mean delays in detecting clinically relevant change in some people. When this is true, a comprise between false positives and false negative rates will be sought. One such comprise was suggested by Lincoln and Weiss [26]. They argued that intervals for cancer screening could be devised under a constraint that no more than a fraction (ϵ) of people should have a delay in detection longer than T years. They did not consider tests that have imperfect detection, but they found that early on (in relation to age) intervals could be longer to allow for the lower incidence rate (low signal), but found that intervals quickly settled down to T as incidence or signal increased. In these early days of screening for cancer, false positives were considered unimportant and they would soon be corrected before any treatment was started and overdiagnosis was rarely discussed. This method could be useful for setting intervals of screening for DR where the harm of a false negative test far outweighs the harm of a false positive. It is still desirable to minimise false positives and hence we should find the maximum interval which means that more than a fraction of people will experience a delay of T years. This does not address the problem of overdiagnosis, which may be increased by such strategies. Retinal photocoagulation can itself reduce visual acuity and visual fields, reduce near and night vision and steroidal treatment for macular oedema can increase intra-occular pressure (IOP) and progression of cataract [108]

Aims and objectives

In this chapter, I aimed to expand the existing methodology to ordinal outcome data using diabetic retinopathy screening as a case study. The model estimated progression and variability of DR over time (on a latent scale) and the likely errors from screening at defined intervals. In addition, I set out to evaluate the properties of different screening strategies and their effects on two criteria; validity (sensitivity and specificity) and responsiveness. To check whether the model made realistic prediction I compared the estimates from the model with equivalent estimates from published studies.

4.2 Data

As part of a Health Technology Assessment (HTA) project on the optimal screening interval for diabetic retinopathy led by Dr Peter Scanlon, I had access to a large data set from the Gloucestershire screening programme for diabetic retinopathy. In April of 2009, and for the National General Practice to Diabetic Retinopathy Screening (GP2DRS) Phase 1 Pilot in Gloucestershire, patients aged 12 and over were identified as having diabetes from 54 GP Practices for diabetic retinopathy screening. The raw data set contained grading data from 14,810 patients. 1,036 had either the duration of diabetes a baseline HbA1c, type of diabetes or baseline serum cholesterol missing and were thereafter excluded from the main analyses. Of the remaining 13,774 patients there were 60,537 observed retinopathy grade constituting 72,533 person-years of follow-up. Thirty seven people had at least one duplicate screening date and these 43 observations were removed from the data set. Twelve people only attended one screening round and were excluded as they could not be included in the model. Risk factor data was available for 12,003 people as 2,266 people had at least one covariate missing. Table 4.4 shows the baseline characteristics of the GP2DRS data.

Variable		N(%)
Gender	Male	7834 (57%)
	Female	5950 (43%)
Type of diabetes	1	873 (6%)
	2	12910 (94%)
	Not known	1 (0%)
HbA1c (mmol/mol)		50 (43,61)
Serum cholesterol (dl/L)		4.3 (3.7, 5.0)
Age at first screen		65 (56, 72)
Duration of diabetes at first screen		2.3 (0.7,6.8)
Grade of retinopathy in worse eye	No dectable DR (R0)	8460 (61%)
	Background DR (R1)	4739 (34%)
	Pre-proliferative DR (R2)	585 (4%)
Maculopathy in worse eye	None (M0)	13041 (95%)
	Any (M1)	743 (5%)

Table 4.4: *Baseline characteristics of the modelling cohort taken from the Gloucester screening data*

4.3 Methods

In this section I describe how the progression and variation of diabetic retinopathy was modelled over time using a hierarchical ordinal model. The model was applied and used to evaluate screening and monitoring strategies under different criteria. The sensitivity, detection time and positive predictive values were compared under strategies with different intervals.

Introduction

In the previous chapter I described how the latent variable interpretation of the logistic regression model could be used to estimate the properties of binary monitoring tests. The model used in this chapter for retinopathy is based on a generalisation of this method to k levels where $k > 2$. Using the same latent variable interpretation we can imagine that the increasing severity of retinopathy is driven by a continuous latent measure z which can be divided up by cutpoints to form ordered grades of retinopathy. To help with the interpretation we might imagine that the latent variable is some (unknown as yet) biochemical marker of retinal microvascular dysfunction, much in the same way as the albumin-creatinine ratio underlies the categories of nephropathy (normo, micro and macroalbuminuria). Alternatively, the latent variable z could represent some function of the total number of microaneurysms, haemorrhages and other functional abnormalities.

In the logistic model, there was one cutpoint fixed at zero ($c_1 = 0$) which meant that if the latent variable z was less than 0, the true state (median result) was normoalbuminuria and if z was greater than 0, the true state was persistent microalbuminuria. The ordinal model for DR extends this idea to $k = 4$ ordered levels of retinopathy: 1) No detectable retinopathy, 2) background retinopathy 3) pre-proliferative retinopathy and 4) proliferative retinopathy. The true

disease categories are related to the latent variable z using the following rules;

$$\text{(true disease)} = \begin{cases} \text{No DR (R0)} & \text{if } z < c1 \\ \text{Background (R1)} & \text{if } z \in \{c1, c2\} \\ \text{Pre-proliferative (R2)} & \text{if } z \in \{c2, c3\} \\ \text{Proliferative (R3)} & \text{if } z > c3 \end{cases} \tag{4.1}$$

where the $k - 1 = 3$ cutpoints $c1, c2, c3$ are estimated along with the usual parameters of the random effects model. The cutpoints do not have a direct interpretation but determine the short-term variation or measurement error. The model for the latent variable z has the same structure as the random effects models used in previous chapters 2 and 3; the variable z is a linear function of time, with subject-specific parameters for the baseline and progression of disease. The model for the ith person is equivalent to;

$$z_i = \alpha + \beta t + \epsilon_i \tag{4.2}$$

where the error terms ϵ_i have a logistic distribution ($logis(0, 1)$). As before, the error term is not included in model estimation. A graphical representation of the general model for K levels is shown in figure 4.1.

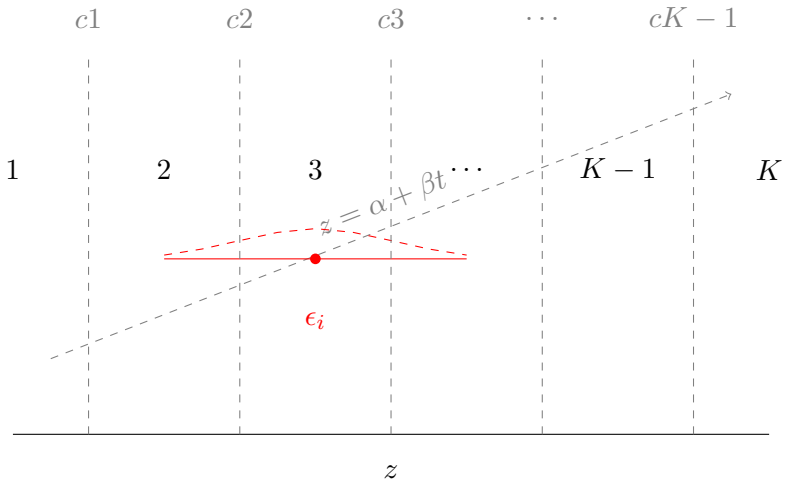
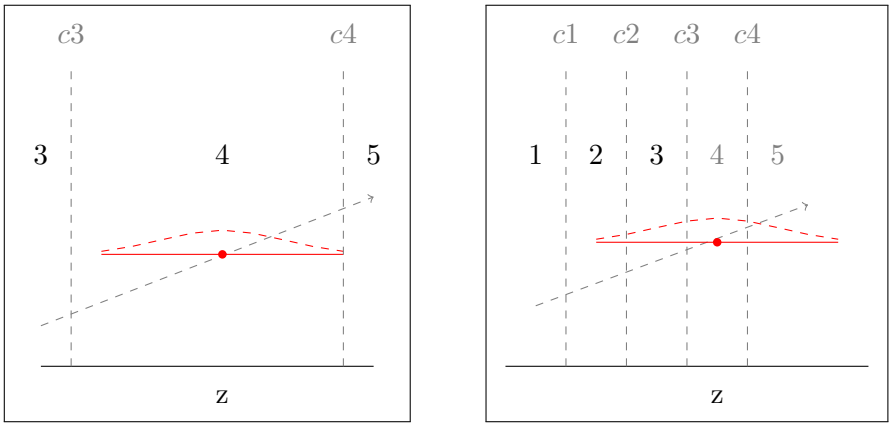


Figure 4.1: *Graphical representation of the latent variable model. Grey dotted line represents the value of the latent variable z_t, the red bell curve as the logistic distribution of the error terms ϵ_i, Numbers from 1 to K are the observed categories and $c_1 - c_k$ the cutpoint parameters.*

Identifiability of the error term

As in the logistic version of the model, the error term has to be fixed and cannot be estimated due to the non-identifiability. Again, this appears as a major restriction as the short-term noise parameter is arguably the most important. The noisiness of the measurement is captured instead by the distance between the estimated cutpoints. Cutpoints spaced far apart represent measurements with good reproducibility and low short-term variation and closely spaced cutpoints represent measures with high variability and poor reproducibility. To help convey this idea I have created a graphic of the ordinal model. Figure 4.2 shows two different models for two different biomarkers, one with small variability and one with large variability, but both with the small error scale. In the model for the biomarker with small variability, the cutpoints are spaced



(a) *Small variability* **(b)** *Large variability*

Figure 4.2: *Representation of two latent models with same size error term (ϵ) but for two biomarker (a) small variability (b) large variability*

far apart in relation to the scale of the latent variable z. In the second model for the biomarker with high variability, the cutpoints are much closer together for the same scaled z.

The hierarchical ordinal model

The four state model for ordered grades of retinopathy is equivalent to a "unidirectional" state model. This implies that a persons' true underlying disease

can only move in one direction as people are assumed to have either a positive or negative gradient. It also assumes that each person passes through each stage in turn, although this does not preclude this happening very quickly.

Estimation of the model

The model was estimated using Monte Carlo Markov Chains (MCMC) routines from the JAGS [75] package in R [72]. The code to fit the model assumes that the data that is arranged in a long format, as in table 4.5. The progression of

id	t	Duration	Grade	HbA1c	Mac. in 1 eye	Mac. in 2 eyes
1	0.00	7.3	1	58	0	0
1	1.21	7.3	1	58	0	0
1	2.51	7.3	2	58	0	0
1	3.52	7.3	2	58	1	0
1	4.59	7.3	2	58	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 4.5: *Example of the formatting of the data required to fit ordinal model for DR. Mac - maculopathy*

the latent variable z is coded into the model as a regression equation adjusting for maculopathy, baseline HbA1c and duration of diabetes. The maculopathy variable takes the value $0, 1$ or 2 and is coded with dummy (boolean) variables. The baseline parameter alpha has mean zero as the cutpoint parameters act as intercept terms. The intercept and slope terms are not correlated and hence this model is of the "increasing variance" form. The interpretation of alpha is a systematic difference from the average in baseline retinopathy level equivalent to adjusting for the level of disease at the first measure and beta is the annual rate of progression of disease between people. As with all JAGS or BUGS models the parameters being estimated are required to have "priors" or probability distributions. As in previous chapters these are specified to be as vague or diffuse as possible. The full code is given in the section B.4 of Appendix B. As with the other random effects models in this chapter, model

checking consists of checking distributional assumptions of the random effects. These will be assessed by histogram plots. In addition, the model will be tested by simulating new outcome data for the cohort used for estimation. The mean expected prevalence's will then be compared to the original and inspected for lack of fit across time and for different grades of retinopathy.

Evaluating screening strategies

I used three criteria to evaluate different strategies for screening and monitoring. I assessed sensitivity, and specificity to check validity of screening for sight-threatening diabetic retinopathy (STDR) and then monitoring for progression from pre-proliferative retinopathy to proliferative disease and the responsiveness of screening for STDR under different intervals of testing.

Sensitivity

To assess sensitivity of different screening and monitoring strategies I used the ordinal model to generate cohorts of people who had no detectable retinopathy at baseline but had progressive retinopathy (positive slope) and then simulated screening visits at different intervals. I then calculated the sensitivity (true positive/true positives + false negatives) for an interval of one, two,three and five years. I did this for three cohorts with low, medium and high risk, where risk was based upon baseline HbA1c, time since diagnosis of diabetes and the presence of maculopathy. In a second analysis, I used the same framework but for cohorts in whom pre-proliferative retinopathy had been detected and were being monitored for progression to proliferative disease.

Specificity

To assess specificity, I use the same cohorts and simulations as for assessing the sensitivity of the test, but calculated the number of false positives, the number of true positives and the ratio of true positives to true + false positives (positive predictive value or PPV).

Responsiveness

To assess the responsiveness of screening for STDR at different intervals I used the model to estimate how long it takes on average for screening to detect true STDR. To do this, I generated a cohort of people with no detectable retinopathy at baseline but who had progressive retinopathy (positive slope) and estimated 1) the time at which the latent variable was equal or exceeded the referral grade of retinopathy and then by simulation calculated the time it took for screening to detect it. The mean and median response time were calculated and as well as the proportions with response time greater than 1, 2 and 3 years.

Validation of the models

Residual plots

Random effect residuals (the differences in individual baseline and slope values from the mean) and the measurement error residuals (difference between the fitted line and the observed) were assessed for normality using histogram plots.

Prevalence plots

In order to check for over-estimation of prevalence I compared observed prevalence in the data with that predicted by the model. This was done by creating a column of simulated observed grades and plotting the proportions of these predicted grades against the observed proportions in years since baseline.

State change or *Bureau* plots

A final and more rigorous check of the model can be made by comparing the time from being observed in one grade of retinopathy until another grade is observed. This is a generalisation of a plot first described by Bureau *et al* for comparing empirical estimates of the waiting times between two states of a model [109]. This method involves calculating the time between observing grade x until grade y is observed for each possible grades or states in the model. When there is no occurrence of grade y following x then the last observed time

is taken and censored in the same way as a survival analysis. Subjects can in effect contribute more than once if they are observed more than once in grade x. As with the survival until referable DR, the event observations are also interval censored and hence this is adjusted for when calculating the product limit estimators by using the parametric functions from the *survival* package [110] in R [72].

External validation

External validation of the model was carried out by comparing predicted errors with external estimates of sensitivity and specificity from diagnostic accuracy studies and studies regrading screening data. Studies were considered for inclusion if they had more than 100 patients and used two-field imaging as the comparator test. In each case, the model predictions of sensitivity and specificity were adjusted so that the prevalence of DR matched that of the prevalence reported in the study.

4.4 Results

Model estimation

Table 4.6 shows the parameter estimates from the fitted ordinal model. The

Parameter	Estimate	SD	lower 95% C.I	upper 95% C.I
sigma.a	2.15	0.03	2.09	2.20
mu.b	0.09	0.01	0.07	0.10
sigma.b	0.29	0.01	0.27	0.31
c1	0.35	0.03	0.30	0.40
c2	6.26	0.06	6.15	6.38
c3	9.72	0.10	9.54	9.91
gamma	0.15	0.00	0.15	0.16
lambda	2.41	0.07	2.28	2.53
delta	3.44	0.10	3.25	3.63
kappa	0.03	0.00	0.03	0.03

Table 4.6: *Parameter estimates from the increasing variance ordinal model for diabetic retinopathy*

parameters of the model do not have a direct interpretation, but are given for completeness. To aid interpretation of the predictors (HbA1c, duration of diabetes and maculopathy) on the progression of DR I computed the cumulative incidence over time for different levels of the covariates. Table 4.7 compares cumulative incidence of referable retinopathy for the base case (covariates set at their means) and then the cumulative incidence with covariates raised while others are fixed.

Transition and error matrices

To further aid interpretation of the fitted model, I calculated the probability of "transitioning" from each grade of retinopathy and made them into transition probability matrices or (TPM). Table 4.8 shows each transition probability as a matrix. I did the same for the probabilities of being observed in one grade conditioning on the true state, for each possible grade of DR. Table 4.9 shows the ordinal model "error" matrix. Table 5.12 shows the error matrix which

Years	Reference case	+ 5 years duration	+ 6 mmol/mol HbA1c	One eye with maculopathy	Both eyes with maculopathy
1	0.00	0.00	0.00	0.00	0.04
2	0.00	0.00	0.00	0.00	0.11
3	0.00	0.00	0.00	0.01	0.30
4	0.01	0.01	0.01	0.06	0.75
5	0.03	0.06	0.04	0.20	1.60
6	0.11	0.19	0.13	0.53	2.92
7	0.30	0.47	0.33	1.12	4.67

Table 4.7: *Predicted cumulative incidence (%) of true referable retinopathy after initial screened showed no DR*

from	*to*			
	No DR	Background DR	Pre-proliferative DR	Proliferative DR
No DR	0.95	0.05	0	0
Background DR	0.03	0.97	0.001	0
Pre-proliferative DR	0	0.09	0.9	0.001
Proliferative Retinopathy DR	0	0	0.13	0.87

Table 4.8: *Transition probability matrix (1 year) for the hierarchical ordinal model*

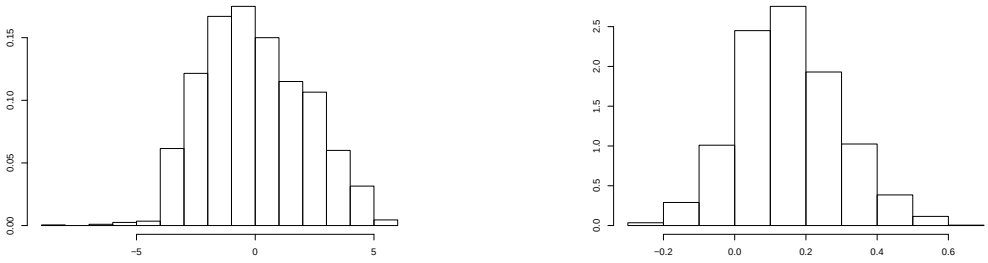
can be interpreted as the probability of observing the column state given the true state is equal to the row state. For example, it is estimated that there is a probability of 1/5 of being graded by screening as having background DR when in fact the true state is no detectable retinopathy.

Diagnostics

The assumption of normality for the random effects is checked using histogram plots with a normal density with mean and standard deviations taken from the parameter estimates (see table 4.6). Figure 4.3 shows histograms for the estimated random effects.

true state	*observed state*			
	No DR	Background DR	Pre-Proliferative DR	Proliferative DR
No DR	0.81	0.19	0.00	0.00
Background DR	0.22	0.76	0.02	0.00
Pre-Proliferative DR	0.00	0.36	0.59	0.05
Proliferative DR	0.00	0.01	0.39	0.60

Table 4.9: *Error matrix for the ordinal model. Estimates represent probability of observing state given the true state (95%) C.I. at baseline.* 0^* *indicates* < 0.001



(a) *Random intercepts* **(b)** *Random slopes*

Figure 4.3: *Histograms of the estimated random effects for the ordinal model (a) the intercept* α_i *and (b) the slopes* β_i

Prediction and internal validation

Comparison of survival time

A minimum requirement of the model is for the modelled survival to closely follow the observed survival in the data. Figure 4.4 plots the Kaplan-Meier estimate derived using the fitted model and the equivalent estimates using the observed data. It shows that that modelled survival from referable level DR does nor correspond to the observed survival. The predicted risk of referral is too high from years 1-3 and too low for years 4 to 6.

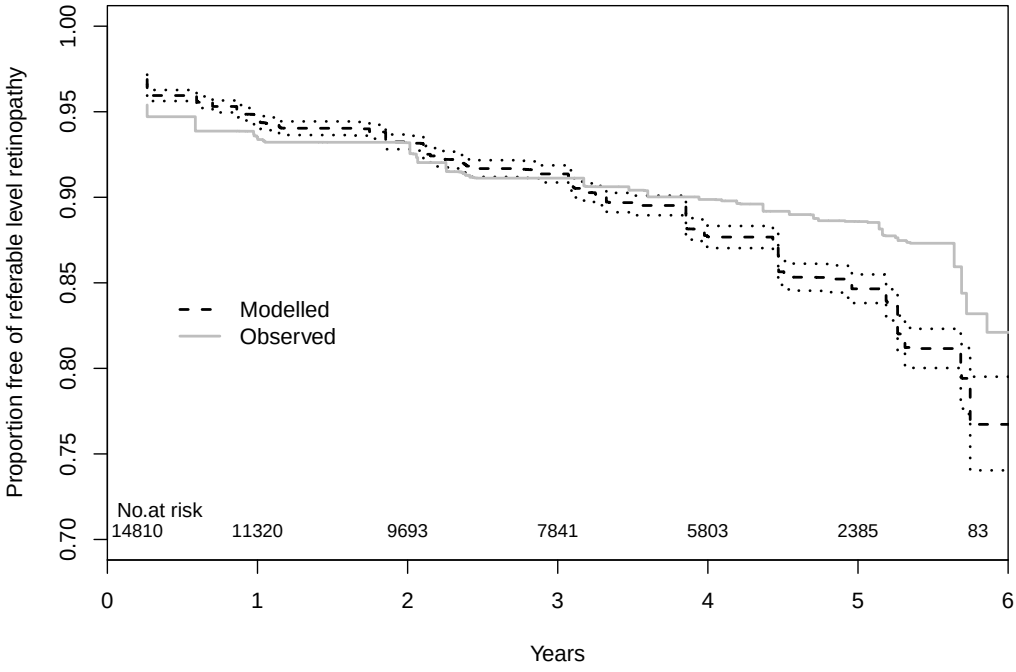


Figure 4.4: *Observed and modelled survival from referable retinopathy*

Prevalence plots

The model was used to predict the prevalence of each grade of DR over time. Figure 4.5 shows the observed and modelled prevalence for each of the four states of DR by years since first screen. The modelled prevalence closely follows the observed prevalence of no detectable DR and background DR, but over-predicts the prevalence of pre-proliferative and proliferative DR consistently over the 6 years.

State change plots

To assess how well the model predicts consecutive screening results, state change plots of the *survival* time until a different grade occurred were constructed using both observed and model predicted times. Results are presented as Kaplan-Meier type plots. Figure 4.6 compares observed and modelled survival times for forward transitions and figure 4.7 shows the equivalent plot for

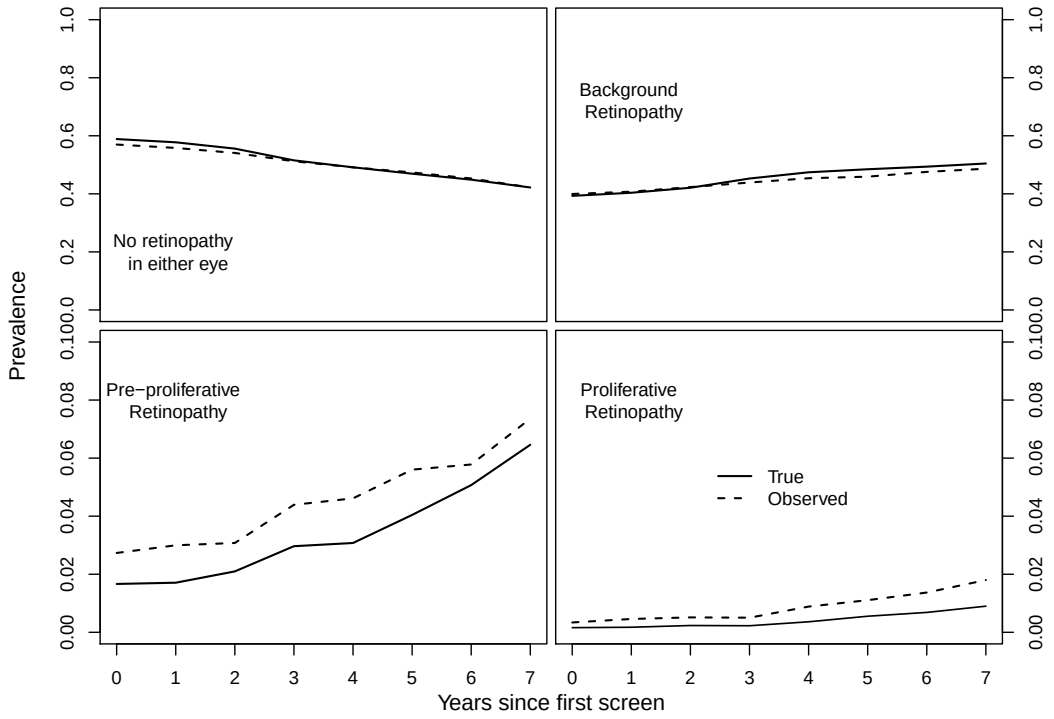


Figure 4.5: *Observed proportions of retinopathy grades by years since first screen with modelled predicted proportions overlaid. Note that the lower panel plots have different y axis scale*

the backward transitions. These state change plots show that there is reasonable agreement with the time to move to an advance state of retinopathy (forward transition) but poor agreement with the time to regress from a higher state. Time to regress back is predicted to be much faster and occurs more often than it is observed in the data.

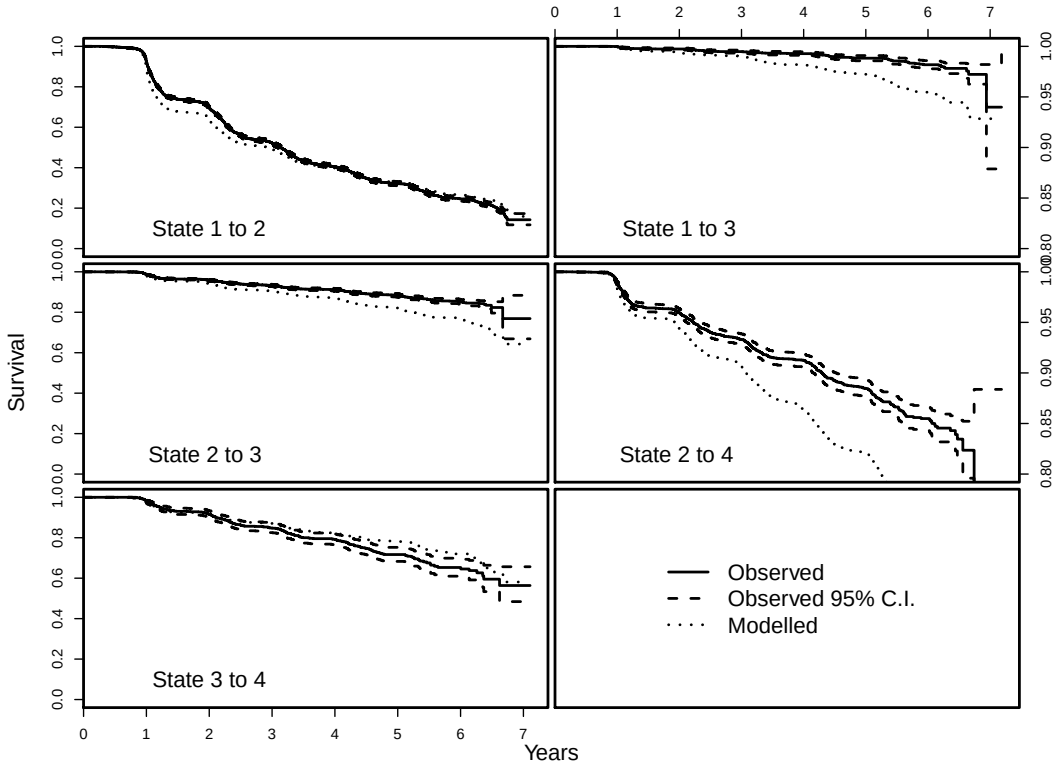


Figure 4.6: *Observed and predicted survival times until change from one screening grade to a more advanced grade*

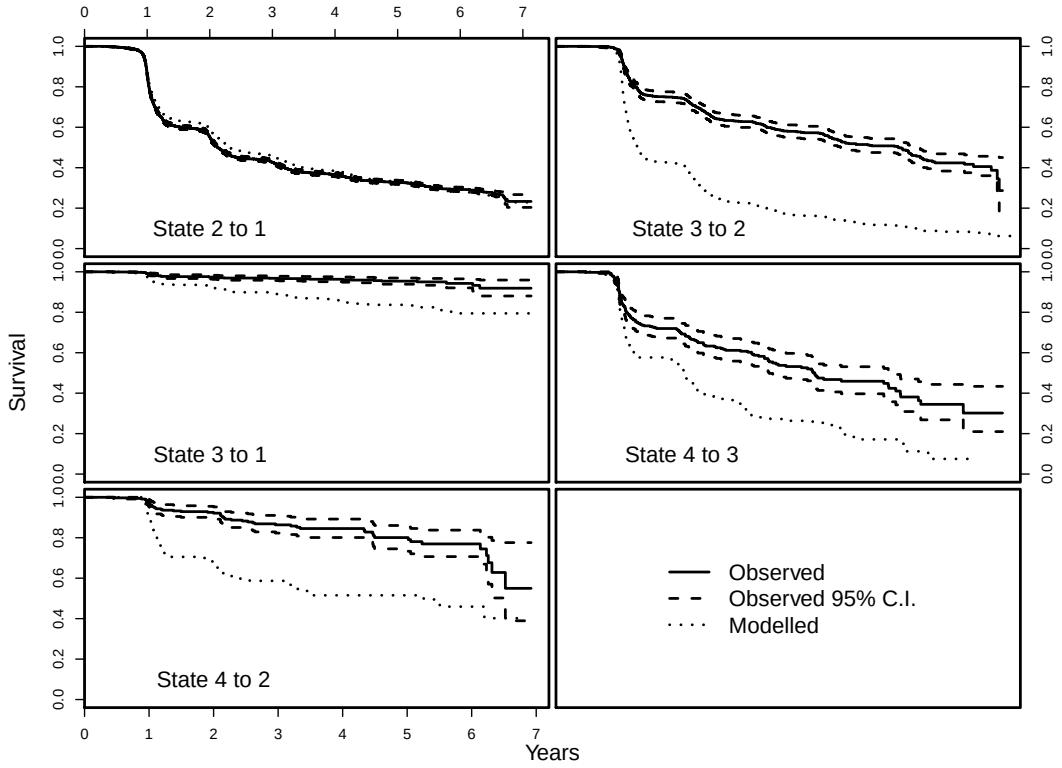


Figure 4.7: *Observed and predicted survival times until change from one screening grade to a less advanced grade*

4.5 Evaluating screening for DR

Sensitivity and specificity

The model was used to assess the sensitivity and specificity of screening for STDR (pre-proliferative DR or proliferative DR) and monitoring for progression from pre-proliferative (R2) to proliferative DR (R3) at different intervals and for different risk groups. Table 4.10 shows the numbers of true positive tests, false negatives tests and the sensitivity of screening if the baseline screen showed no detectable DR (R0). Table 4.11 shows equivalent results for *monitoring* for development of proliferative retinopathy. This assumes that pre-proliferative retinopathy has been detected at the previous screen. Table 4.12 shows the rates of true positive tests, false positive tests and the ratio of false positive to the total number of positive tests for the screening scenario and Table 4.13 for the monitoring scenario.

	Screening for STDR (R2 or R3)					
	Low risk		Medium risk		High risk	
Time from first screen	FN/TP	Sens.	FN/TP	Sens.	FN/TP	Sens.
1	0.0/0.0	56.3	0.0/0.0	56.6	0.2/0.2	57.5
2	0.0/0.0	57.1	0.0/0.0	57.5	0.5/0.7	58.5
3	0.0/0.1	58.4	0.2/0.2	58.8	1.9/2.9	60.0
5	1.4/2.3	62.1	3.2/5.4	62.6	16.6/29.6	64.0

Table 4.10: *True positive and false negatives from screening for DR after initial screen showed no retinopathy. Numbers are per 10,000 screened. FN-false negatives, TP - true positives, Sens - Sensitivity (TP/(TP + FN)). Low risk - HbA1c = 60 mmol/mol, duration of diabetes = 5 years, No maculopathy. Medium risk = HbA1c = 70 mmol/mol, duration of diabetes = 10 years, no maculopathy, High risk - Medium risk plus maculopathy in one eye*

Responsiveness of screening

The time taken to detect referable DR was calculated for different intervals of screening. Table 4.14 shows the mean and median response time for intervals

	Monitoring for Proliferative Retinopathy (R3)					
	Low risk		Medium risk		High risk	
Time from first screen	FN/TP	Sens.	FN/TP	Sens.	FN/TP	Sens.
1	0.7/1.0	58.6	2.3/3.3	59.2	22.0/34.2	60.9
2	1.8/2.6	59.5	4.9/7.4	60.2	36.3/59.2	62.0
3	5.0/7.8	61.0	11.7/18.9	61.7	63.1/110.6	63.7
5	29.5/54.6	64.9	51.8/99.6	65.8	157.4/334.4	68.0

Table 4.11: *True positive and false negatives from monitoring for proliferative retinopathy after detection of pre-proliferative retinopathy. Numbers are per 10,000 screened. Legend as per table 4.10*

	Screening for STDR (R2 or R3)					
	Low risk		Medium risk		High risk	
Time from first screen	FP/TP	PPV	FP/TP	PPV	FP/TP	PPV
1	4.4/0.0	0.0	9.9/0.0	0.1	47.2/0.2	0.4
2	6.8/0.0	0.1	14.6/0.0	0.3	63.6/0.7	1.1
3	11.7/0.1	0.6	23.7/0.2	1.0	91.0/2.9	3.1
5	36.1/2.3	5.9	63.5/5.4	7.9	180.1/29.6	14.2

Table 4.12: *True positive and false positive from screening for STDR after initial screen showed no retinopathy. Numbers are per 10,000 screened. Legend as per table 4.10*

	Monitoring for Proliferative Retinopathy (R3)					
	Low risk		Medium risk		High risk	
Time from first screen	FP/TP	PPV	FP/TP	PPV	FP/TP	PPV
1	81.7/1.0	1.0	140.3/34.2	2.3	405.7/34.2	7.8
2	105.9/2.6	2.6	174.8/59.2	4.1	462.4/59.2	11.3
3	144.0/7.8	5.1	224.9/110.6	7.8	529.1/110.6	17.3
5	254.4/54.6	17.7	351.1/334.4	22.1	641.0/334.4	34.3

Table 4.13: *True positive and false positives from monitoring for proliferative retinopathy after detection of pre-proliferative retinopathy. Numbers are per 10,000 screened. Legend as per table 4.10*

of screening of 1, 2, 3 and 5 years and also the proportion whose response time was greater than 1,2 or 3 years. This is based upon a hypothetical cohort who had no detectable retinopathy (R0) at baseline, had duration of diabetes of 5 years and HbA1c was 60 mmol/mol

	Interval of screening			
	1 year	2 years	3 years	5 years
Mean time	1.0	1.8	2.6	4.1
Median time	0.9	2.3	3.2	2.6
% > 1 year	47%	73%	82%	82%
% > 2 years	19%	44%	62%	62%
% > 3 years	7%	29%	41%	41%

Table 4.14: *Responsiveness of screening for STDR from programmes with different length intervals.*

4.6 Clinical validation of models

Comparison with estimates from diagnostic accuracy studies

I found five diagnostic accuracy studies that met the inclusion criteria. A number of studies either had too few patients or used only a single field image and thus were not included. Table 4.15 shows the reported prevalence, sensitivity and specificity for classification of background DR and STDR and the predicted sensitivity and specificity when the prevalence is matched to the study prevalence.

	External Study				Model Estimate	
Retinopathy Grade	Study	Prevalence	Sens.	Spec.	Sens.	Spec.
Background DR	Kalm [111]	37.8%	92%*	N.A.	78%	87%
Referable DR	Stellingwerf [112]	4.5%	95%	95%	75%	85%
	Schachat [113]	7%	87%	98%	77%	86%
	Olson [114]	8.3%	93%	93%	78%	86%
	Scanlon [115]	49%	83%	93%	98%	97%

Table 4.15: *Comparison of estimates of sensitivity and specificity for detection of retinopathy from four studies and the hierarchical ordinal model estimates. * Average of left and right eye sensitivity. N.A. - not reported*

Comparison with estimates from screening studies

I found two studies that reported results from regrading screening images. Manjunatha *et al.* [116] reported the results of independent secondary grading and arbitration of digital photographs initially graded as having no retinopathy by the Wolverhamption screening service. Table 4.16 compares the estimates from the Wolverhampton regrading errors with those predicted by the ordinal model. Sallam *et al.* [107] investigated reasons for disagreement between retinopathy

Level Retinopathy	Secondary Grading	Model Estimates
None	2343 (86%)	2419 (89%)
Background	367 (13.5%)	456 (16.8%)
Pre-proliferative	1 (0.04%)	0 (0.07%)
Proliferative	0 (0%)	0 ($< 0.0001\%$)
Total	**2716**	**2716**

Table 4.16: *Comparison of regraded images with initial screen result of no retinopathy with model predictions of screening errors.*

grades assigned by the Gloucester screening service and retinopathy grades from the hospital eye service. In eyes where the screening service and the hospital grading differed for either retinopathy or maculopathy, screening photographs were jointly reviewed by a senior ophthalmologist and a senior grader to form a consensus grade [107]. I used their reported consensus grades as the reference test and formed estimates of the screening service operational sensitivity and specificity. Table 4.17 compares the false positive and false negative rates from the Gloucester screening service and those predicted by the model, as reported by Sallam *et al.* [107].

		Empirical Estimates	Modelled Estimates
Retinopathy	False positives	6.7%	3.5%
	False negatives	3.7%	2.75%

Table 4.17: *Table comparing empirical estimates of referral errors (false negatives and false positives) with those predicted by the ordinal model*

4.7 Conclusion

In this chapter I investigated whether the monitoring model could be usefully extended to ordinal measures arising from screening or monitoring programmes. Using data from an established screening programme for diabetic retinopathy I devised a hierarchical ordinal model to estimate the natural history of diabetic retinopathy and then evaluated different screening strategies. To validate the models, I compared independent estimates of the rates of errors expected due to screening from published studies and compared them to predicted errors using the modelling approach.

Restatement of findings

Modelling diabetic retinopathy

The creation of an ordinal monitoring model was motivated by diabetic retinopathy (DR) screening. The model was relatively easy to estimate, albeit using specialist MCMC software, something that may limit its use among non-statisticians. As maculopathy did not fit into an ordering of DR it was incorporated into the model as a covariate and risk factor for progression of DR. The presence of maculopathy in one or both eyes, higher HbA1c and longer duration of diabetes were all positively associated with increased risk of developing true referable DR. The model predicted that at baseline, over-grading *no DR* as *background DR* and vice-versa occurred in 1 in 5 photographs and over-grading *background DR* as *pre-proliferative DR* occurred in 2% of such cases, and over-grading *pre-proliferative DR* as *proliferative DR* occurred in 5% of cases. The percentage of photographs that over-graded *background DR* as *pre-proliferative DR* (referable grade) increased to 5% after five years and 11% after ten years. The probability of under-grading *pre-proliferative DR* as *background DR* was 36% and under-grading *proliferative DR* as *pre-proliferative DR* at 39%.

Statistical validation of the model

Statistical validation of the model was done by checking assumptions of normality in the random effects, comparisons of the predicted and observed prevalence, survival from referable DR and the time to change in DR state. The ordinal model made good predictions of the prevalence of the *no detectable DR* and *background DR* retinopathy but over predicted the prevalence of STDR (pre-proliferative and proliferative DR). This was also reflected in the survival comparison which clearly showed that predicted risk of referral was higher for years 1-3 to and lower for years 4-6 than the observed survival. The state change or *Bureau* plots for the forward transitions showed reasonable agreement with the observed times but some of the backward transitions showed comparatively poor fit against the observed times. This was particularly the case for the time taken to regress from states 3 (pre-proliferative retinopathy) and state 4 (proliferative retinopathy) where the model suggested a much shorter time to transit backwards. This lack of fit may be due to the fact that the model cannot explicitly model regression in the underlying condition.

Clinical validation

Clinical validation (validation in external data) of the error component of the model was done by comparing published estimates from diagnostic accuracy studies comparing two-field digital photography against a recognised reference standard and also from studies which retrospectively regraded screening images. The results were mixed, with the estimated error rates from the model showing poor agreement with the diagnostic accuracy studies but good agreement with estimates from the screening based studies. The poor agreement with the diagnostic accuracy studies may be due to the fact that the sensitivity and specificity of the monitoring test under the ordinal model are not constant but change proportionally with the underlying disease. This is in contrast to the assumption commonly used in diagnostic accuracy studies. It is

therefore imperative that the case-mix is closely matched. The model predicted accurately the consensus grades of images from the Wolverhampton DR screening service and reasonably well with the consensus grading from the Gloucester DR screening photographs.

Evaluating monitoring and screening strategies

I used the model to compare sensitivity and specificity of screening for STDR and further monitoring for proliferative DR and how they changed with respect to the interval of surveillance and baseline risk. I did this by estimating the sensitivity for intervals of 1,2,3 and 5 years for cohorts of people deemed to be low, medium and high risk. As expected, the sensitivity improved as the interval was extended reflecting the fact that the "signal" increases over time. The gain in sensitivity over the five years was estimated at $6-7\%$ for all risk groups. As the risk increased, the sensitivity of screening for DR increases by approximately 1% consistently, which is much smaller than the gain from increasing the interval. In this model, risk did not interact significantly with time with respect to sensitivity. A similar pattern was observed with monitoring for proliferative disease, the sensitivity increased but the overall signal was weaker suggesting development from pre-proliferative DR to proliferative DR is slow. For both screening and monitoring, sensitivity is estimated to be low (60%-70%) which implies a number of people with STDR are being missed by screening. I also compared the specificity of screening and monitoring for DR. Again, as the interval increased the proportion of positive tests which were true positives increased. The positive predictive value is very low reflecting the fact that most positive tests from screening are false positives. The improvement in PPV by increasing the interval of screening and monitoring is significant but even after five years remains low. Unlike sensitivity, risk (as measured by baseline covariates) had a noticeable effect on false positive rates, again reflecting the stronger signal to noise ratio. Finally, I assessed responsiveness of screening for pre-proliferative retinopathy by simulating people who had

no detectable retinopathy at baseline and had a positive gradient (implying that at some point in the future they will develop pre-proliferative DR). the mean response time was 1 year, and over half of those with referable DR being detected by first screening visit. With biennial screening the mean response time increased to 1.8 and the majority of people with referable DR do not get detected within one year. This pattern continues as the interval increases and although precision is improved by screening less often (better sensitivity), longer intervals mean longer delays in patients to be referred.

Discussion

Regular eye examinations to detect or monitor DR has been justified because the disease is common and treatment is effective if given early on. The interval and frequency of examinations has been the subject of much research and a number of studies have examined whether the interval could be extended for low risk groups [117][118][119] and be more frequent for people with poor control of systemic risk factors [97]. Using novel methodology, I modelled progression of retinopathy over time and developed simulation models to assess the effectiveness of screening. The findings are broadly in line with other studies and suggest that for people with good control of risk factors and no evidence of retinopathy, screening could take place between 3 and 5 years as the risk of developing proliferative disease is less than 1 per 10,000. The modelling shows that even screening people every three years, there are many more false positives than true positives screening results. The positive predictive value (the likelihood that a positive finding is true) is less than 5% until the interval exceeds four years. The sensitivity of screening is also low ($\sim$ 60%) which indicates a significant false negative rate but due to the very low rate of true referable disease in the Gloucester cohort, the absolute number of people missed by screening is very small, less than 1 per 10,000 for low and medium risk groups. Once background retinopathy has been detected screening more

frequently (1-2 years) depending on risk factors, is more effective but still requires that 200 people need to be screened to find one person with referable DR when risk is moderate.

Comparison with the literature

The argument for extending the interval of screening and monitoring for diabetic retinopathy rests upon the three factors, the rate at which disease progresses, the specificity and the sensitivity of screening. Changes to these parameters brings about significant changes to the overall effectiveness. For the evaluation to be credible the values used in the simulation models should be comparable to estimates from natural history and cohort studies. Looker *et al.* [117] recently published estimates of the rate of referable retinopathy and concluded that for those with recently diagnosed type 2 diabetes and no retinopathy at their last screening the percentage developing referable level retinopathy was less than 0.03% at one year and 0.07% at two years. The estimated number of people observed with referable level retinopathy from the ordinal model is 0.04% at one year and 0.07% at two years, but the number with true referable disease is much less, $< 0.01\%$ and 0.01% respectively. In the Gloucester data there were 3088 people with newly diagnosed diabetes with no retinopathy or maculopathy in either eye at the first screening visit and none of these were detected with referable retinopathy at the following visit, or at the third screen.

Estimates from the literature of the sensitivity and false negative rate from screening DR vary. In a study from 1995 [120], which compared community based mydriatic photography with slit-lamp biomicroscopy reported there were 8 false negatives from 15 slit-lamp biomicroscope confirmed cases with sight-threatening diabetic retinopathy. The sensitivity (95% C.I) was estimated at 47 (21 to 93). The false negatives tended to have peripheral venous beading or cotton wool spots with artefacts that rendered grading of peripheral retina difficult [120]. More contemporary studies have reported much lower false negative rates. Results from the Gloucester screening service re-grading study found

that screening missed only 5 from 139 referable cases (False negative = 3.6%). The study of Manjunatha [116] found only one consensus graded image which was a pre-proliferative grade, suggesting either that the false negative rate is low or that very few of the false negatives have *no DR* screen results. In a study designed specifically to test accuracy, two-field mydriatic digital photography missed 48 from 242 seven-field stereo photography confirmed cases (sens. = 80%)[115].

Evidence of the scale of false positives was more consistent although reports are few in number. Sallam *et al*[107] counted 20 cases in which a referable grade was recorded by the screening programme in the absence of the type of lesions that qualify a pre-proliferative grade. This was from 300 who on consensus did not have referable DR (FP = 6.7%). Percentages of false positives from two other diagnostic accuracy studies were 5% [112],and 7% [114].

Limitations

The model did not permit maculopathy as an outcome and to overcome this I included it as a risk factor. It would have been preferable to model it as a competing risk, given that diabetic retinopathy and maculopathy are managed differently and represent different disease processes. Ideally, the model predictions would be validated against repeated measures of both an index screening test and gold standard test (seven field digital photography or ophthalmologist examination). As no such data exists (as far as I am aware) I compared predictions with published estimates from trials investigating diagnostic accuracy of screening tests. The total lack of false positive results is in stark contrast to all of the model based findings presented in this chapter and may be due either to model mis-specification or discrepancies due to the design of this study. Finally, this analysis did not include any attempt to work out how to quantify the trade off between false positive and false negative findings. As there may be little physical harm from a false positive screening test, and potentially catastrophic harm from failing to identifying sight-threatening DR any future

investigations should attempt to tackle this, although it may require health economic expertise.

In summary

The evidence presented adds weight to the argument that monitoring and screening intervals for DR can be extended because the risk of developing sight-threatening DR is very low and screening detects a disproportionate number of false positives which lead to people being referred to ophthalmology unnecessarily. Intervals could be personalised and may already be happening in practice [97]. Intervals and frequency could be set depending on control of blood sugar, length of duration diabetes and any evidence of retinopathy. Research has shown that these approaches can be cost-effective [121]. However, longer intervals can be associated with delayed intervention for the few that are at risk of developing proliferative retinopathy. As the consequences of inaction for these people is so serious (vision loss, blindness), and the chance of physical harm arising from false positive results is negligible means any decisions to examine less frequently would need to be taken carefully.

Chapter 5

Models for categorical data: Diabetic retinopathy

In this chapter I describe how a continuous time hidden Markov (HM) model can be used to estimate the rates of true and false positive assessments under different monitoring intervals and for different baseline risk. The model has two states to represent the early stages of retinopathy and two *absorbing* states for referable retinopathy, one with maculopathy and one without maculopathy. Transition probabilities for the progression of retinopathy and misclassification probabilities were estimated using the retinopathy screening data. The model was used to simulate a cohort of people with diabetes being screened over a period of six years under different monitoring protocols. The model was validated internally using predictive model checking and externally by comparing prevalence estimates and empirical estimates of misclassification probabilities.

5.1 Introduction

Markov models

Markov models are often used to describe natural histories of disease. These models are particularly suited to conditions or diseases that have an interpretation in terms of stages of disease. For example, Markov models have been applied to diabetic nephropathy [94], the progression of human immunodeficiency virus disease [92],remission of functional disability following stroke [122]

and Alzheimer's disease [123]. Diabetic retinopathy has also previously been modelled as a Markov process [124][125] and more recently by Looker *et al*[117].

The defining property of the Markov process is what is sometimes referred to as *lack of memory*. This says that the future state or prognosis is determined entirely by the knowledge of the present state and potentially, relevant covariates. The history up to the current state does not affect prognosis under the Markov assumption. In these models, the probability of progression or regression from one state or stage to another is often referred to as a transition probability and these can be estimated directly using individual person data or from aggregate data. Markov models are very flexible and can be adapted to suit many different conditions. If a particular condition is understood to be a progressive condition from which recovery is not possible then the model can be configured to represent this. Alternatively, if remission is deemed possible from early stage disease then this can also be accommodated, in each case by either fixing the relevant transition probabilities to zero or to let them be estimated. States from which one cannot move backward or forward are known as absorbing states. Death is one such example, but any state which a person does not *escape* from can also be modelled in this way. Figure 5.1 shows a simple (bi-directional) model for a condition with three states, well, ill or dead. In this diagram, an arrow or *edge* from one state to another signifies a non-zero

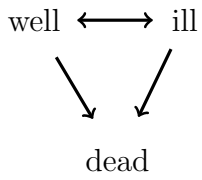


Figure 5.1: *Example of a simple multistate model*

probability or possible transition. The model in figure 5.1 permits recovery from the illness as there is an arrow from *ill* to *well*, although for many applications this may seem unreasonable and this route will be blocked. Death is

treated as an absorbing state and this is specified in the graphic by the lack of arrows back from *dead*. The transition from *well* to *dead* acts as a surrogate for modelling competing risks from causes other than that of the illness being modelled. For each arrow in the diagram there is a corresponding transition probability.

Discrete time Markov models can be used when the outcome have been observed at regular intervals. Transition probabilities will then reflect the interval at which the data was observed. For example, if the data comprises annual measurements, then estimated transitions probabilities will represent probabilities of moving from one state to another in one year. The transition probabilities are presented as a matrix which I will refer to hereafter as a transition probability matrix or *tpm* for short. The positioning of the numbers in the matrix is significant and the row position indicates the current state and the columns index the next state. Hence, the number in the first row of the first column of the tpm (p_{11}) is the probability of being well in a year's time if they are well today. The number in the first row, second column (p_{12}) gives the probability of becoming ill in one year given the person is well today. The upper-diagonal elements of the matrix p_{12}, p_{13}, p_{23} represent probabilities of moving forward or progressing. The lower-diagonal states p_{21}, p_{31}, p_{32} represent regression or remission. In this model, the regressing probabilities from state 3 p_{31} and p_{32} are zero representing no chance of recovery from this state.

	Well	Ill	Dead
Well	p_{11}	p_{12}	p_{13}
Ill	p_{21}	p_{22}	p_{23}
Dead	0	0	1

Table 5.1: *Hypothetical transition probability matrix for a three state model*

Continuous time Markov models

Although, screening and monitoring is scheduled to takes place at regular intervals, it is often the case that patient visits can vary widely. A continuous

time generalisation of the discrete time Markov model has wide appeal for modelling monitoring and screening data. In addition to relaxing the assumption of consistent follow-up times, a further advantage with continuous time models is that they can be used to calculate instantaneous risk or transition *intensities* as well as transition probabilities for any interval of time.

Transition intensities are also formatted into a matrix with rows and column positions representing movement from one state to another. The setting out of the structure of the intensity matrix; (which transitions are feasible, which are not), is an important step in building the model. It will often require expertise and theoretical justifications as empirical estimates may not be available. When working across disciplines to define the structure of the intensity matrix it is worth emphasising that intensities represent instant transitions and that this has a different interpretation to transition probabilities. For example, in a model with many recognised stages it may not be thought possible for someone to die of the disease without passing through all the stages in turn. Hence, the transition intensity from *well* to *death* will be zero. However, this does not mean that the equivalent transition probability will also be non-zero, as within the interval, people may pass through each stage. It is unlikely that empirical evidence will be found to support or refute whether this can happen in practice and so some subjective decision will need to made and sensitivity of the choices made to the structure of the intensity matrix.

Hidden Markov models

It is often reasonable to assume that the test used for monitoring or screening is subject to error. A hidden Markov (HM) model is an extension of the discrete or continuous time Markov model for the case in which the outcome is observed with error and the true state is hidden. HM models have an additional set of probabilities to represent probabilities of being observed in state x given the true state is y. These *misclassification* or *error* probabilities are bundled into a misclassification or error matrix. Table 5.2 is a hypothetical error matrix

for the illness-recovery model. The elements of the error matrix correspond

	Well	Ill	Dead
Well	e_{11}	e_{12}	0
Ill	e_{21}	e_{22}	0
Dead	0	0	1

Table 5.2: *Hypothetical error probability matrix for a three-state model*

to having a true state indexed by rows whilst simultaneously being observed in a state indexed by the column. In other words, e_{11} is the probability of being observed well given that you are well, e_{12} is the probability that the assessment or test gives a result that the person is ill when in fact there is nothing wrong with them. The probability of being misclassified dead given that the person is well or ill is zero, as is the probability of being observed well and alive given the person is truly dead, this will be a reasonable assumption in most models. This structure allows estimation of two different prevalence's; the true prevalence and the observed prevalence.

Estimating the true state and probability of misclassification: An example

Diabetic retinopathy (DR) is graded from R0 - R3 as an ordinal scale, R0 indicating no detectable retinopathy through to R3 proliferative retinopathy. Suppose a sequence of annual screening results is observed as R1, R0, R2, R1. A relevant question is; does the observed sequence represent real change in the underlying condition with some initial evidence of improvement (R1 - R0), but later deterioration (R0 - R2) and finally back to the grade observed at the start or does it just reflect short-term variation or a noisy measurement? Figure 5.2 shows a network graph of the observed (bottom row) and true (top row) sequence with question marks in the true nodes to represent that these are unknown. If we assume that the screening test never makes an error then the true sequence is just the same as the observed sequence. However, if we do

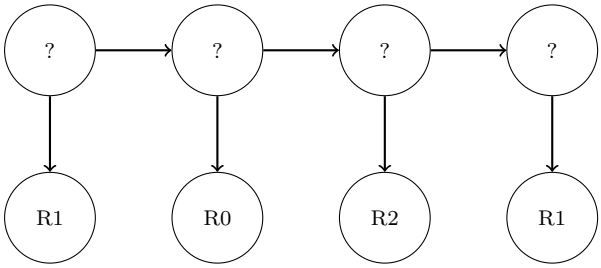


Figure 5.2: *Directed graph of a hidden Markov model for a sequence of observed retinopathy grades for a single person*

not make that assumption then the true sequence may not match the observed sequence and a task is to estimate the true grades. If we assume, for now, that we know the population transition probabilities and error probabilities, then we can select the true sequence of retinopathy grades most likely to have generated the observed sequence. For example, if we assume a progression only model with known annual transition probability; matrix P;

$$P = \begin{bmatrix} 0.8 & 0.15 & 0.04 & 0.01 \\ 0 & 0.8 & 0.15 & 0.05 \\ 0 & 0 & 0.9 & 0.1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where the elements of P are the probabilities of transition from row to column states in one year. It can be interpreted as follows, the probability that at a person remains with no retinopathy (R0) after one year is 0.8, the probability a person will develop background retinopathy (R1) having been R0 is 0.15, progress to pre-proliferative retinopathy (R2) 0.04 and proliferative retinopathy (R3) 0.01. As this is a progression only model, the lower diagonal numbers are all zero. The misclassification or error matrix E supplies the probability of an observed grade given the true grade. Suppose that the misclassification or error matrix for this model is;

$$\begin{bmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0.1 & 0.8 & 0.1 & 0 \\ 0.05 & 0.1 & 0.8 & 0.05 \\ 0 & 0.1 & 0.1 & 0.8 \end{bmatrix}$$

The error matrix can be interpreted as follows. Reading the rows as the "true" grade and the columns as the observed grade, the probability of observing any

True state sequence	$Pr(True)$	$Pr(2,1,3,2\|Tr)$	Product
$\{1,1,1,1\}$	$1*0.8*0.8*0.8$	$0.1*0.8*0.1*0.1$	0.0004
$\{1,1,1,2\}$	$1*0.8*0.8*0.15$	$0.1*0.8*0.1*0.8$	0.0006
$\{1,1,1,3\}$	$1*0.8*0.8*0.04$	$0.1*0.8*0.1*0.1$	0.00002
$\{1,1,1,4\}$	$1*0.8*0.8*0.01$	$0.1*0.8*0.1*0.1$	0.000005
$\{1,2,1,1\}$	$1*0.15*0*0.8$	$0.1*0.1*0.1*0.1$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\{4,4,4,4\}$	0 * 1 * 1 * 1	0.1 *0 * 0.1 * 0.1	0

Table 5.3: *Example of a likelihood computation for the four state model*

one of the four grades when the true state is R1 is given by the second row of E. In other words, the probability of misclassifying a person as a R0 when there true grade is R1 is 0.1 meaning that one in 10 will be misclassified.

The *likelihood* of the observed sequence can be calculated for every possible true sequence. For every true sequence there is a probability for the observed sequence. The computation can be broken down into two components, the true sequence probability and the observed probability conditioning on the true, *or*.

$$[\text{True}|\text{Observed}] \propto [\text{Observed}|\text{True}][\text{True}] \tag{5.1}$$

The probability of a true sequence is the product of the relevant probabilities from P, i.e (for a transition $1 \rightarrow 1$, the relevant probability is the 1st row, 1st column of P, for the transition $1 \rightarrow 2$, we require the 1st row, 2nd column of P. The true probability is multiplied by the product of observed probabilities conditioning on the true sequence. The observed probability conditioning on the true is given by the appropriate element from E (e.g. the probability of observing a grade 2 given a true state of 1 is taken from the 1st row, 2nd column of E). Table 5.3 shows how likelihood calculations for a 4 state HMM model are constructed for one sequence.

For a sequence of 4 grades there are $4^4 = 256$ possible paths and hence 256 computations. The path which maximises the likelihood in this example is row 2 in table 5.3 and is the sequence 1,1,1,2. This is the most likely true sequence for the observed sequence $O = \{2,1,3,2\}$ and the transition and

misclassification matrix P and E. This method of finding the most likely path is known as global decoding in HM models. For this estimated true sequence there are two observed grades that disagree with the true grade, the first and third observations. This illustrates how the model estimates the probability of misclassification.

The computations for each sequence are onerous but not difficult. However, we will often need to estimate the transition and misclassification matrices given sequences of data rather than assume they are known. This requires finding the intensity and error matrices that maximise the likelihood across all the data.

Aims and objectives

In this chapter I set out to see if Markov models could help quantify grading classification error from screening for diabetic retinopathy. This is done by first constructing and estimating a model for the progression of underlying disease and the errors due from screening using a HM model. Then as before, using the estimates from the model, use simulation to assess the effects of different screening intervals for different risk profiles.

5.2 Data

Modelling data

The data set to derive the model was the Gloucester screening data. For details, see section 4.2, page number 130.

Validation data

As part of the HTA retinopathy project, I was given access to repeated measures data from the Lambeth diabetic retinopathy screening service. The data set had repeated retinopathy grades in 49,954 people from the period 2007-2010, constituting 120,331 person-years of follow-up.

5.3 Exploratory analysis

A useful way to summarise categorical responses over time is by a frequency table of pairs of consecutive states. This counts over all individuals in the dataset, for each pair of states (r- row, c - column), the number of times state c has been followed by state r. Table 5.4 is such a table for the four state outcome.

State at prior visit	State at next screening assessment			
	No Detectable Retinopathy	Background Retinopathy	Referable Maculopathy	Referable Retinopathy w/o Maculopathy
	n (% of the row total)			
No Detectable Retinopathy	20783 (76)	6440 (24)	24 (0)	5 (0)
Background Retinopathy	6089 (30)	13330 (65)	751 (4)	254 (1)
Referable Maculopathy	30 (2)	608 (37)	869 (53)	140 (9)
Referable Retinopathy without Maculopathy	1 (0)	170 (27)	117 (19)	337 (54)

Table 5.4: *Observed transitions between states of retinopathy and maculopathy in people undergoing diabetic retinopathy screening.*

Table 5.4 illustrates that the majority of the screening grades represent the early stages of diabetic retinopathy (undetectable and background stages). It is shows that it is unlikely ($< 0.1\%$ that a referable grade of retinopathy or maculopathy will follow a screen which detected no retinopathy in either eye the previous year. These probabilities increases to 4% for maculopathy and 1 for retinopathy when the previous screen detected background retinopathy. The lower diagonal cells (row 2, column 1) of the table represents consecutive screens where background retinopathy has appeared to regress. In this data, most screens showing background retinopathy in one year will be followed by the same result the next year, although a significant proportion will regress (30%). As we have good reason to believe that the screen process will over-grade and under-grade in some cases these numbers represent the mixtures of

people whose disease has progressed and the misgraded photographs. It is aim of this analysis to estimate the occurrence of these errors.

In the English National Screening Programme for Diabetic Retinopathy (ENSPDR), referral to ophthalmology is indicated if maculopathy is detected or pre-proliferative or proliferative retinopathy is detected by screening. People will be referred to ophthalmology and may receive intravitreal injections or macular laser if maculopathy is found or panretinal photocoagulation for retinopathy. They may then be referred back to the screening programme. Table 5.4 shows that in 608 times a referable grade of maculopathy was followed by screens detecting only background retinopathy and in 30 cases, no detectable retinopathy at all. Again, these cases could due to misclassification by screening or genuine remission bought about by treatment. As we know that the screening process does not always reflect the true state of disease as judged by a suitable reference standard (7 field stereophotography or Ophthalmologist examination), these observed transitions will be a mixture of true progression and regression of disease, under or over-grading by the screening service and potentially, improvements due to treatment after referral. The hidden Markov model attempts to solve which transitions are real and which are spurious. To successfully implement the model a number of assumptions about the pattern of true progression of disease need to be made and it has to be decided what type of errors are possible or not conceivably possible.

5.4 Methods

This section describes the methods used to construct the HM Model for screening diabetic retinopathy and maculopathy, how it was estimated, used to evaluate different screening intervals and finally how it was validated.

Introduction

In screening for diabetic retinopathy eyes are graded into one of four degrees of retinopathy, starting with no retinopathy (R0) up to proliferative disease (R3) and for the presence or absence of maculopathy, *M0* or *M1*. This gives $4*4*2*2 = 64$ potential combinations per person and is far too many to model as states in a Markov model. This is for a number of reasons. Firstly, there is an issue of identifiability of the model. An inherent feature of HM models is that the hidden or latent states are interchangeable and any permutation of the hidden states has the same likelihood [126]. Therefore, there are many plausible sets of estimates for each set of data and this problem is exacerbated with bigger and bigger models. A second related problem is that the number of parameters to estimate grows exponentially with increasing number of states, for example a basic HM model with two states has five parameters (one to represent the initial state probability, and two each for the intensity and error matrices), three states, 14 and four states 27 parameters. An eight state Markov model model has 105 parameters. Although there are plenty of observations (over 80,000 grades in the screening data) the length of the individual sequences is also important. In the case of unevenly spaced observations, sequences of two observations with at least two different time intervals occurring in the data are needed to provide enough degrees of freedom to identify a simple two-state model [126]. The maximum follow up in the GDRSS data is seven years and thus there is a maximum of seven observations per person. Another issue which is particular to the GDRSS data is that people would have been referred to ophthalmology if either eye had pre-proliferative or proliferative

State	Retinopathy	Maculopathy
1	No detectable retinopathy in either eye	None
2	Background retinopathy	None
3	Any	Any level of maculopathy
4	Pre-proliferative or proliferative	None

Table 5.5: *The four states representing retinopathy and maculopathy progression*

retinopathy or there was evidence of maculopathy. Whether any treatment was given after referral is not recorded in the screening data and hence we do not know whether they received any treatment. This has implications for trying to model errors of over-referral (screening results which indicate a referral grade, but the true grade is lower) because we cannot identify whether apparent improvements after a referral grade are a result of treatment. Establishing whether any treatment was administered would help remedy this problem.

The states of the Markov model do not need to be ordered as they were in the ordinal model of the previous chapter and this gives us more flexibility to combine different degrees of referable maculopathy and retinopathy without having to justify one being more or less severe than the other. After much discussion with the lead consultant ophthalmologist and health economists on the HTA project a consensus was arrived at with a model describing the early progression of retinopathy and absorbing states which separate referral involving maculopathy from retinopathy without maculopathy. The reasoning for this was based upon defining utilities in the health economic modelling. This is because, maculopathy is more likely to affect vision loss and thus quality of life where as pre-proliferative or proliferative retinopathy rarely causes visual impairment. Table 5.5 shows the 4 states constructed for the purpose of modelling the screening data.

Description of the model

The first step in building the Markov model is to set the structure of the model. That is, to decide whether certain transitions (movement to other states) should be allowed and estimated or blocked and set to zero. An alternative approach is a data driven approach, along these lines;

1. Allow all possible transitions
2. Estimate the model
3. Block transitions estimated as having near-zero probability
4. Re-estimate the model

I chose to build the model upon theoretical considerations of the pathophysiology of retinopathy. For example, there is good evidence that background retinopathy, mainly categorised by micro-aneurysms, can come and go without intervention [127][128], therefore it made good sense to allow regression back from the early stage of retinopathy. This makes this model bi-directional and sets it apart from the ordinal linear model in the previous chapter. Due to uncertainty around whether treatment was administered after referral, I decided that states 3 (maculopathy) and 4 (retinopathy) would have to be absorbing states with no progression or regression possible. This means that observations occurring after a referral grade (for the same person) are censored. As it is not considered theoretically possible to progress to pre-proliferative retinopathy without first passing through a background stage (State 1 - 3), transitions from no detectable retinopathy in either eye (State 1) to pre-proliferative (State 4) were set to zero. Table 5.6 is a table showing which transitions are permitted under the model. To indicate which of the $4 * 4 = 16$ transitions are possible the cells corresponding to the rth row and sth column are given a q subscripted with the row and column numbers, hence if transition from state 1 to state 3 is permissible the cell corresponding to the 1st row and third column of the

matrix has $q_{13} = 0$. Cells with zero's indicate that a transition is blocked, hence if transition from state 1 to 4 is blocked, the 1st row, 4th column of Q should have a zero. Diagonals given 1's indicate absorbing states. Figure 5.3

	to			
from	No DR	Background DR	Referable Maculopathy	Referable Retinopathy
No DR	q_{11}	q_{12}	q_{13}	0
Background DR	q_{21}	$q_{2,2}$	q_{23}	q_{24}
Referable Maculopathy	0	0	1	0
Referable Retinopathy w/o Maculopathy	0	0	0	1

Table 5.6: *Intensity matrix for the four state retinopathy model. The q_{ij} notation refers to transition that are estimated by the model, and 0's to transitions which are blocked*

is a graphical illustration of the four state model.

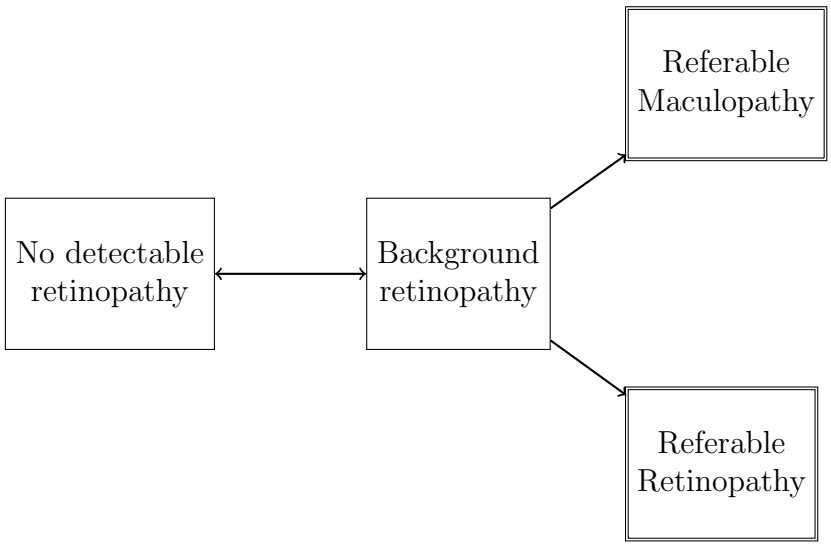


Figure 5.3: *Illustration of the four state model for diabetic retinopathy screening*

Estimation of the model

The HM model's in this chapter were fitted using the *msm* [129] function in R [72]. Successful convergence to maximum likelihood is not guaranteed and may depend on a number of factors. In attempting to fit a number of different models I found the following steps helped obtain useful estimates.

1. Use plausible guesses as starting values for the intensity matrix or random number constrained in a reasonable range.

2. Use Nelder-Mead simplex option to quickly refine the starting values - switch off hessian matrix calculation save considerable amounts of time

3. Rescale the likelihood using the *fnscale* option to avoid over-flow problems (set *fnscale* to same as -2 log likelihood)

4. When looking to finalise model and obtain convergence - use conjugent gradient method and a large number of iterations (e.g. 2000)

5. Allow time (24 hours)!

Internal validation

Assessing the fit of a HM model is not straightforward for a number of reasons. As with the linear model method, the component of the model which is arguably of the most interest; the true prevalence is unobservable and hence it is often not possible to assess how well the model fits this crucial component of the model. In addition, using screening data to fit the models means that the transition times will not be known precisely and only observed to have occurred within a (potentially large) interval of time (at least the interval of screening). Also, depending on the design of the study, the observation times may also be irregular and unbalanced making many goodness-of-fit tests invalid or excessively complicated. However, simple methods of model assessment can be effective to some extent [49]. In a comprehensive review article of diagnostic

methods for multi-state models, Titman and Sharples (2009) [130] appraised tools for diagnosing both multi-state Markov and multi-state Markov models with classification error. From this article, I found three methods which would be feasible and useful for assessing the fit of the HM model.

Prevalence plots

The first of the internal validation methods compares the observed screening prevalence in the data with that estimated by the model. The consists of "binning" observations by suitable blocks of time and counting up the number observed in each state. For the GDRSS data I decided to count observed states by periods of one year as although there is a large amount of data the rates of referable maculopathy and retinopathy are low. The counts in each state are divided by the number of people observed within the same time frame. Confidence intervals for the observed prevalences are computed using an "exact" method to prevent confidence intervals for the prevalence exceeding 0 or 1. The same procedure is done for the model prediction by simulating a fake data set using model parameters but the same structure of the data. The observed prevalence and 95% confidence intervals are compared graphically using line plots.

Survival from referable disease

If the assumptions in the Markov model are correct then there should be close agreement between the observed survivor function and the survivor function predicted by the model [109]. Defining survival as not being referred due to either maculopathy or retinopathy I will compare modelled and observed survival separately using standard survival functions in R [72]. Because data from screening is scheduled with annual and not random measurements I used a parametric distribution (Weibull) to fit the survival curve that takes into account that event time is not known exactly but only known to occur within a

certain time interval. The predicted survival curve was estimated by simulating new data based upon the structure of the derivation data set.

State-change plots

A more stringent test of the model are comparisons of the time to change in state using graphical methods. State change plots were first proposed by Bureau *et al* in their analysis of a two-state recurrent model and extended to multi-state models by Titman and Sharples [49]. These are plots of the observed and predicted Kaplan-Meier estimates of the time to transit from one state to another. Even though they compare the observed state and not explicitly the hidden state they reflect both the underlying latent and observed process.

For all possible transitions in the observed data, I calculated the time from the first occurrence of the first state until the first time of the occurrence of the *end* state if it occurred. If the *end* did not occur then the event is censored. A person may contribute more than one observation if they enter the first state more than once. Kaplan-Meier curves estimated from the data were compared to model predicted curves by first simulating a data set using the characteristics and intervals of the original data. A second set of curves was then estimated using the same method as for the actual data.

Evaluation of different screening intervals

The best fitting model was used to evaluate the effects of increasing the interval of screening. For different observed baseline states I calculated the true state or retinopathy/maculopathy for those referred because of (a) maculopathy and (b) retinopathy. An advantage with Markov models, is these calculations can be made without the need for simulation and are quicker and not susceptible to Monte Carlo error, although they can be calculated this way if needed. Two tables were created for each baseline retinopathy status and represented numbers per 10,000 screened in each state. At each interval of screening, the proportion of people who are observed to have no retinopathy at baseline,

background retinopathy in one eye, or background in both eyes at baseline and observed with referable disease at the subsequent screen were derived by calculation using the estimated transition probability matrix (tpm) and error probability matrix.

Assumptions

In order to derive the true and observed states of eye disease a number of assumptions are needed. Firstly, that all screening appointments are kept or equivalently that if they are not it is not related to the level of disease. Secondly, the condition does not manifest itself and become clinical leading to a diagnosis either through primary or secondary care. Lastly, all photographs are gradeable or equivalently that images if ungradeable, the reasons are not related to the level of eye disease.

External validation

External validation of the model was done by comparing predictions from the model with an external source of screening data and published estimates of sensitivity and specificity of digital photography against a recognised reference standard. Comparing state change predictions from the model tested how well both the latent and observed components of the model function and the comparison with estimates of diagnostic accuracy will asses the "error component" of the model .

State change comparison

Using the same method as described in section 5.4 the time to change in state observed in the Lambeth data will be compared with that predicted by the model.

Comparisons with diagnostic accuracy studies

A number of cohort studies have reported incidence or prevalence of proliferative or retinopathy requiring photocoagulation treatment. The Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR) reported estimates of the incidence of proliferative retinopathy in people without any retinopathy in a first retinal examination [41]. The United Kingdom Prospective Diabetes Study (UKPDS) reported the proportion of people requiring photocoagulation treatment at 3, 6 and 9 years [127]. More recently, Jones *et al* reported incidence and progression of diabetic retinopathy from 17 years of screening data [131]. Very recently, there have been two studies in which retinopathy has been modelled using the same HMM methodology and these estimates are broadly comparable to estimates reported here.

The direct comparison of model estimates and published estimates is complicated by the structure of my model and whether the papers have reported incidence of maculopathy.

The model estimates the rates of pre-proliferative retinopathy by multiplication of the initial state probabilities by the transition matrix and finally the error matrix.

External validation of the error matrix

To validate the error matrix in the model I compared estimates published in the literature from (1) diagnostic accuracy studies of diabetic retinopathy using suitable gold standard methods (either full ophthalmic examination, slit-lamp biomicroscopy or seven-field stereophotography) with two-field mydriatic digital photography and (2) in studies where screening images were regraded and given a consensus reference grade. I restricted inclusion to studies with a minimum of 100 participants and comparable populations (i.e. not people who cannot attend screening such as people who reside in care homes). Sensitivity and specificity for detection of background and referable retinopathy and

referable maculopathy can be compared directly to the appropriate cells in the estimated error matrix. To compare regrading images where the observed result is known and the regrade is assumed to be the true grade then by using Bayes theorem we can recover the appropriate conditional probability. For example, if images that were initially graded as no retinopathy and were later re-graded as background retinopathy (false negatives), the appropriate probability from the model is found using Bayes thereom, i.e.;

$$[\text{observed DR}|\text{True DR}] \propto [\text{True DR}|\text{observed DR}][\text{observed DR}] \qquad (5.2)$$

5.5 Results

Model estimation

There were 14,810 patients in total and 13769 with usable repeated retinopathy grades and covariate information. The model was adjusted for duration of diabetes which was standardised (x - mean(x)/sd(x)) to obtain convergence to maximum likelihood. Transition and error probabilities were estimated using a hidden Markov model. The development of retinopathy was adjusted for duration of diabetes and HbA1c, but not gender. The three main outputs from the model are the intensity matrix which governs the progression of the eye disease, the error matrix which gives the rates of misclassification and the hazard ratios for the covariates. Table 5.7 shows the fitted intensity matrix for average HbA1c (54 mmol/mol) and duration of diabetes (5.4 years).

	to			
from	No DR	Background DR	Maculopathy	Retinopathy
No DR	-0.09 $(-0.10, -0.08)$	0.09 $(0.08, 0.10)$	0.00005 $(0, 0.001)$	0
Background DR	0.08 $(0.07, 0.10)$	-0.10 $(-0.11, -0.08)$	0.01 $(0.009, 0.01)$	0.002 $(0.001, 003)$
Maculopathy	0	0	0	0
Retinopathy	0	0	0	0

Table 5.7: *Intensity matrix for the unadjusted four state retinopathy model. Numbers represent transition intensities (*95%*) C.I.*

The average length of stay (sojourn) in each state of retinopathy can be calculated directly from the intensity matrix by taking minus the reciprocal of each of the diagonal elements. The average sojourn represents the true course of disease and does not include the spurious transitions that represent over-grading by screening. Table 5.8 gives the estimated sojourn time and 95 % confidence intervals for the *no detectable* and *background* stages of retinopathy.

	Estimate	lower (95% C.I.)	upper (95%) C.I
No DR	11.20	10.35	12.10
Background DR	10.50	9.11	12.00

Table 5.8: *Average sojourn time for the first two states of retinopathy*

Transitions can be presented in terms of intervals of time. Table 5.9 gives the transition probability matrix for an interval of 1 year. Each cell represents the probability of moving from state r to state c in any given year which corresponds to consecutive screens. For example, the number 0.92 in the first

	to			
from	No DR	Background DR	Maculopathy	Retinopathy
No DR	0.92	0.08	0.0005	0.0001
Background DR	0.08	0.91	0.01	0.002
Maculopathy	0	0	1	0
Retinopathy	0	0	0	1

Table 5.9: *Transition probability matrix (1 year) for the four state model.*

row, first column is the estimated probability of remaining free from detectable eye disease over the course of a year. The off-diagonals represent the probability of moving states in any one year, for example, row 1,column 2 (0.08) is the probability of developing background retinopathy and row 1 column 3 (0.0005) is the probability of developing referable maculopathy if the current state is no detectable retinopathy. The lower diagonal numbers represent the probability of regressing. This is only estimated for the early stages of retinopathy and is set to zero for the latter stages. The estimates of regression from background in one eye are larger than the probability of further development. For example, the probability of background retinopathy regressing to no detectable retinopathy within a year is 0.08.

Effects of covariates on the transitions are expressed as hazard ratios. The effect of standardised duration of diabetes is shown in table 5.10. The hazard

Transition	Hazard	Lower 95% C.I.	Upper 95% C.I.
No DR - Background	1.34	1.28	1.41
No DR - Maculopathy	1.29	0.26	6.37
Background DR - No DR	0.19	0.15	0.24
Background DR - Maculopathy	1.32	1.23	1.41
Background DR - Referable DR	1.47	1.30	1.67

Table 5.10: *Hazard ratios for the effect of standardised duration of diabetes (equivalent to a change of 7.8 years)*

Transition	Hazard	Lower 95% C.I.	Upper 95% C.I.
No DR - Background	1.41	1.34	1.48
No DR - Maculopathy	0.32	0.02	4.71
Background DR - No DR	0.87	0.80	0.96
Background DR - Maculopathy	1.46	1.33	1.59
Background DR - Referable DR	1.47	1.23	1.77

Table 5.11: *Hazard ratios for the effect of standardised HbA1c (equivalent to change of 15.2 mmol/mol)*

ratios can be interpreted as proportional to the probability of progressing when compared to the reference case. Hence for duration of diabetes, an increase of nearly five years is associated with a 30% risk of progression from no detectable retinopathy to background retinopathy. The probability of regressing however is 80% lower for duration of ten years compared to mean duration in the Gloucester cohort (5.4 years). Progression from background DR to referable maculopathy and retinopathy is 30% and 47% higher for people with duration of ten years compared to mean duration. I have chosen to ignore hazard ratios that have confidence intervals that cross 1 (No DR - maculopathy). A similar pattern is observed for HbA1c. Increasing HbA1c is associated with an increased risk of progressing from no DR to background DR and to referable retinopathy and maculopathy. An increase in one standard deviation of HbA1c

(15 mmol/mol) is equivalent to 8 years duration with respect to progression of eye disease.

A marked difference between the effect of HbA1c and duration is that duration seems to have a much stronger effect on reducing the rate of regression when compared to the effects of higher HbA1c.

The estimated error matrix is shown in table 5.12. The error matrix is

	observed state			
true state	No DR	Background DR	Maculopathy	Retinopathy
No DR	0.85 (0.007, 0.999)	0.15 (0.147, 0.163)	0^*	0^*
Background DR	0.04 (0.04, 0.06)	0.96 (0.03, 0.999)	0^*	0^*
Maculopathy	0^*	0^*	0.9998 (0.3585, 1)	0^*
Retinopathy	0.0003 (0, 0.03)	0.0003 (0, 0.04)	0.0003 (0, 0.02)	0.9993 (0.112, 1)

Table 5.12: *Error matrix for the four state model. Estimates represent probability of observing state given the true state (95%) C.I.* 0^* *indicates* < 0.001

interpreted as follows. The rows represent the true state and the columns the observed state. This is a generalisation of the standard two by two table used in diagnostic testing. Hence, upper diagonals are false positives (screening over-diagnoses the true state) and lower-diagonals are false negatives (screening under diagnoses). Of particular interest are the upper diagonal elements (rows 1-2, column 3-4) as these are the probability that screening refers a person to ophthalmology erroneously. Of equal importance are the lower diagonals (row 3-4, columns 1-2) as these give the probability of screening failing to detect referral level maculopathy or retinopathy and hence missed treatment opportunities. The model estimates that up-grading no detctable DR to background DR occurs in 15% of screens involving people who have no detectable DR. The probability of down-grading as no detectable DR is much smaller at 4%. Errors which would lead to under or over referral are practically zero.

Prediction and internal validation

Given the propensity of these models to find many different sets of parameter estimates with similar likelihood it is important that goodness of fit tests are carried out so that best models can be selected. The following graphics are different ways in which to test the fit of the model, the first compares estimates of prevalence (numbers in each state) with the model predicted prevalence. The second method compares survival from each absorbing state with modelled survival. The last method compares survival for each possible transition under the model. Although not shown, I have already ruled out a number of possible models for the screening data because they have produced poor predictions of the observed prevalence by state.

Prevalence

The estimated intensity and error matrix were used to re-create the screening data, replacing the observed grades with modelled predicted grades. Figure 5.4 shows the model predicted prevalences for each of the four states and the observed prevalences with exact 95% confidence intervals.

Survival from referable disease

Observed survival from referable maculopathy or referable retinopathy was compared to modelled survival by recreating the outcome data as per the prevalence method (described above). Figure 5.5 shows the Kaplan-Meier estimates for both the observed and modelled survival from maculopathy. Figure 5.6 shows the observed and modelled survival from referable retinopathy. Both figures represent the proportion remaining free from receiving a referable grade as a result of being screened annually.

State change plots

Figure 5.7 shows the state change plot validation for the four state model for maculopathy and retinopathy.

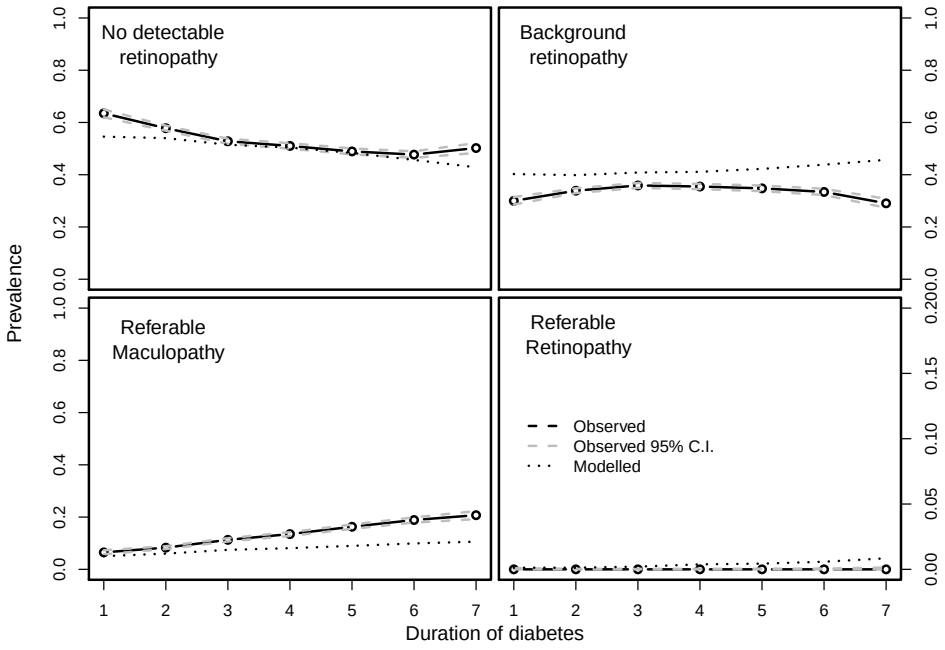


Figure 5.4: *Modelled screening prevalence versus the observed screening prevalence for the first three states of retinopathy*

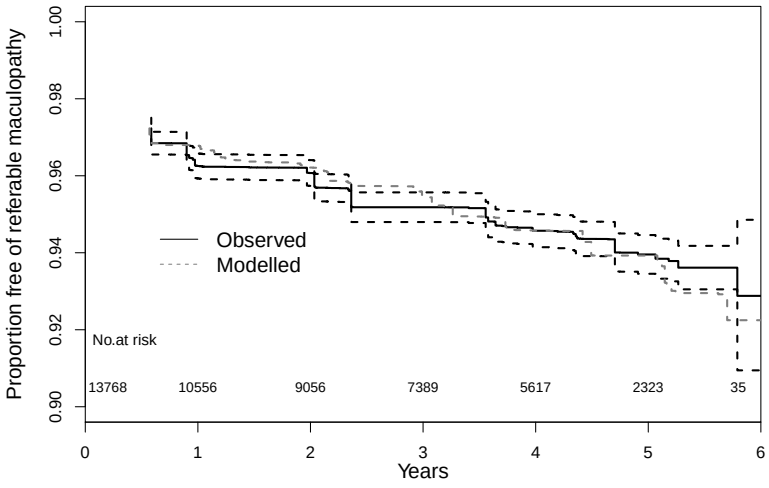


Figure 5.5: *Observed and modelled survival from referable maculopathy*

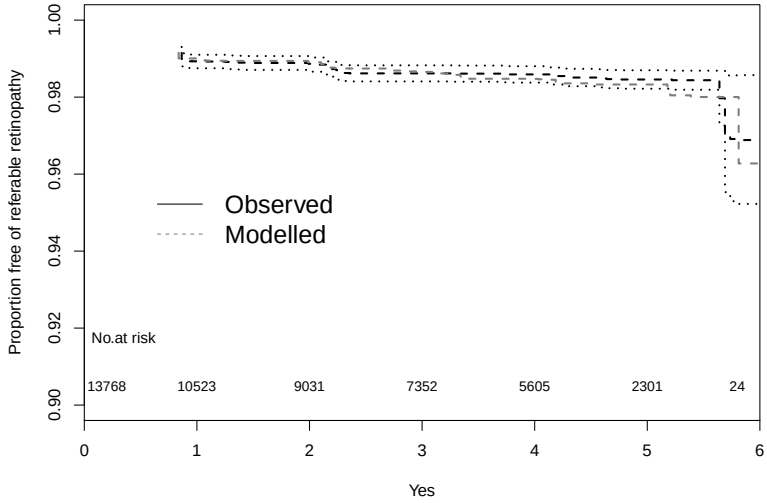


Figure 5.6: *Observed and modelled survival from referable retinopathy*

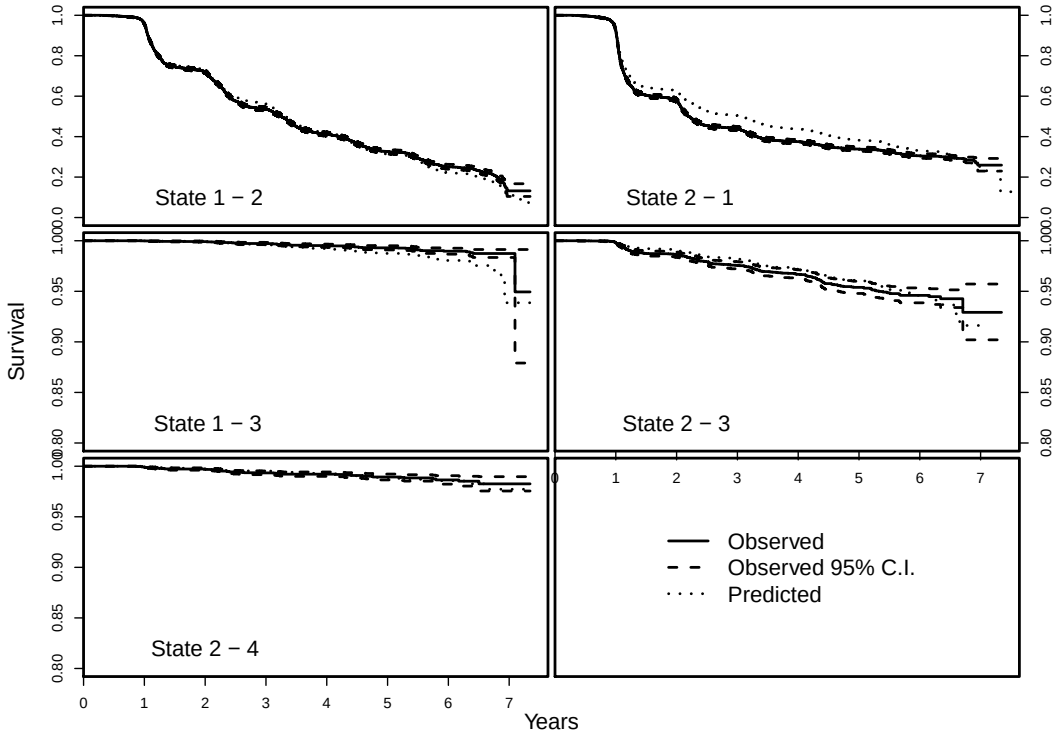


Figure 5.7: *Statechange plot of the observed and predictions from the model*

5.6 Screening for retinopathy and maculopathy

I used the HM model estimates of the intensities and the errors to derive the expected number of referrals from the screening service and their true states if the interval between the baseline and the following assessments were 1,2,3,4 or 5 years. Tables 5.13 and 5.14 show referrals for maculopathy and retinopathy when the baseline screen showed no detectable retinopathy.

No retinopathy at baseline

Interval	Referred at next assessment	True states			
		No detectable DR	Background DR	Any Maculopathy	Referable DR
	N (%)	N (%) of referred			
1	7 (0.07)	0 (0.7)	0 (0.1)	7 (99.2)	0 (0)
2	21 (0.21)	0 (0.2)	0 (0.1)	21 (99.7)	0 (0)
3	41 (0.41)	0 (0.1)	0 (0.1)	41 (99.8)	0 (0)
4	66 (0.66)	0 (0.1)	0 (0)	65 (99.9)	0 (0)
5	94 (0.94)	0 (0)	0 (0)	94 (99.9)	0 (0)

Table 5.13: *Expected referrals for maculopathy and the true unobserved level of disease per 10,000 people screened, who had no retinopathy at baseline. Screening intervals of 1 - 5 years*

The tables can be interpreted in the following way. Each row corresponds to the expected numbers for a specific interval (given in the first column) between assessments. The second columns gives the number (%) of referrals and columns 3 to 7 gives the number in each true state (percentage of the referrals). Hence, for three yearly monitoring we should expect 41 eferrals for maculopathy per 10,000 screened, all of which represent true maculopathy.

Background retinopathy at baseline

Tables 5.15 and 5.16 are for referral for maculopathy and retinopathy when background retinopathy was detected at baseline.

Interval	Referred at next assessment	True states			
		No detectable DR	Background DR	Any Maculopathy	Referable DR
	N (%)	N (%) of referred			
1	1 (0.01)	0 (3)	0 (0.7)	0 (0)	1 (96.2)
2	4 (0.04)	0 (0.9)	0 (0.4)	0 (0)	4 (98.7)
3	8 (0.08)	0 (0.4)	0 (0.3)	0 (0)	8 (99.3)
4	13 (0.13)	0 (0.2)	0 (0.2)	0 (0)	13 (99.5)
5	19 (0.19)	0 (0.2)	0 (0.2)	0 (0)	19 (99.7)

Table 5.14: *Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had no retinopathy at baseline. Screening intervals of 1 - 5 years*

Interval	Referred at next assessment	True states			
		No detectable DR	Background DR	Any Maculopathy	Referable DR
	N (%)	N (%) of referred			
1	70 (0.7)	0 (0)	0 (0.1)	70 (99.9)	0 (0)
2	136 (1.36)	0 (0)	0 (0)	136 (99.9)	0 (0)
3	200 (2)	0 (0)	0 (0)	199 (100)	0 (0)
4	260 (2.6)	0 (0)	0 (0)	260 (100)	0 (0)
5	319 (3.19)	0 (0)	0 (0)	319 (100)	0 (0)

Table 5.15: *Expected referrals for maculopathy and the true unobserved level of disease 10,000 per people screened, who had background retinopathy at baseline. Screening intervals of 1 - 5 years*

Tables 5.15 and 5.16 are interpreted in the same way as tables 5.13 and 5.14. Each row corresponds to the referral after an interval of either 1,2,3,4 or 5 years and the greyed out column represents the number of referrals which are correct. When the baseline screen detects background retinopathy, we can expect 70 referrals for maculopathy after an interval of one year, 136 after two and 200 after an interval of three years. The referral yield is between 4 and 12 times the number of referrals when no retinopathy is detected at baseline. As before, all referrals are correct assignments.

Interval	Referred at next assessment	True states			
		No detectable DR	Background DR	Any Maculopathy	Referable DR
	N (%)	N (%) of referred			
1	15 (0.15)	0 (0.1)	0 (0.4)	0 (0)	15 (99.4)
2	29 (0.29)	0 (0.1)	0 (0.2)	0 (0)	29 (99.7)
3	42 (0.42)	0 (0)	0 (0.1)	0 (0)	42 (99.8)
4	55 (0.55)	0 (0)	0 (0.1)	0 (0)	55 (99.8)
5	67 (0.67)	0 (0)	0 (0.1)	0 (0)	67 (99.9)

Table 5.16: *Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had background retinopathy at baseline. Screening intervals of 1 - 5 years*

Adjustment for duration of diabetes

The model can be adjusted for longer or shorter durations of diabetes and varying levels of HbA1c. The previous analyses assumed duration of diabetes was the same as the mean in the Gloucester data (5.4 years) and the same mean HbA1c (54 mmol/mol). In this model the covariates only change the rate of true progression and do not affect the screening error rates, therefore they have a limited impact on the misclassification rates. For example, adjusting the model to assume a duration of diabetes of 10 years, produces 29 referrals for retinopathy after an interval of five year, an increase of 50% when compared to the base case but there is little change in the positive predictive value (99.9% vs 99.8%). A similar result is found when adjusting for HbA1c. When HbA1c is assumed to be 15 mmol/mol above the mean, equivalent to 1 standard deviation then the number of referrals for retinopathy almost doubles but remain relatively low at 37 per $10,000$ screened. Adjusting for both simultaneously increases the number of referrals further but remains low 55 per $10,000$ screened.

Sensitivity analysis

I modified the error matrix used in the previous calculations to include false positive and false negative rates that better reflected empirical estimates from the literature. In the first of the sensitivity analysis, I changed the false positive referral rate for retinopathy to 5% if the true state is no detectable retinopathy and to 10% false positive rate when the true state is background retinopathy and 5% if the true state was maculopathy. This is equivalent to an overall false positive rate of 6% and specificity of 94%. These were the figures reported by Sallam *et al* [107]. For the second sensitivity analysis I changed both the false positive rate and the false negative rate. I kept the true positive rate at 0.94 as in the first sensitivity analysis but lowered the false negative rate to 0.965 which is close to the rate seen in Sallam *et al*.

Table 5.17 shows the expected number of referrals for retinopathy and their true states if the true screen detection error rate was 6%, equivalent to specificity of 94%. This analysis assumes that the first assessment showed no detectable retinopathy. Table 5.18 shows the expected number of referrals for

Interval	Referred at next assessment	True states			
		No detectable DR	Background DR	Any Maculopathy	Referable DR
	N (%)	N (%) of referred			
1	552 (5.52)	449 (81.4)	101 (18.3)	0 (0.1)	1 (0.2)
2	587 (5.87)	416 (70.9)	165 (28.2)	1 (0.2)	4 (0.7)
3	617 (6.17)	388 (62.9)	219 (35.4)	2 (0.3)	8 (1.3)
4	644 (6.44)	365 (56.6)	263 (40.8)	3 (0.5)	13 (2.1)
5	668 (6.68)	345 (51.6)	299 (44.8)	5 (0.7)	19 (2.9)

Table 5.17: *Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had no retinopathy at baseline. For screening intervals of 1 - 5 years and specificity equivalent to 0.94*

retinopathy and their true states if the true screen detection error rate was as table 5.17. This analysis assumes that the first assessment detected background

retinopathy. For the second set of sensitivity analyses I altered the false posi-

Interval	Referred at next assessment	True states			
		No detectable DR	Background DR	Any Maculopathy	Referable DR
	N (%)	N (%) of referred			
1	825 (8.25)	184 (22.3)	624 (75.7)	3 (0.4)	14 (1.7)
2	825 (8.25)	193 (23.3)	599 (72.7)	6 (0.8)	27 (3.2)
3	826 (8.26)	200 (24.1)	578 (70)	9 (1.1)	39 (4.7)
4	829 (8.29)	205 (24.7)	560 (67.6)	12 (1.5)	51 (6.2)
5	832 (8.32)	210 (25.2)	544 (65.4)	15 (1.8)	63 (7.6)

Table 5.18: *Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had background retinopathy at baseline. For screening intervals of 1 - 5 years and specificity equivalent to 0.94*

tive rate and the false negative rates. Table 5.20 shows the expected number of referrals for retinopathy and their true states if the true screen detection error rate was 6%, equivalent to specificity of 94% and the sensitivity was 96.5%. This analysis assumes that the first assessment showed no detectable retinopathy. Table 5.20 shows the expected number of referrals for retinopathy and

Interval	Referred at next assessment	True states				False Negatives
		No detectable DR	Background DR	Any Maculopathy	Referable DR	
	N (%)	N (%) of referred				n (%)
1	552 (5.52)	449 (81.4)	101 (18.3)	0 (0.1)	1 (0.2)	
2	586 (5.86)	416 (70.9)	165 (28.2)	1 (0.2)	4 (0.7)	
3	617 (6.17)	388 (62.9)	219 (35.4)	2 (0.3)	8 (1.3)	
4	644 (6.44)	365 (56.6)	263 (40.8)	3 (0.5)	13 (2)	
5	667 (6.67)	345 (51.6)	299 (44.8)	5 (0.7)	19 (2.8)	

Table 5.19: *Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had no retinopathy at baseline. For screening intervals of 1 - 5 years, specificity equivalent to 94% and sensitivity equivalent to 96.5%*

their true states if the true screen detection error rate was as table 5.19 but

assumes that the first assessment detected background retinopathy.

Interval	Referred at next assessment	True states				False Negatives
		No detectable DR	Background DR	Any Maculopathy	Referable DR	
	N (%)	N (%) of referred				N (%)
1	825 (8.25)	191 (23.2)	605 (73.4)	5 (0.6)	23 (2.8)	0 (0.5)
2	835 (8.35)	196 (23.5)	584 (70)	9 (1.1)	45 (5.4)	1 (0.9)
3	846 (8.46)	200 (23.7)	565 (66.8)	14 (1.6)	66 (7.9)	1 (1.3)
4	857 (8.57)	203 (23.7)	548 (64)	18 (2.1)	87 (10.2)	2 (1.7)
5	868 (8.68)	206 (23.7)	533 (61.3)	22 (2.6)	107 (12.4)	2 (2.2)

Table 5.20: *Expected referrals for retinopathy and the true unobserved level of disease per 10,000 people screened, who had background retinopathy at baseline. For screening intervals of 1 - 5 years, specificity equivalent to 94% and sensitivity equivalent to 96.5%*

5.7 Validation

Comparison with estimates from diagnostic accuracy studies

I found five studies that met the inclusion criteria. A number of studies either had too few patients or used only a single field image and thus were not included. Table 5.21 shows the reported estimates of sensitivity and specificity for detecting different levels of retinopathy and maculopathy.

Thresholds	Stelling-werf	Kalm	Olson	Schachat	Scanlon	HMM
	Sensitivity (%)					
Background	-	92	-	-	-	95
Any	83	-	83	87	-	97
Referable retinopathy	95	-	93	-	83	99
Maculopathy	-	80.5	83	-	-	99
	Specificity (%)					
Background	-	92	-	-	-	83
Any	83	-	83	87	-	83
Referable retinopathy	95	-	93	-	83	99
Maculopathy	-	80.5	83	-	-	99

Table 5.21: *Comparison of estimates of sensitivity and specificity for detection of retinopathy or maculopathy. HMM - Hidden Markov Model*

Comparison with estimates from screening studies

I found two studies that regraded screening images. Manjunatha *et al* [116] reported the results of independent secondary grading and arbitration of digital photographs initially graded as having no retinopathy by the screening service. Table 5.22 breaks down the 2716 images initially graded as clear of retinopathy by the grade attributed by consensus alongside numbers of false-negatives expected under the HM model. The predicted numbers and percentages in table

Level Retinopathy	Secondary Grading	HMM estimates
None	2343 (86%)	2668(98%)
Background	367 (13.5%)	48(2%)
Pre-proliferative	1 (0.04%)	0 $(0 * \%)^1$
Maculopathy	5 (0.2%)	0(0%)
Total	**2716**	2716

Table 5.22: *Comparison of regraded images with initial screen result of no retinopathy with model predictions of screening errors.* [1] *includes any referable retinopathy.* $* = 0$ *to two decimal places.*

5.22 were derived directly or by summing probabilities from the model error matrix (see table 5.12, section 5.5).

Sallam *et al* reported reasons for discrepancy between the Gloucester screening service and the hospital eye service. In eyes where the screening service and the hospital grading differed for either retinopathy or maculopathy, screening photographs were jointly reviewed by a senior ophthalmologist and a senior grader [107] to form a consensus grade. For retinopathy grading there were 102 discrepant images, 25 of which were considered errors by the screening service, 64 considered errors by the hospital eye service. Thirteen images were either unassessable or made errors that would not affect whether someone was referred at all. Table 5.23 shows a cross-tabulation of screening and consensus grades for referable retinopathy. From tables 5.23 and table 5.24, I formed

DRSS	Consensus		Total
	Non-Referral Grade	Referral Grade	
Non-Referral Grade	280	5	285
Referral Grade	20	134	154
Total	300	139	439

Table 5.23: *Table showing agreement between screening service grading and consensus grading for referable retinopathy*

DRSS	Consensus		Total
	Non-Referral Grade	Referral Grade	
Non-Referral Grade	247	1	248
Referral Grade	32	172	204
Total	279	173	452

Table 5.24: *Table showing agreement between screening service grading and consensus grading for referable maculopathy*

estimates of the false positive and false negative rate that can be attributed to screening with digital photography. For referable retinopathy, the screening service graded 5 images as non-referral grade retinopathy from a total 134 and 20 eyes as referable retinopathy from a total of 300 images considered normal by consensus. For referable maculopathy screening missed just one 1 from 173 photographs with referable maculopathy but over-referred 32 from 279 normal eyes. Table 5.25 compares the false positive and false negative rates for referral with those predicted by the HM model.

		Empirical Estimates	Modelled Estimates
Retinopathy	False positives	6.7%	0.002%
	False negatives	3.6%	0.09%
Maculopathy	False positives	11.5%	0.09%
	False negatives	0.6%	0.02%

Table 5.25: *Table comparing empirical estimates of referral error rates with rates predicted by the model*

Comparisons of State Change with External Data

Table 5.26 shows the raw counts of consecutive screen grades for the Lambeth data, divided by the row totals to give the percentages (or probabilities if divided by 100). The Lambeth percentages can be compared to the raw counts of transitions observed in the Gloucester data (see table 5.4).

State at prior visit	State at next screening assessment			
	No Detectable Retinopathy	Background Retinopathy	Referable Maculopathy	Referable Retinopathy w/o Maculopathy
	n (% of the row total)			
No Detectable Retinopathy	56695 (83)	10459 (15)	774 (1)	47 (0)
Background Retinopathy	8247 (28)	17160 (59)	3231 (11)	346 (1)
Referable Maculopathy	392 (6)	1967 (29)	4263 (62)	272 (4)
Referable Retinopathy without Maculopathy	24 (7)	65 (19)	139 (41)	112 (33)

Table 5.26: *Observed transitions between states of retinopathy and maculopathy in people undergoing diabetic retinopathy screening in Lambeth.*

Comparison of survival from referable disease

Kaplan-Meier product limit estimates of survival from referable retinopathy or maculopathy were derived from the Lambeth data. Curves were compared to the predicted curves from the Hidden Markov Model and the observed curves estimated from the Gloucester data. Figure 5.8 shows the estimates of survival from retinopathy by year and figure 5.9 shows the equivalent graphic for referable maculopathy.

Comparison of transition times

The time until a change in state was compared between that predicted by the model and that observed in the Lambeth data. Figure 5.10 shows Kaplan-Meier estimates of "survival" or time to event curves plotted together for forward and backwards transitions in the early stage of retinopathy.

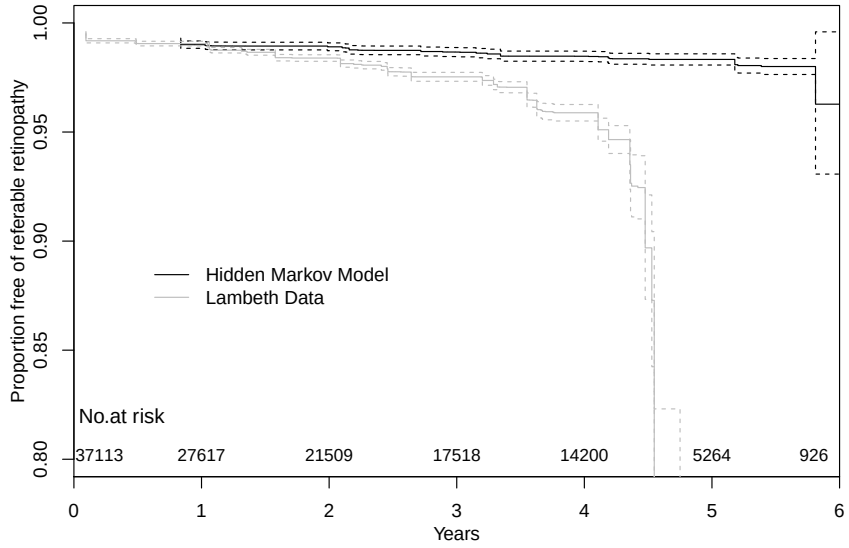


Figure 5.8: *Comparison of Kaplan-Meier estimates of survival from referable retinopathy for the hidden Markov model, and the observed survival from the Gloucester and Lambeth data*

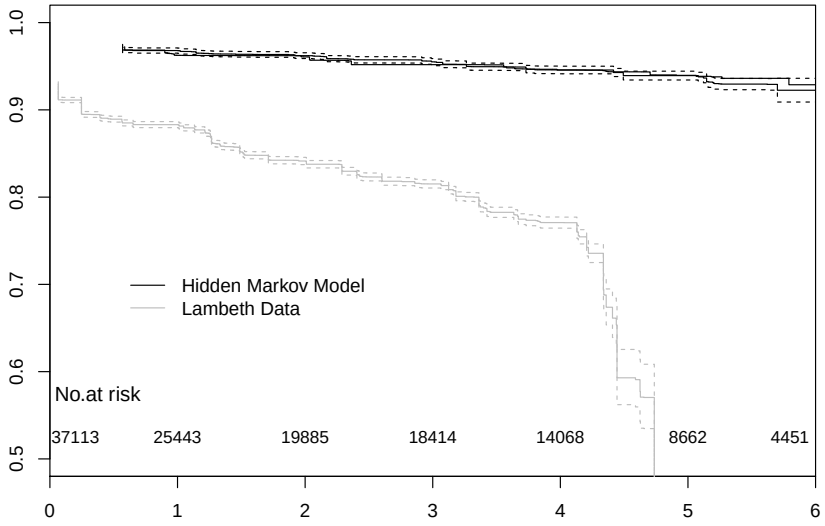


Figure 5.9: *Comparison of Kaplan-Meier estimates of survival from maculopathy for the hidden Markov model, and the observed survival from the Gloucester and Lambeth data*

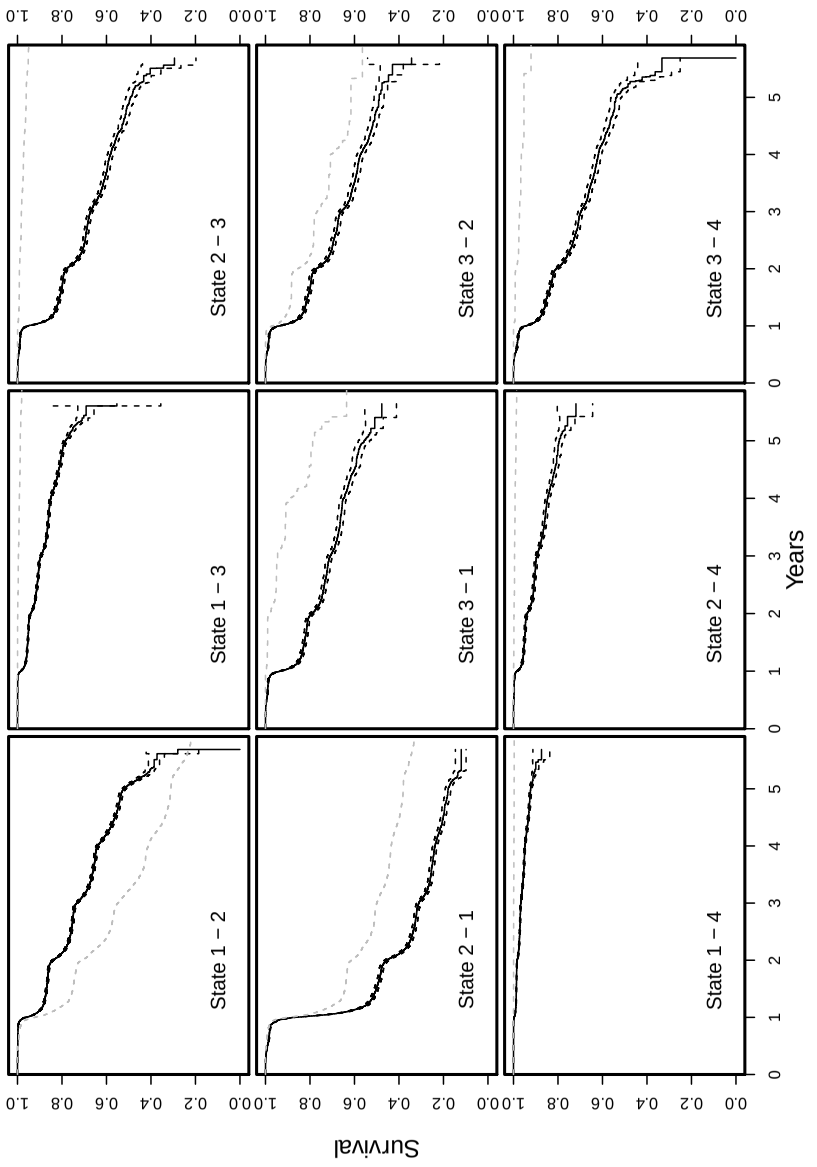


Figure 5.10: *State change plots comparing time to event between those observed in the Lambeth data and those in the modelled data*

5.8 Conclusion

In this chapter I used a *Hidden Markov* (HM) model as the basis for evaluating screening intervals for diabetic retinopathy. Hidden Markov approaches have not been used extensively to model screening and monitoring problems although there is one very recent example of this analysis, which is also been applied to diabetic retinopathy screening [117]. Using data from the Gloucester screening programme, I constructed and estimated a model for progression of retinopathy and maculopathy. Using this model, I evaluated the impact of changing the screening interval for different initial grades of retinopathy and calculated the proportion of referrals to ophthalmology that represented over-referrals (true state was background or no detectable retinopathy). To assess the validity of the model I compared the predictions from the model to published estimates from diagnostic accuracy studies and an independent source of screening data.

Modelling diabetic retinopathy

Fitting a HM model to the screening data was challenging. It took a considerable amount of time to find a model that produced estimates that passed all of the internal statistical validation. Using the same methodology, Looker *et al* also reported problems with finding maximum likelihood estimates for their HM model and they conceded that they did manage to obtain maximum likelihood estimates and settled instead for "a solution that could be interpreted as a misclassification model" [117]. Initial attempts to fit a larger six state model were largely unproductive and although the reference unadjusted model converged to a maximum likelihood estimate, adjusted models routinely failed and attempts to solve this problem proved futile. I considered this to be unacceptable as the current understanding is that duration of diabetes and glycemic control are important predictors of progression to sight-threatening retinopathy. By reducing the number of states to the model I was able to resolve these issues and adjust for both HbA1c and duration of diabetes.

The HM model estimated that there was significant over-grading *no detectable DR* as *background DR* (15%) and some under-grading of *background DR* as *no detectable DR*, but virtually no over or under referring for maculopathy or retinopathy (STDR). As a result of this, the cumulative incidence of true and observed referrals for STDR are practically identical. The length of duration of diabetes was shown to be associated with the progression of retinopathy, with longer durations of diabetes increasing the risk of forward progression and reducing the clearance of background retinopathy. Baseline glycated heamoglobin was also shown to be positively associated with increased risk of progression and lower chances of regression from background retinopathy. The HM model allowed for clearance of background retinopathy and this was shown to be $(0.08/0.01 = 8)$ times as likely as progression to maculopathy and many more time likely as progression to referable retinopathy.

Statistical validation of the model

Statistical or internal validation of the model was done by comparing predictions from the model with the observations. The comparison of predicted and observed prevalence showed that under the assumption of a Markov process for the progression of true retinopathy over-predicted the prevalence of background DR and under predicted the development of referable maculopathy. Predictions of the prevalence of no detectable retinopathy and referable retinopathy were consistent with the observed prevalences. Comparison of survival from referable disease reiterates this finding. Further inspection of the model was done by examining observed and predicted change from each state of DR. The state change plots showed excellent agreement with the time until a change in state observed in the data.

Clinical validation of the model

To clinically validate the model I compared estimates of sensitivity and specificity from diagnostic accuracy studies and studies which regraded images from

screening programmes. I found evidence that screening is liable to over- and under grade DR, although the exact scale of it is uncertain as estimates varied. For example, Sallam *et al* reported that in 4 (3.9%) photographs, the screening service misinterpreted dust spots or light reflections as microaneurysms or exudates and pigment spots as microaneurysms. Milton et al [132] also reported intra-reader variability could lead to over-grading no retinopathy as minimal retinopathy, but this happened in only one of the 54 (2%) photographs judged to have no retinopathy in the first reading. The estimated probability of missing background retinopathy using the HM model is 4.5% and this figure is directly comparable to empirical estimates from the Wolverhampton regrading analysis (see table 5.22. The model estimates are lower than that reported by Manjunatha *et al* [116] and this difference is probably too large to be just random variation and may reflect systematic variation between the Wolverhampton and Gloucester screening services.

In Sallams data, the screening service misdiagnosed drusen as exudates in 13 (10.4%) of photographs and in 9 (7%) images a referral for maculopathy was made when microaneurysms or haemorrhages were present but visual acuity was $> 6/12$ (a requirement under the English National Screening Programme) and hence they were over-referred. As expected, the model completely under estimated the rates of errors compared to those reported by Sallam *et al*. I believe this is due to the study design or data collection process rather than the model, although it could be a combination of the two. The reason for this is that because we had to censor all observations after a referral grade to ensure we did not mix recovery due to treatment with apparent recovery because of over-referral. Because we could not distinguish which observations were over-referrals and which were real improvements, the model under-estimates the rates of over-referrals. This problem could be averted if the hospital eye service data was linked to the GDRSS data so that we could distinguish which patients received treatment and which did not and which did not attend the eye service even if they were referred. With this information it would be possible to include

grades post referral. I was unable to attain this data within the timeframe of this thesis and had to rely upon using a sensitivity analysis approach.

I compared referral rates for maculopathy and retinopathy predicted by the model (which were in good agreement with the Gloucester data) with referral rates from the Lambeth screening service. The rates of referral of retinopathy and in particular maculopathy were in stark contrast to the predicted rates. Comparison of the times to change in state bore little resemblance to those observed in the Lambeth data. These findings places doubt on the usefulness of the HM model to make predictions and evaluate screening in different DR programmes.

Evaluating intervals of screening

I used the HM model to evaluate the effectiveness of screening for DR under different conditions. I examined the effect of changing the interval of screening when the initial screen showed no detectable retinopathy and when the initial screen showed background retinopathy. I did this for people with average risk (average duration of diabetes and average blood sugar) and for those with higher risk of progressive DR. For an interval of one year, if the initial screen showed no detectable retinopathy we could expect just 7 referrals for maculopathy for every 10,000 screened, rising to 21 for an interval of two years, 41 for three years, 66 in four years and 94 if the interval was five years. As the model estimates a very small rate of over-referrals, it is predicted that nearly all referrals represent true maculopathy. The number of referrals for retinopathy are lower still and we could expect 1 per 10,000 screened after one year, 4 after two years, 8 after three and 19 after 5 years. People with higher HbA1c at baseline and have had diabetes longer are at more at risk of developing referable DR, with an increase of 15 mmol/mol in baseline HbA1c leading to a doubling in the rate of retinopathy referrals and a duration of 10 years producing a 50% increase in referrals compared to people with duration of 5 years. When the ini-

tial screen showed evidence of background retinopathy, the number of referrals increases significantly for both maculopathy and retinopathy referrals.

Sensitivity analysis

At the time of writing this I was unable to resolve the lack of fit in error part of the model. To overcome this, I used a sensitivity type approach to recover what would happen under different screening intervals with referrals errors similar to those reported in the literature. This was done by substituting the model estimated false positive error rates with rates corresponding to those seen in diagnostic accuracy studies. In the first of these sensitivity analyses I changed the false positive referral rate from no retinopathy to 5% and from background to 10%. This was equivalent to a false positive referral rate of 6% overall which was the figure I obtained from numbers reported in Sallam *et al* [107].

A moderate adjustment to the false positive error rate has a large influence on (1) the number of referrals and (2) the likelihood that a referral is a true referral. This is a simple matter of "error amplification" by screening as Shuster coined it [51]. For example, for 10,000 people whose last screen showed no detectable retinopathy, 8982 will still have no DR a year later, and 1008 will have developed background retinopathy. Assuming a false positive rate of 5% from no DR and 10% from background DR we should expect (8982*0.05 = 449) referrals from those who still have no DR and $(1008 * 0.1 = 101)$ from those with background DR. This is set against the true rate of development of $1 per 10,000$. Hence, most referrals will be false referrals. Changing the false negative rate by a similar amount does not have the same effect as there are so few that develop true referable disease.

Irrespective of the false positive or false negative rate screening people for DR when the previous screen showed no detectable DR leads to an unacceptable low yield (less than 50 per 10,000 screened for any referable disease). Increasing the interval improves both the return and the specificity of screening with a negligible impact on the number of delayed referrals. Evidence of background

retinopathy, a longer duration of diabetes and higher baseline HbA1c is also associated with higher risk of developing referable disease and screening intervals could be adjusted according to these factors.

Discussion

In this chapter, I have shown that HM models can be used to find evidence for and evaluate screening and monitoring strategies. Although, estimates of referral error rates were unrealistically low and were at odds with published data this has been subsequently found to be dependent on the decision to censor patients after a referral grade has been observed. This would suggest that accurate estimation can only be obtained when there is sufficient data after the point of referrals.

There are a number of advantages to this approach over and above the linear model based method. Firstly, the ease with which the pathophysiology of disease can be modelled, in particular the ability to model actual regression of disease represents a distinct advantage for a number of conditions (retinopathy, nephropathy, tumour growth) over a method which restricts true disease to only progress. Secondly, the Markov method is better at modelling categorical states as it not restricted to ordinal levels of disease and can accommodate competing risks. This made it possible to model referrals to maculopathy and retinopathy separately, something that I couldn't do with the linear version. Lastly, estimates of progression and misclassification from the Markov method are easier to embed within a health economic analysis as Markov model are routinely used for such analyses [133].

There are also limitations to using this approach when compared to the linear model method. The Markov assumption may not be hold for diseases where the speed of progression through early stages of disease is predictive of the speed of further progression. Estimation of the model is not straightforward, is computationally intensive and larger models may not converge to maximum likelihood which may be considered a minimum criteria for some researchers.

Finally, estimation of the errors seems to depend on choices about censoring of people in the cohort. In the models shown in this chapter, censoring all screening sequences at the point of referral weakened estimation of over- and under-referral, something which did not occur with the linear model method.

Chapter 6

Summary

This thesis was motivated by the need to find optimal intervals for monitoring diabetic nephropathy and screening for diabetic retinopathy. Evidence for the timing and frequency of monitoring complications of diabetes as for many other important chronic conditions is often based on consensus rather than hard evidence. Randomised control trials are considered the gold-standard approach for evaluating monitoring strategies [25] but require very large sample sizes and have limited scope to evaluate many strategies, hence modelling plays an important role in building the evidence base.

In this thesis I have extended the existing methods to monitoring tests that have highly skewed, dichotomous or categorical outcomes. These adaptations to the methodology represent an original contribution to the area as previously, the method had only been used on normally distributed continuous variables. In addition, I have identified two new metrics for evaluating monitoring strategies; the advance and delay time. The advance time takes into account the fact that a false positive may represent an advanced detection and could in fact be beneficial, whilst the delay time quantifies the inevitable delay in detection and treatment change that comes with increasingly longer intervals of monitoring. The methodological developments could be useful for finding evidence for monitoring and screening strategies where the evidence is poor or absent. Monitoring eGFR in people with chronic kidney disease (CKD) and screening people at risk of cirrhosis are examples where the evidence is weak and the

methods described in this thesis could be used to evaluate different strategies.

Previous applications of the monitoring model to cholesterol, blood pressure, CD4 count and HbA1c monitoring had identified a trend for substantial numbers of false positive tests when monitoring intervals in low-risk populations were too frequent. In this thesis, a similar pattern emerges; when the signal is weak, frequent testing detects mainly false positives. Allowing time for changes to develop improves the signal and reduces the occurrence of false positives. The identification of this problem should help in the development of future screening and monitoring guidelines.

Monitoring diabetic nephropathy

The simulation modelling showed that frequent monitoring leads to a disproportionate number of false positive tests, increased operational costs (assays, clinic appointments, etc) and more people starting therapy and starting therapy earlier in life. A proportion of those false positives will go on to develop microalbuminuria but some will not and will be overdiagnosed (they will be treated but will not benefit). Increasing the interval of monitoring and therefore reducing the frequency of monitoring leads to less false positive results, fewer people overdiagnosed and less operational costs. However, when the interval of monitoring is increased more people have a false negative test and experience a longer delay in the time until microalbuminuria is detected. The increased false negative rate and delay represent a minority of patients but the clinical implications of being detected later need to be understood.

The efficacy of the current strategy for monitoring for microalbuminuria in people with type 1 diabetes is of questionable value. However, there is potential for improving current monitoring strategies identifying higher risk and lower risk patients with the intention of close monitoring of the high risk groups and less monitoring for the low risk group. Future research could involve evaluating strategies which stratify on risk and potentially lead to full scale RCT.

Conclusions about how often monitoring should be done in adolescents and young adults are limited by the fact that the simulation models used in this thesis could not account for the long-term clinical implications, such as the incidence of cardiovascular disease, proteinuria or end-stage renal disease. This is mainly because a full health economic model is outside the scope of this thesis, but the lack of direct evidence for the effectiveness of therapy in adolescents is a contributing factor. The simulation model has highlighted that under the current strategy there is potential for many patients to be overtreated, even when regression of microalbuminuria is not taken into account and hence current monitoring strategies for microalbuminuria in people with type 1 diabetes should be revised.

Screening for diabetic retinopathy

I used a ordinal latent model and later a Markov model to describe the progression of retinopathy in people with diabetes. Simulation models were then used to evaluate different aspects of screening and monitoring. Annual or biennial screening of people with no evidence of retinopathy appears to have limited value as very few people will develop sight-threatening DR in this time and although the specificity of the screening test is reasonably high, false positives far outweigh true positives. Screening people with higher systemic risk factors improves the sensitivity and reduces the false positive rate but not to the same extent as increasing the interval between screens. Monitoring for further deterioration (from pre-proliferative to proliferative) is a similar story; longer intervals improve the sensitivity and lower the overall false positive rate.

The estimates of errors from the statistical models although not directly comparable do seem to correspond to estimates from other studies. However, the rates of progression to proliferative disease are substantially lower than some reports. A higher and faster rate of progression to proliferative DR would have a significant impact on the effectiveness of monitoring further progression.

The reason lower transition rates seen in the Gloucester data may be three-fold. Firstly, there were very few people in the Gloucester data who progressed to proliferative retinopathy and so the estimate may be low due to sampling error. Secondly, there are few reports of the incidence from pre-proliferative DR to proliferative DR and what does exist is from older studies. Lastly, there is a growing body of evidence that suggests the incidence of proliferative DR is much lower than before due to improvements in diabetes care [118].

There were some limitations to the methods used to evaluate screening for diabetic retinopathy. We did not have access to any post-referral data. Data on how many of those referred from screening to ophthalmology that were not treated because screening had over-referred would have been invaluable for validating the model predicted error rates or could have been used in the model itself to improve estimation. The evaluation of screening did not include balancing the harms and benefits from being false positive and false negative. False positives are unlikely to produce any physical harm (although there is little in the way of evidence for this) but false negatives are potentially devastating, leading to irreversible vision loss. Future work in this area should extend the models presented here to include benefits and harms of over and under referral, the potential costs to the health service and the long-term implications for patients.

Final remarks and indications for future research

Future Research

I have come to the conclusion that there are two ways in which future research could go. One of these involves developing the statistical modelling and decision analytic approach, improving the way we synthesise evidence from a wide range of sources, find better data sources to model the true course of disease and construct comprehensive decision analytic models that fully assess cost-effectiveness. The other pathway is the interventionist approach. This involves

testing the effectiveness of monitoring in novel interventional trials. Both have strengths and limitations, which I will now discuss.

The statistician George Box said, "To find out what happens to a system when you interfere with it, you need to interfere with it (and not passively observe it)" [134], he also is quoted as saying that, "all models are wrong, some are useful" .... Mathematical models of biological systems can work well but can go wrong when important detail has been omitted or misunderstood; you cannot model what you don't know.

The outcome of the Cardiac Arrhythmia Suppression Trial (CAST) made me think differently about how far statistical modelling can be useful in providing evidence for clinical interventions and practice. CAST was a trial designed to evaluate the effects of antiarrythmic therapy in people with asymptomatic or mildly symptomatic ventricular arrythmia after myocardial infarction [135]. Patients were randomised to receive either active therapy or matching placebo. They were monitored using ambulatory ECG (Holter monitors). Gilbert Welch, was a medical resident whose role was to respond to the monitor's findings [1]. He recalls that,

> The patients were well, but their monitors were always going off. It seemed as though every one of them had an abnormality. And we were treating them. We were constantly adjusting medications to try to find the right drug or the right dose to make the rhythm normal.

The trial was stopped prematurely because there was an excess of deaths (two and half times higher) in the active therapy group compared to the placebo group. This in itself, is not direct evidence that monitoring is harmful in this case, but if pro-arrhythmic effects of the drug had a dose-response relationship, then the numerous false alarms leading to increased doses of medication would have contributed to the excess deaths. However, the expectation prior to the trial was that the therapy would lead to a 30% reduction in mortality, and the study team had planned to test this hypothesis using a one-sided test, so

sure were they that such treatment could only lead to benefit. This optimism persisted even after the first trial failed but must surely have been abated when CAST 2; a similar protocol but with a different drug and in high risk patients, produced the same result [1]. The CAST trial is a good example of where modelling on the basis of pre-existing evidence and common sense would have led to wrong conclusions.

There are however, a number of obstacles to the interventionist route. Firstly, it is difficult to generate interest and patronage for existing and routine practice that "*everyone knows*" works. There will also be resistance to randomising people to "less" or "no" monitoring as the presumption of benefit will make it appear unethical to withhold it. This problem was identified by Wilson and Junger [37] in their treatise on screening. They said that "if the use of drugs for mild hypertension became general, it would probably no longer be thought ethical to randomise treatment"[1]. Lastly, there is the problem of scale. Whilst this has not been rigorously researched for monitoring strategies specifically, a theoretical paper on planning clinical trials to evaluate early detection programs investigated sample sizes needed to detect a mortality difference between a control and screening group. The sample sizes required differed dramatically, depending on number of exams and study length but were all in the order of tens of thousands. For a trial with five year follow up, a sample size of 75,000 would be required to attain power equivalent to standard trials ($\sim 80\%$) [137].

The alternative route is to develop and optimise the modelling approach. By optimise, I mean, ensuring that the models and the assumptions within them are robust and represent accurately all of the different aspects of monitoring in practice. In addition, the effectiveness should be measured, not only by indirect measures (false positives, false negatives, or biochemical changes)

[1] All though there have been RCT's, some forty years later, a Cochrane review [136] could find no evidence that drug therapy reduced mortality, stroke or heart disease in people with mild hypertension and concluded that more research is needed

of benefit, but also in *hard* outcomes. This will often mean that a full cost-effectiveness analysis will need to be done to accurately model costs, benefits and harms. The natural history of the underlying condition needs to be understood and described accurately by the model. Data for the progression of disease, unaltered by intervention can be sourced from placebo arms of control trials (although, these will often not have enough repeat measurements), historical cohort studies (before treatment was mandated) or in some exceptional cases, clinical databases. The models used should be able to not only separate signal from noise or pathological change from physiological variations but should also be configured to allow for the fact, not all of the early pathological changes will lead to consequential disease. In both, microalbuminuria and retinopathy, early changes did not imply an inexorable path to clinical disease.

In practice

Realistically, both approaches to building the evidence base for monitoring and screening will co-exist and the way in which the evidence accumulates will depend on the disease and the expertise or skills of the team's leading the research. Bossuyt [25] has suggested a staged evaluation for the development of the evidence base for monitoring strategies. In Bossuyt's system the most appropriate design depends on the level of evidence. He suggests that there are four distinct phases in building the evidence base for monitoring and designates a study design to each one; The first phase consists of identifying a risk marker or early

Phase	Aim	Design
1	Collect evidence to justify monitoring	Various
2	Identify the most effective strategies	Modelling
3	Determine empirical estimates of effectiveness	RCT's
4	Control quality of monitoring	Surveillance

Table 6.1: *Bossuyt's staged evaluation for the development of monitoring. Adapted from [25]*

stage which predicts the outcome, finding a treatment or intervention that is

effective and a suitable test to separate those at high risk and those at low risk. Cost-effectiveness may follow and inform an RCT and if shown to be effective, the monitoring strategy will be rolled out and its delivery monitored (quality checked). Bossuyt's scheme represents an idealistic view of how the research should progress. In reality, it is very different. Depending on the actual clinical area, the evidence for monitoring has progressed in a variety of ways. For example, Dinnes *et al* reviewed the evidence for prostate-specific antigen (PSA) monitoring after radical treatment for prostate cancer. They found that guideline recommendations were inconsistent, and the scheduling of examinations was based on convenient time-frames rather than evidence and only one guideline cited a primary study in support of its recommendations for the interval of monitoring [138]. Clearly, PSA monitoring is stuck in phase 1 of Bossuyt's scheme. This is in contrast to the way evidence has accumulated for the monitoring carcinoembryonic antigen (CEA) for recurrence of colorectal cancer. In the seventies, it was shown that CEA detected the asymptomatic recurrence of colorectal cancer, and in 1981 the National Institute for Health called for a trial of CEA. A large randomised trial was started but never published until very recently [139], but even before the trial had finished, CEA testing had become commonplace [139]. Its usefulness is still disputed and modelling different strategies of CEA testing is an active area of research. CEA monitoring includes evidence from phases 1 and 3 but 2 (modelling has been omitted, up to now). Monitoring of ACR in adolescents with type 1 diabetes is different again. In this case, it is recommended that ACR is monitored but there is no consensus as to the most appropriate management of subjects when microalbuminuria is detected and research is currently under way to assess benefit [88].

The case against less monitoring

All the findings from the statistical evaluation of screening and monitoring show that longer intervals between monitoring tests are associated with more effec-

tive monitoring, reducing false positives tests, resource use and over-treatment. However, in practice, monitoring intervals are more likely to based upon pragmatic considerations. For political and economic reasons less frequent monitoring, irrespective of whether it is more effective on average, may still be unacceptable to either patient or clinician. Firstly, monitoring of many chronic conditions in the UK is incentivised by the quality outcomes framework (QOF) and general practitioners are paid to test and monitor their patients. For example, doctors in the UK are currently paid according to the percentage of patients in whom their last IFCC-HbA1c is 59 mmol/mol in the preceding 15 months [140]. This will undoubtedly influence the timing of monitoring HbA1c making it unlikely that GP's will wait more than 15 months to remeasure. The QOF incentives could in fact lead to more frequent monitoring of people who are well-controlled and less frequent monitoring of those people unlikely to meet the target if GP's *game* the system. This would be exactly the opposite of what the current evidence says should happen [9].

Less frequent monitoring may lead to the perception that control of their condition is of little concern to the caregiver or clinician and subsequently they may also take a more relaxed attitude to their condition. Frequent contact, though not necessarily to test, but to reinforce the importance of diet and exercise and the importance of adhering to the medication may lead to better outcomes. However, the opposite could also be true, and more contact and more *nagging* may not necessarily produce a positive response in the patient. There is certainly a point at which the intensity of management becomes a burden. An example of this comes from a paper by May and colleagues [141].

> A man being treated for heart failure in UK primary care rejected the offer to attend a specialist heart failure clinic to optimise management of his condition. He stated that in the previous two years he had made 54 visits to specialist clinics for consultant appointments, diagnostic tests, and treatment. The equivalent of one full day every two weeks was devoted to this work.

"The work of being a patient" includes much more than just remembering to take the medication, it also includes organising visits to the doctors and co-ordinating care between different parts of the health service. In a study of 8661 people with diabetes and 63 provider groups, although more intensive disease management improved processes of care (including retinal examinations, renal screening, etc), they did not lead to improvements in intermediate outcomes, such as systolic blood pressure, serum cholesterol and HbA1c [2].

Finally, there is the issue of litigation. Take for example, screening for diabetic retinopathy. Suppose a person attends retinal screening and no retinopathy is found, the person is asked to attend again in three year's time. In the meantime, the patient experiences a sudden and irreversible loss of sight in one eye. The plaintiff could argue that the failure to detect pre-proliferative retinopathy (which would have led to closer monitoring and later sight saving therapy) was because the screening interval was too long and thus the screening service was negligent. I really don't know if this would actually stand up in a court of law, it is beyond my expertise but clinical negligence cases involving failures to order tests to detect cancer are not uncommon and may represent the thin end of the wedge.

Final remarks

Monitoring, or more generally surveillance, strategies as with other aspects of clinical care, should be based upon solid evidence rather than opinion and optimism. In theory, they should not be adopted unless it can be shown that they improve and maintain patient health and that they do so in reasonable balance with costs[4]. Monitoring when the signal is weak against a background of noise from short-term biological or analytical error is prone to over-detection, leading to more *medicine*, more costs and more burden on health care systems.

[2]Except for the structured care management comparison in which HbA1c was significantly lower (4 mmol/mol)

However there are limitations to how far mathematical models can guide monitoring intervals. In practice, decisions about monitoring are based on more than just considerations of effectiveness. There are many other forces at play. Pressures to test and monitoring often stem from political interference, fear of litigation, being named and shamed, and general defensive approaches to modern medicine. The task of changing basic clinical practice is a formidable one and should not be under-estimated. Studies such as these add to the growing body of evidence identifying the potential problems over-testing and over-treating and present the first step towards finding solutions.

Appendices

Appendix A

Model estimation using hierarchical models

A.1 Hierarchical models

All of the random effects models used in this thesis can be formulated and estimated using a hierarchical model approach. The term hierarchy comes from the way these models are constructed in layers or levels. In the first level, we have measurements on an individual, at the second level we have individuals as part of a population and a third level we have hyper-parameters which represent the uncertainty in the second level. Although, hierarchical models are more complex and as a result more time-consuming they are flexible and can be adapted to a wide range of outcomes and data structures. Moreover, uncertainty estimates can be easily found for any transformations or combinations of the model parameters, which means confidence or credible intervals for many of the outcomes relating to the evaluation of monitoring can be computed relatively easily.

Hierarchical models are bayesian in origin, but can be used with relatively informative priors aligning them closely with equivalent models using maximum likelihood or restricted maximum likelihood. I have used MCMC based programs, BUGS [74] and later JAGS [75] to estimate hierarchical models and I will describe how to go from hierarchical equations to model code. To illustrate how the model is formulated and later estimated, I use the example of

the linear model with increasing variance.

Hierarchical Model for Cholesterol

The increasing variance model used to evaluate frequency of monitoring cholesterol in people with dyslipidemia used in Glasziou *et al* [5] can be re-formulated as a hierarchical model. The first level of the model constitutes the measurement of cholesterol on an individual. The second level the population of people with dysplidemia and the third level the uncertainty around the estimates of the population mean and standard deviation. Glasziou *et al.* considered cholesterol tests repeated on an individual to be normally distributed around a true mean μ with short-term variation or noise σ_W^2, hence. In the hierarchical model we write;

$$Chol_t \sim N(\mu, \sigma_w^2) \tag{A.1}$$

where the true mean μ is a linear function of the baseline value T and rate of change β; hence;

$$\mu = T + \beta t \tag{A.2}$$

These two equations could have been expressed as a single equation; $Chol \sim N(T + \beta, \sigma_w^2)$ but they are seperated due to the way the model estimation code is written. The first line of code, indexing over each data point i is written verbatim as;

```
Chol[i] ~ N(mu[i], tau.omega)
```

with the exception of the parameter tau.omega in place of sigma.omega. This is because the MCMC routines parameterise normal distributions in terms of the mean and precision (1/variance). Later on, code is written to invert the 'tau' parameter to a standard deviation. The second line of code is again written in a similar fashion to the equation, as;

```
mu[i] = T[id[i]] + beta[id[i]]*t[i]
```

The code used here assumes that the data is in a long format with individuals contributing multiple rows with an indicator (id) numbered from 1:J, where J is number of people in the data set. I assume that a data set has $j = 1, 2, \ldots, J$ people each measured at $i = 1, 2, \ldots, t$ times and that there are n observations in total. It is not necessary for the number n of observations to be the same for each person nor at regular intervals, nor for n ot t to be of a certain size. Missing outcome data is imputed directly using the model structure, but missing covariate data cannot be included unless given probability models. Objects in heirarchical model code are either modelled data (such as 'Chol') or unmodelled (such as 't'), modelled parameters (such as 'T') or unmodelled parameters, derived quantities ('mu') or looping indexes (such as 'i'). Modelled data is uncertain and requires a probability model, whereas unmodelled data is assumed fixed and known and does not need a probability model. Modelled parameters also have probability model and unmodelled parameters do not. Derived quantities are often used to link parts of the model. The model code for the first level is complete by writing a loop function over the all the 'n' observations in the data. therefore;

```
for (i in 1:n){
    Chol[i] ~ N(mu[i], tau.omega)
    mu[i] = T[id[i]] + beta[id[i]]*t[i]
}
```

The second level of the model describes the population distribution. It is assumed that each true baseline cholesterol 'T' is drawn from a population distribution with mean 'mu.T' and standard deviation 'sigma.T'. It is also assumed that the gradients 'beta' are also drawn from a normal distribution but with mean 'mu.beta' and standard deviation 'sigma.beta'. This is written algebrically as;

$$\begin{aligned} T_i &\sim N(\mu_T, \sigma_T^2) \\ \beta_i &\sim N(\mu_\beta, \sigma_\beta^2) \end{aligned}$$

and is coded in the MCMC code as;

```
for (j in 1:J){
T[j] ~ dnorm(mu.T, tau.T)
beta[j] ~ dnorm(mu.beta,tau.beta)
}
```

where again, the precision is used in place of the variance in the specification of the normal distribution. The third level of the model is used to complete the model. Each of the modelled parameters, must be given "prior" distributions. Starting with the precision terms, we first convert them to standard deviations by doing

```
tau.omega <- 1/sqrt(sigma.omega)
sigma.beta <- 1/sqrt(sigma.beta)
sigma.T <- 1/sqrt(sigma.T)
```

and then give the standard deviation parameters uniform priors;

```
sigma.omega ~ dunif(0,100)
sigma.beta ~ dunif(0,100)
sigma.T ~ dunif(0,100)
```

These uniform priors give equal weight to values of the standard deviation in the range 0 to 100 and 0 otherwise. If the marker being modelled has a large standard deviation, this upper bound can easily be made much bigger. I choose uniform priors for the standard deviations rather than gamma distributions for the precisions as they are less prone to bias [142]. Priors are also needed for the mean parameters, these are given uniformative normal distributions, with mean 0 and small precision (= large variance), which also gives equal weight to all realistic values for these parameters;

```
mu.T ~ dnorm(0,0.0001)
mu.beta ~ dnorm(0,0.0001)
```

Appendix B

JAGS MCMC model code

The following code was used to estimate the four main hierarchical models used in this thesis. Each are designed to be used as the *model.file* argument in a *jags* model in R. They can be saved as plain text files with a *.bug* extension.

They assume that the data is stored in the form of a list such as

```
data <- list(n = n,J = J, id = patient ,x = x,y = y)
```

where n is the length of the response vector y, J is the number of unique *ids'* in the *patient* variable and x is a vector of times at which y is observed. Starting values for the variance parameters are supplied to the *jags* call as a function;

```
inits <- function(){
  list(sigma.a = runif(1),sigma.b = runif(1))}
```

In addition, a character vector is used to tell *jags* which parameters to save;

```
parameters <- c("mu.b","sigma.a","sigma.b")
```

Finally, the call to jags is completed by specifying the number of chains, the number of iterations and the thinning rate of the chain, i.e.;

```
ivmodel.jags <- jags(data = data, jags.seed=1,inits=inits,
model.file="ivmodel.bug",n.chains=3,parameters.to.save=parameters,
n.burnin=1000,n.iter = 15000,n.thin=5)
```

B.1 Flexible variance model

```
model{

  for (i in 1:n){
    y[i] ~ dnorm(y.hat[i], tau.omega)
    y.hat[i] <- b[patient[i],1] + b[patient[i],2]*x[i]
  }
  # the residual variance
  sigma.omega ~ dunif(0,100)
  tau.omega <- pow(sigma.omega,-2)

  for (j in 1:J){
     b[j,1:2] ~ dmnorm(mu.b[1:2], Tau.B[,])
     }

  # The group level variance
  Tau.B[1:2,1:2] <- inverse(Sigma.B[,])
  Sigma.B[1,1] <- pow(sigma.a,2)
  Sigma.B[2,2] <- pow(sigma.b,2)
  Sigma.B[2,1] <- rho*sigma.a*sigma.b
  Sigma.B[1,2] <- Sigma.B[2,1]

  # priors
  sigma.a ~ dunif(0,100)
  sigma.b ~ dunif(0,100)
  rho ~ dunif(-1,1)
  mu.b[1] ~ dnorm(0,0.001)
  mu.b[2] ~ dnorm(0,0.001)
  }
```

B.2 T model

```
model
{
    for (i in 1:n) {
        y[i] ~ dnorm(y.hat[i], tau.omega[id[i]])
        y.hat[i] <- beta[id[i],1] + beta[id[i],2] * x[i]
    }
    a <- nu.omega/2
    b <- (nu.omega/2) * pow(sigma.omega, 2)
    sigma.omega ~ dunif(0.00000E+00, 100)
    nu.omega <- 1/nu.inv.omega
    nu.inv.omega ~ dunif(0.00000E+00, 1)
    for (j in 1:J) {
        beta[j, 1:2] ~ dmnorm(mu.b[1:2], Tau.B[, ])
        tau.omega[j] ~ dgamma(a, b)
        # random effects
        rfa[j] <- beta[j,1]
        rfb[j] <- beta[j,1]
    }

    ### Variance Level 2
    Tau.B[1:2, 1:2] <- inverse(Sigma.B[, ])
    Sigma.B[1, 1] <- pow(sigma.a, 2)
    Sigma.B[2, 2] <- pow(sigma.b, 2)
    Sigma.B[2, 1] <- rho * sigma.a * sigma.b
    Sigma.B[1, 2] <- Sigma.B[2, 1]

    ### priors
```

```
    sigma.a ~ dunif(0.00000E+00, 100)
    sigma.b ~ dunif(0.00000E+00, 100)
    rho ~ dunif(-1, 1)
    mu.b[1] ~ dnorm(0.00000E+00, 1.00000E-04)
    mu.b[2] ~ dnorm(0.00000E+00, 1.00000E-04)

   }
```

B.3 Binary flexible variance model

```
model{
  for (i in 1:n) {
    y[i] ~ dbern(p[i])
    logit(p[i]) <- b[id[i],1] + b[id[i],2] * x[i]
  }
  for (j in 1:J){
    b[j,1:2] ~ dmnorm(mu.b[1:2], Tau.B[,])
  }

  # The group level variance
  Tau.B[1:2,1:2] <- inverse(Sigma.B[,])
  Sigma.B[1,1] <- pow(sigma.a,2)
  Sigma.B[2,2] <- pow(sigma.b,2)
  Sigma.B[2,1] <- rho*sigma.a*sigma.b
  Sigma.B[1,2] <- Sigma.B[2,1]

  # priors
  sigma.a ~ dunif(0,100)
  sigma.b ~ dunif(0,100)
  rho <- 0
```

```
  # rho ~ dunif(-1,1) # change to estimate rho
  mu.b[1] ~ dnorm(0,0.001)
  mu.b[2] ~ dnorm(0,0.001)
}
```

B.4 Hierarchical ordinal model

```
model {
    for (i in 1:N) {
        mu[i] <- a[id[i]] + b[id[i]]*t[i] + gamma*dur[i] +
        lambda*Mac1[i] + delta*Mac2[i] + kappa*hba1c[i]
        logit(Q[i,1]) <- c[1] - mu[i]
        logit(Q[i,2]) <- c[2] - mu[i]
        logit(Q[i,3]) <- c[3] - mu[i]

        P[i,1] <- Q[i, 1]
        P[i,2] <- Q[i,2] - Q[i,1]
        P[i,3] <- Q[i,3] - Q[i,2]
        P[i,4] <- 1 - Q[i,3]
        y[i] ~ dcat(P[i,])
    }

for (j in 1:m) {
        a[j] ~ dnorm(0,tau.a)
        b[j] ~ dnorm(mu.b, tau.b)
    }
    ## thresholds
    for(j in 1:3){
    c0[j] ~ dnorm(0, .01)}
```

```
c[1:3] <- sort(c0)

tau.a <- pow(sig.a,-2)
sig.a ~ dunif(0,40)
mu.b ~ dnorm(0.00000E+00,0.001)
gamma ~ dnorm(0,0.001)
tau.b <- pow(sig.b,-2)
sig.b ~ dunif(0,40)
lambda ~ dnorm(0,0.001)
delta ~ dnorm(0,0.0001)
kappa ~ dnorm(0,0.0001)
}
```

Appendix C

Variogram Method

The origins of the variogram are as a tool to describe spatial variation and correlation and has been used extensively in geostatistics for more than 40 years [143]. Diggle [68] defines the variogram $\gamma(u)$ for a stochastic process Y as;

$$\gamma(u) = \frac{1}{2}E\left[\{Y(t) - Y(t-u)\}^2\right] \tag{C.1}$$

Haslett [143] states that Diggle's definition of the variogram is incorrect and should be;

$$var(Y_t - Y_{t-u}) = 2\gamma(u) \tag{C.2}$$

hence the variogram is twice the variance of the change in Y at time t and Y at time $t-u$;

$$\gamma(u) = \frac{1}{2}\mathrm{Var}(Y_t - Y_{t-u}) \tag{C.3}$$

The difference is subtle but important depending on the assumed form of Y. Haslett and Diggle do agree that the variogram is not a useful tool to describe correlation when a process does not have constant variance. This is because as the variogram is linked intrinsically to the correlation function $\rho(u)$ by;

$$\gamma(u) = Var\left\{1 - \rho(u)\right\} \tag{C.4}$$

then the variogram becomes a useful way to estimate the correlation over time as long as the variance is stationary (constant or stable). However, the direct method does not use the variogram for its original intended purpose; to describe

correlation but it is used as a way to estimate the short-term variation (σ_W^2). This is done by calculating a special case of the variogram, namely, only the changes from time zero and not for all the possible pairs of measurements. This is because the modified variogram evaluated at time $t = 0$ is equal to the short-term variance;

$$\gamma(t = u, u \to 0) = \frac{1}{2}\text{Var}(Y_0 - Y_{u\to 0}) = \sigma_w^2 \tag{C.5}$$

This rule follows for a number of statistical models for repeated measures data that have independent measurement errors. This is due to the fact that the covariance between changes from baseline is equal to the variance of Y and hence cancel each other out leaving only the measurement error. This still holds for models with many random factors because they also cancel out. For example, a model with average baseline μ_α and standard deviation σ_α,average slope μ_β with standard deviation σ_β^2, and with slopes and intercepts co-varying by an amount ϕ then the variance at time t is;

$$V_t = \sigma_\alpha^2 + \sigma_\beta^2 t^2 + 2\phi t + \sigma_W^2$$

and the variance at time $t = 0$ is;

$$V_0 = \sigma_\alpha^2 + \sigma_W^2$$

and the covariance between two time points 0 and t is given by;

$$Cov(Y_0, Y_t) = \sigma_\alpha^2 + \sigma_\beta^2 0t + \phi(0 + t) = \sigma_\alpha^2 + \phi t$$

then the variogram or half the squared variance of change $\gamma(t)$ is equal to;

$$\begin{array}{ll} \gamma(t) = & \frac{1}{2}(V_0 + V_t - 2\text{Cov}(Y_0, Y_t)) \\ = & \frac{1}{2}\left\{\sigma_\alpha^2 + \sigma_W^2 + \sigma_\alpha^2 + \sigma_\beta^2 t^2 + 2\phi t + \sigma_W^2\right\} - 2\left\{\sigma_\alpha^2\right\} \\ = & \frac{1}{2}\left\{2\sigma_\alpha^2 + \sigma_\beta^2 t^2 + 2\sigma_W^2 - 2\sigma_\alpha^2 + 2\phi t\right\} \\ = & \frac{1}{2}\left\{\sigma_\beta^2 t^2 + 2\phi t + 2\sigma_W^2\right\} \end{array} \tag{C.6}$$

and when $t = 0$, all terms involving t cancel leaving $\gamma(0) = \sigma_w^2$. This works for models without random slopes (the random intercept model) and models with more random effects because they cancel out in the calculation.

It will often be the case that measurements taken in close proximity will not be available and hence estimating the variogram at $t = 0$ will not be possible directly from the data. Glasziou *et al* suggested that the variogram be calculated for all time points available and then extrapolate $\gamma(0)$ using linear regression. Hence,

1. estimate the variogram $\gamma(t)$ for all the time points $t = 1, 2, \ldots, n$
2. regress the variogram on t i.e. $\gamma(t) = \alpha + \beta t$
3. take α to be $\gamma(0)$.

I have found that the choice of regression method is relevant here. For a number of candidate models, the variogram will not be a linear function of time, and hence a standard linear regression my not be appropriate. Figure C.1 shows the theoretical variograms for a range of candidate models. Figure C.1 suggests

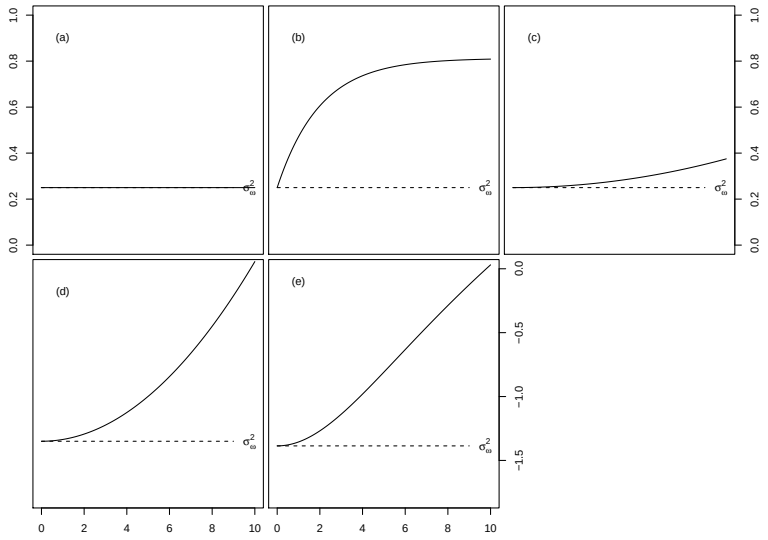


Figure C.1: *Variograms for a range of plausible models for longitudinal marker data. (a) a random intercept model, (b) random intercept with serial correlation, (c) random slope with small σ^2_β (d) random slope with larger s σ^2_β, (e) log variogram of model (d)*

that a linear model could be used when the between-person variation in slopes is small, but a log transformation of the variogram is more suitable when there is a

greater degree of between person slope variation ($\sigma^2_\beta > 0$). For serial correlation models, a non-parametric smoother (such as lowess) may be appropriate but cannot be used to extrapolate to zero or values of x for which there is no data. It may still be useful to assess the magnitude of the difference between the lowess estimate at the lowest value of x with the extrapolated linear regression estimate.

Appendix D

Publications from the thesis

The work in Chapter 2 of this thesis was published in Diabetes Research and Clinical Practice, 99(3), 30714 [62]

Work from Chapter 2 and 3 contributed to a HTA monograph entitled *Optimal strategies for identifying kidney disease in diabetes: properties of screening tests, progression of renal dysfunction and impact of treatment - systematic review and modelling of progression and cost-effectiveness.* [76]

Work on evaluating diabetic retinopathy screening (Chapter 4 and 5) is being written up as a chapter in the forthcoming HTA monograph (in preparation) *Development of a cost-effectiveness model for optimization of the screening interval in diabetic retinopathy screening*, Scanlon *et al.*

Bibliography

[1] Welch DHG, Schwartz DLM, Woloshin DS. *Over-diagnosis: Making People Sick in the Pursuit of Health*. Boston: Beacon Press; 2012.

[2] Black WC, Welch HG. Screening for disease. *American Journal of Roentgenology*. 1997 Jan;168(1):3–11.

[3] Glasziou P, Irwig L, Mant D. Monitoring in chronic disease: A rational approach. *BMJ (Clinical research ed)*. 2005 Mar;330(7492):644–8.

[4] Glasziou P, Irwig L, Aronson J. *Evidence-Based Medical Monitoring*. 1st ed. Blackwell Publishing; 2008.

[5] Glasziou P, Irwig L, Heritier S, Simes RJ, Tonkin A. Monitoring cholesterol levels: measurement error or true change? *Annals of Internal Medicine*. 2008 May;148(9):656–61.

[6] Cooper A, Nherera L, Calvert N, OFlynn N, Turnbull N, Robson J et al . *Clinical Guidelines and Evidence Review for Lipid Modification: Cardiovascular risk assessment and the primary and secondary prevention of cardiovascular disease*. Royal College of General Practioners; 2008.

[7] Doll H, Shine B, Kay J, James T, Glasziou P. The rise of cholesterol testing: how much is unnecessary. *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*. 2011 Feb;61(583):e81–8.

[8] Keenan K, Hayen A, Neal BC, Irwig L. Long term monitoring in patients receiving treatment to lower blood pressure: analysis of data from placebo controlled randomised controlled trial. *BMJ*. 2009 Apr;338(30 1):b1492.

[9] Oke JL, Stevens RJ, Gaitskell K, Farmer AJ. Establishing an evidence base for frequency of monitoring glycated haemoglobin levels in patients with Type 2 diabetes: projections of effectiveness from a regression model. *Diabetic Medicine : A Journal of the British Diabetic Association*. 2012 Feb;29(2):266–71.

[10] Hodgkinson J, Mant J, Martin U, Guo B, Hobbs FDR, Deeks JJ, et al. Relative effectiveness of clinic and home blood pressure monitoring compared with ambulatory blood pressure monitoring in diagnosis of hypertension: systematic review. *BMJ*. 2011 Jun;342(jun24 1):d3621–d3621.

[11] Inside Health, BBC Radio 4. Overdiagnosis: Chronic Kidney Disease;. Available from: http://www.bbc.co.uk/programmes/b01lt2vn.

[12] Kamerow D. How to decrease overtreatment in cancer. *BMJ (Clinical research ed)*. 2013 Jan;347(aug14_5):f5071.

[13] Farmer A, Wade A, Goyder E, Yudkin P, French D, Craven A, et al. Impact of self monitoring of blood glucose in the management of patients with non-insulin treated diabetes: open parallel group randomised trial. *BMJ (Clinical research ed)*. 2007 Jul;335(7611):132.

[14] Heneghan C, Alonso-Coello P, Garcia-Alamino JM, Perera R, Meats E, Glasziou P. Self-monitoring of oral anticoagulation: a systematic review and meta-analysis. *Lancet*. 2006 Feb;367(9508):404–11.

[15] Ryan D, Price D, Musgrave SD, Malhotra S, Lee AJ, Ayansina D, et al. Clinical and cost effectiveness of mobile phone supported self

monitoring of asthma: multicentre randomised controlled trial. *BMJ (Clinical research ed)*. 2012 Jan;344:e1756.

[16] Cremer OL, van Dijk GW, van Wensen E, Brekelmans GJF, Moons KGM, Leenen LPH, et al. Effect of intracranial pressure monitoring and targeted intensive care on functional outcome after severe head injury. *Critical care medicine*. 2005 Oct;33(10):2207–13.

[17] Dyson DC, Danbe KH, Bamber JA, Crites YM, Field DR, Maier JA, et al. Monitoring Women at Risk for Preterm Labor. *New England Journal of Medicine*. 1998 Jan;338(1):15–19.

[18] Watkinson PJ, Barber VS, Price JD, Hann A, Tarassenko L, Young JD. A randomised controlled trial of the effect of continuous electronic physiological monitoring on the adverse event rate in high risk medical and surgical patients. *Anaesthesia*. 2006 Nov;61(11):1031–9.

[19] Moschetti I, Brandt D, Perera R, Clarke M, Heneghan C. Adequacy of reporting monitoring regimens of risk factors for cardiovascular disease in clinical guidelines: systematic review. *BMJ (Clinical research ed)*. 2011 Jan;342:d1289.

[20] Fink HA, Ishani A, Taylor BC, Greer NL, MacDonald R, Rossini D, et al. Screening for, monitoring, and treatment of chronic kidney disease stages 1 to 3: A systematic review for the U.S. Preventive Services Task Force and for an American College of Physicians Clinical Practice Guideline. *Annals of Internal Medicine*. 2012 Apr;156(8):570–81.

[21] Royal College of Physicians. National Collaborating Centre for Chronic Conditions. Type 2 diabetes: National clinical guideline for management in primary and secondary care; 2008.

[22] American Diabetes Association. Standards of medical care in diabetes–2013. Diabetes Care. 2013 Jan;36 Suppl 1(Supplement_1):S11–66.

[23] Gotzsche PC, Nielsen M. Screening for breast cancer with mammography. *The Cochrane Database of Systematic Reviews*. 2011 Jan;(1):CD001877.

[24] Mandelblatt JS. Effects of Mammography Screening Under Different Screening Schedules: Model Estimates of Potential Benefits and Harms. *Annals of Internal Medicine*. 2009 Nov;151(10):738.

[25] Bossuyt PMM. Evaluating the effectiveness and costs of monitoring. In: Glasziou PP, Irwig L, Aronson JK, editors. *Evidence-based medical monitoring*. 1st ed.; 2008. p. 362.

[26] Lincoln TL, Weiss GH. A Statistical Evaluation of Recurrent Medical Examinations. *Operations Research*. 1964 Apr;12(2):187–205.

[27] Barlow RE, Hunter LC, Proschan F. Optimum Checking Procedures. *Journal of the Society for Industrial and Applied Mathematics*. 1963 Dec;11(4):1078–1095 CR – Copyright © 1963 Society for I.

[28] Zelen M, Feinleib M. On the theory of screening for chronic diseases. *Biometrika*. 1969;56(3):601–614.

[29] Prorok PC. The Theory of Periodic Screening I: Lead Time and Proportion Detected. *Advances in Applied Probability*. 1976 Mar;8(1):127–143.

[30] Prorok PC. The Theory of Periodic Screening II: Doubly Bounded Recurrence Times and Mean Lead Time and Detection Probability Estimation. *Advances in Applied Probability*. 1976 Sep;8(3):460–476.

[31] Gardner MJ, Heady JA. Some Effects of Within-Person Variability in Epidemiological Studies. *Journal of Chronic Disease*. 1973;26:781–795.

[32] Wilson PD, Hebel JR, Sherwin R. Screening and diagnosis when within-individual observations are Markov-dependent. *Biometrics*. 1981 Sep;37(3):553–65.

[33] Johnson WD, George VT. Effect of regression to the mean in the presence of within-subject variability. *Statistics in Medicine*. 1991 Aug;10(8):1295–302.

[34] Thompson SG, Pocock SJ. The variability of serum cholesterol measurements: Implications for screening and monitoring. *Journal of Clinical Epidemiology*. 1990 Jan;43(8):783–9.

[35] Kimmel AD, Goldie SJ, Walensky RP, Losina E, Weinstein MC, Paltiel aD, et al. Optimal frequency of CD4 cell count and HIV RNA monitoring prior to initiation of antiretroviral therapy in HIV-infected patients. *Antiviral therapy*. 2005 Jan;10(1):41–52.

[36] JM Burr, P Botello-Pinzon YT, R Hernandez, M Vazquez-Montes AE, R Asaoka, K Banister JvdS, C Fraser, A King, H Lemij, R Sanders, S Vernon, A Tuulonen, A Kotecha, P Glasziou, D Garway- Heath, D Crabb, L Vale AAB, R Perera, M Ryan JD, Cook J. Optimal frequency of monitoring of intraocular pressure and visual fields. *Health Technology Assessment*. 2012 Jun;16(29).

[37] Wilson J, Junger G. *Principles and practice of screening for disease*. Royal College of General Practitioners; 1968.

[38] Macaskill P. Control charts and control limits in long-term monitoring. In: Glasziou, Paul P , Irwig, Les, Aronson, Jeffery K, editor. *Evidence-based medical monitoring*. 1st ed.; 2008. p. 90–101.

[39] Mant D. A framework for developing and evaluating a monitoring strategy. In: Glasziou, Paul P , Irwig, Les, Aronson, Jeffery K, editor. *Evidence-based medical monitoring*. 1st ed.; 2008. p. 382.

[40] Dunger DB, Schwarze CP, Cooper JD, Widmer B, Neil HAW, Shield J, et al. Can we identify adolescents at high risk for nephropathy before the development of microalbuminuria? *Diabetic Medicine : A Journal of the British Diabetic Association*. 2007 Feb;24(2):131–6.

[41] Klein R, Klein BEK, Moss SE, Davis MD, DeMets DL. The Wisconsin Epidemiologic Study of Diabetic Retinopathy: II. Prevalence and Risk of Diabetic Retinopathy When Age at Diagnosis Is Less Than 30 Years. *Archives of Ophthalmology*. 1984 Apr;102(4):520–526.

[42] Kilpatrick ES. Haemoglobin A1c in the diagnosis and monitoring of diabetes mellitus. *Journal of Clinical Pathology*. 2008 Sep;61(9):977–82.

[43] Perkins BA, Ficociello LH, Silva KH, Finkelstein DM, Warram JH, Krolewski AS. Regression of Microalbuminuria in Type 1 Diabetes. *New England Journal of Medicine*. 2003 Jun;348(23):2285–2293.

[44] Levy J, Atkinson AB, Bell PM, McCance DR, Hadden DR. Beta-cell deterioration determines the onset and rate of progression of secondary dietary failure in type 2 diabetes mellitus: the 10-year follow-up of the Belfast Diet Study. *Diabetic Medicine : A Journal of the British Diabetic Association*. 1998 Apr;15(4):290–6.

[45] Jackson CH, Sharples LD. Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*. 2002 Jan;21(1):113–28.

[46] Chen HH, Duffy SW, Tabar L. A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer

screening. *Journal of the Royal Statistical Society Series D (The Statistician)*. 2012;45(3).

[47] Guihenneuc-Jouyaux C, Richardson S, Longini IM. Modeling markers of disease progression by a hidden Markov process: application to characterizing CD4 cell decline. *Biometrics*. 2000 Sep;56(3):733–41.

[48] Keiding N, Klein JP, Horowitz MM. Multi-state models and outcome prediction in bone marrow transplantation. *Statistics in Medicine*. 2001 Jun;20(12):1871–85.

[49] Titman AC. *Model diagnostics in multi-state models of biological systems*; 2007.

[50] Saint-Pierre P, Combescure C, Daurès JP, Godard P. The analysis of asthma control under a Markov assumption with use of covariates. *Statistics in Medicine*. 2003 Dec;22(24):3755–70.

[51] Shuster S. Malignant melanoma: how error amplification by screening creates spurious disease. *The British Journal of Dermatology*. 2009 Nov;161(5):977–9.

[52] Diggle PJ. Estimating prevalence using an imperfect test. *Epidemiology Research International*. 2011;2011:1–5.

[53] Goldberg JD, Wittes JT. The estimation of false negatives in medical screening. *Biometrics*. 1978 Mar;34(1):77–86.

[54] The Lancet. The screening muddle. The Lancet. 1998 Feb;351(9101):459.

[55] Petticrew MP, Sowden AJ, Lister-Sharp D, Wright K. False negative results in screening programmes: systematic review of impact and implications. *Health Technology Assessment Programme*. 2000;4(5).

[56] The Diabetes Control and Complications Trial (DCCT) Study Group. Effect of intensive therapy on the development and progression of diabetic nephropathy in the Diabetes Control and Complications Trial. The Diabetes Control and Complications (DCCT) Research Group. *Kidney International*. 1995 Jun;47(6):1703–20.

[57] Warram JH, Gearin G, Laffel L, Krolewski AS. Effect of duration of type I diabetes on the prevalence of stages of diabetic nephropathy defined by urinary albumin/creatinine ratio. *Journal of the American Society of Nephrology*. 1996 Jun;7(6):930–7.

[58] Coresh J, Astor BC, Greene T, Eknoyan G, Levey AS. Prevalence of chronic kidney disease and decreased kidney function in the adult US population: Third National Health and Nutrition Examination Survey. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*. 2003 Jan;41(1):1–12.

[59] de Jong PE, Curhan GC. Screening, monitoring, and treatment of albuminuria: Public health perspectives. *Journal of the American Society of Nephrology : JASN*. 2006 Aug;17(8):2120–6.

[60] Parving HH, Lehnert H, Bröchner-Mortensen J, Gomis R, Andersen S, Arner P. The effect of irbesartan on the development of diabetic nephropathy in patients with type 2 diabetes. *The New England Journal of Medicine*. 2001 Sep;345(12):870–8.

[61] Palmer AJ, Annemans L, Roze S, Lamotte M, Lapuerta P, Chen R, et al. Cost-effectiveness of early irbesartan treatment versus control (standard antihypertensive medications excluding ACE inhibitors, other angiotensin-2 receptor antagonists, and dihydropyridine calcium channel blockers) or late irbesartan treatment in patien. *Diabetes Care*. 2004 Aug;27(8):1897–903.

[62] Oke J, Farmer A, Neil A, Dalton RN, Dunger D, Stevens R. Monitoring young people with type 1 diabetes for diabetic nephropathy: Potential errors of annual ACR testing. *Diabetes Research and Clinical Practice*. 2013 Mar;99(3):307–14.

[63] Husdan H, Rapoport A. Estimation of creatinine by the Jaffe reaction. A comparison of three methods. *Clinical Chemistry*. 1968 Mar;14(3):222–38.

[64] Schultz CJ, Konopelska-Bahu T, Dalton RN, Carroll TA, Stratton I, Gale EA, et al. Microalbuminuria prevalence varies with age, sex, and puberty in children with type 1 diabetes followed from diagnosis in a longitudinal study. Oxford Regional Prospective Study Group. *Diabetes Care*. 1999 Mar;22(3):495–502.

[65] Nagrebetsky A, Jin J, Stevens R, James T, Adler A, Park P, et al. Diagnostic accuracy of urine dipstick testing in screening for microalbuminuria in type 2 diabetes: a cohort study in primary care. *Family Practice*. 2013 Apr;30(2):142–52.

[66] Gaddum JH. Lognormal distributions. *Nature*. 1945;156(3964):463–466.

[67] Bland JM, Altman DG. Measurement error proportional to the mean. *BMJ (Clinical research ed)*. 1996 Jul;313(7049):106.

[68] Zeger L, Diggle P, Heagerty P, Liang KY, Scott. *Analysis of Longitudinal Data*. London: Oxford University Press; 2001.

[69] Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Boston: Cambridge University Press; 2007.

[70] Rasbash, J , Charlton, C , Browne, W J , Healy, M and Cameron, B. MLwiN Version 2.02. Centre for Multilevel Modelling, University of Bristol; 2005.

[71] Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: Linear and Nonlinear Mixed Effects Models; 2013. R package version 3.1-113.

[72] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria; 2012.

[73] Pinheiro J, Bates D. *Mixed Effects Models in S and S-Plus*. Springer; 2009.

[74] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*. 2000;10(4):325–337.

[75] Plummer M. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*; 2003.

[76] Farmer AJ, Stevens R, Hirst J, Lung T, Oke J, Clarke P, et al. Optimal strategies for identifying kidney disease in diabetes: properties of screening tests, progression of renal dysfunction and impact of treatment - systematic review and modelling of progression and cost-effectiveness. *Health Technology Assessment*. 2014 Feb;18(14):1–128.

[77] Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine*. 2000 Feb;19(4):453–73.

[78] Montgomery, Douglas C . *The Design and Analysis of Experiments*. 5th ed. Wiley; 2005.

[79] Kreft IGG, de Leeuw J, Aiken LS. The Effect of Different Forms of Centering in Hierarchical Linear Models. Multivariate Behavioral Research. 1995 Jan;30(1):1–21.

[80] Cowles MK, Carlin BP. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. Journal of the American Statistical Association. 1996 Jun;91(434):883–904.

[81] de Boer IH, Rue TC, Cleary PA, Lachin JM, Molitch ME, Steffes MW, et al. Long-term renal outcomes of patients with type 1 diabetes mellitus and microalbuminuria: an analysis of the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications cohort. Archives of internal medicine. 2011 Mar;171(5):412–20.

[82] Cho YH, Craig ME, Hing S, Gallego PH, Poon M, Chan A, et al. Microvascular complications assessment in adolescents with 2- to 5-yr duration of type 1 diabetes from 1990 to 2006. Pediatric diabetes. 2011 Dec;12(8):682–9.

[83] Orchard TJ, Secrest AM, Miller RG, Costacou T. In the absence of renal disease, 20 year mortality risk in type 1 diabetes is comparable to that of the general population: a report from the Pittsburgh Epidemiology of Diabetes Complications Study. Diabetologia. 2010 Nov;53(11):2312–9.

[84] Salgado PPCdA, Silva IN, Vieira EC, Simões e Silva AC. Risk factors for early onset of diabetic nephropathy in pediatric type 1 diabetes. Journal of pediatric endocrinology & metabolism : JPEM. 2010 Dec;23(12):1311–20.

[85] Andersen AR, Christiansen JS, Andersen JK, Kreiner S, Deckert T. Diabetic nephropathy in Type 1 (insulin-dependent) diabetes: an epidemiological study. Diabetologia. 1983 Dec;25(6):496–501.

[86] Alleyn CR, Volkening LK, Wolfson J, Rodriguez-Ventura A, Wood JR, Laffel LMB. Occurrence of microalbuminuria in young people with Type

1 diabetes: importance of age and diabetes duration. Diabetic medicine : a journal of the British Diabetic Association. 2010 May;27(5):532–7.

[87] Parving HH, Hommel E, Mathiesen E, Skø tt P, Edsberg B, Bahnsen M, et al. Prevalence of microalbuminuria, arterial hypertension, retinopathy and neuropathy in patients with insulin dependent diabetes. *British medical journal (Clinical research ed)*. 1988 Jan;296(6616):156–60.

[88] The Adolescent Type 1 Diabetes Cardio-Renal Intervention Trial Research Group. Adolescent type 1 Diabetes Cardio-renal Intervention Trial (AdDIT). *BMC Pediatrics*. 2009 Jan;9(1):79.

[89] Hirst JA, Taylor KS, Stevens RJ, Blacklock CL, Roberts NW, Pugh CW, et al. The impact of renin-angiotensin-aldosterone system inhibitors on Type 1 and Type 2 diabetic patients with and without early diabetic nephropathy. *Kidney International*. 2012 Apr;(7):674–83.

[90] Kiberd BA, Jindal KK. Screening to prevent renal failure in insulin dependent diabetic patients: an economic evaluation. *BMJ (Clinical research ed)*. 1995 Dec;311(7020):1595–9.

[91] National Institute for Health and Care Excellence. NICE clinical guideline 85: Diagnosis and management of ohronic open angle glaucoma and ocular hypertension; 2009. April.

[92] Satten GA, Longini Ira M J. Markov Chains With Measurement Error: Estimating the 'True' Course of a Marker of the Progression of Human Immunodeficiency Virus Disease. *Applied Statistics*. 1996;45(3):275–309.

[93] Borch-Johnsen K, Wenzel H, Viberti GC, Mogensen CE. Is screening and intervention for microalbuminuria worthwhile in patients with insulin dependent diabetes? *BMJ (Clinical research ed)*. 1993 Jun;306(6894):1722–5.

[94] Adler AI, Stevens RJ, Manley SE, Bilous RW, Cull CA, Holman RR. Development and progression of nephropathy in type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS 64). *Kidney International*. 2003 Jan;63(1):225–32.

[95] Watts GF, Kubal C, Chinn S. Long-term variation of urinary albumin excretion in insulin-dependent diabetes mellitus: some practical recommendations for monitoring microalbuminuria. *Diabetes Research and Clinical Practice*. 1990;(2):169–177.

[96] Forsblom CM, Groop PH, Ekstrand A, Groop LC. Predictive value of microalbuminuria in patients with insulin-dependent diabetes of long duration. *BMJ*. 1992 Oct;305(6861):1051–1053.

[97] Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet*. 2010 Jul;376(9735):124–36.

[98] Scanlon P, Wilkinson C, Aldingtion S, Matthews D. *A Practical Manual of Diabetic Retinopathy Management*. First edition ed. Wiley-Blackwell; 2009.

[99] Porta M, Bandello F. Diabetic retinopathy. A clinical update. *Diabetologia*. 2002 Dec;45(12):1617–34.

[100] Stefansson E, Bek T, Porta M, Larsen N, Kristinsson JK, Agardh E. Screening and prevention of diabetic blindness. *Acta Ophthalmologica Scandinavica*. 2000 Aug;78(4):374–385.

[101] Scanlon PH. The English national screening programme for sight-threatening diabetic retinopathy. *Journal of Medical Screening*. 2008 Jan;15(1):1–4.

[102] Moore K, Agarawal P, Edgar DF. Visual acuity screening for Diabetic Maculopathy. *Practical Diabetes*. 1985;2(6):30 – 32.

[103] Aldington SJ, Kohner EM, Meuer S, Klein R, Sjø lie AK. Methodology for retinal photography and assessment of diabetic retinopathy: the EURODIAB IDDM complications study. *Diabetologia*. 1995 May;38(4):437–44.

[104] Hutchinson A, McIntosh A, Peters J, O'Keeffe C, Khunti K, Baker R, et al. Effectiveness of screening and monitoring tests for diabetic retinopathy–a systematic review. *Diabetic Medicine : A Journal of the British Diabetic Association*. 2000 Jul;(7):495–506.

[105] Moss S, Meuer S, Klein R, Hubbard L, Brothers R, Klein B. Are seven standard photographic fields necessary for classification of diabetic retinopathy? *Invest Ophthalmol Vis Sci*. 1989 May;30(5):823–828.

[106] Ruamviboonsuk P, Teerasuwanajak K, Tiensuwan M, Yuttitham K. Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening. *Ophthalmology*. 2006 May;113(5):826–32.

[107] Sallam A, Scanlon PH, Stratton IM, Jones V, Martin CN, Brelen M, et al. Agreement and reasons for disagreement between photographic and hospital biomicroscopy grading of diabetic retinopathy. Diabetic Medicine : A Journal of the British Diabetic Association. 2011 Jul;28(6):741–6.

[108] Gillies MC, Sutter FKP, Simpson JM, Larsson J, Ali H, Zhu M. Intravitreal triamcinolone for refractory diabetic macular edema: two-year results of a double-masked, placebo-controlled, randomized clinical trial. *Ophthalmology*. 2006 Sep;113(9):1533–8.

[109] Titman AC, Sharples LD. Model diagnostics for multi-state models. *Statistical Methods in Medical Research*. 2010 Dec;19(6):621–51.

[110] Therneau T. *survival: A Package for Survival Analysis in S. R package version 2.37-4*; 2013.

[111] Kalm H, Egertsen R, Blohmé G. Non-stereo fundus photography as a screening procedure for diabetic retinopathy among patients with type II diabetes. Compared with 60D enhanced slit-lamp examination. *Acta ophthalmologica*. 1989 Oct;67(5):546–53.

[112] Stellingwerf C, Hardus PLLJ, Hooymans JMM. Two-Field Photography Can Identify Patients With Vision-Threatening Diabetic Retinopathy: A screening approach in the primary care setting. *Diabetes Care*. 2001 Dec;24(12):2086–2090.

[113] Schachat AP. Comparison of Diabetic Retinopathy Detection by Clinical Examinations and Photograph Gradings. *Archives of Ophthalmology*. 1993 Aug;111(8):1064.

[114] Olson JA, Strachan FM, Hipwell JH, Goatman KA, McHardy KC, Forrester JV, et al. A comparative evaluation of digital imaging, retinal photography and optometrist examination in screening for diabetic retinopathy. *Diabetic Medicine*. 2003 Jul;20(7):528–534.

[115] Scanlon PH, Malhotra R, Greenwood RH, Aldington SJ, Foy C, Flatman M, et al. Comparison of two reference standards in validating two field mydriatic digital photography as a method of screening for diabetic retinopathy. *The British journal of ophthalmology*. 2003 Oct;87(10):1258–63.

[116] Manjunatha RS, Manjunatha NP, Baskar V, Headon MP, Singh BM, Viswanath aK. Quality assurance in the diabetic retinopathy screening programme: evaluating the benefit of universal regrading of the normal primary grade. *Diabetic Medicine : A Journal of the British Diabetic Association*. 2012 Feb;29(2):287–8.

[117] Looker HC, Nyangoma SO, Cromie DT, Olson JA, Leese GP, Philip S, et al. Predicted impact of extending the screening interval for diabetic retinopathy: the Scottish Diabetic Retinopathy Screening programme. *Diabetologia*. Aug.;(8):1716–25.

[118] Stratton IM, Aldington SJ, Taylor DJ, Adler AI, Scanlon PH. A simple risk stratification for time to development of sight-threatening diabetic retinopathy. *Diabetes Care*. 2013 Mar;36(3):580–5.

[119] Agardh E, Tababt-Khani P. Adopting 3-year screening intervals for sight-threatening retinal vascular lesions in type 2 diabetic subjects without retinopathy. *Diabetes Care*. Jun.;(6):1318–9.

[120] Harding SP, Broadbent DM, Neoh C, White MC, Vora J. Sensitivity and specificity of photography and direct ophthalmoscopy in screening for sight threatening eye disease: the Liverpool Diabetic Eye Study. *BMJ (Clinical research ed)*. 1995 Oct;(7013):1131–5.

[121] van der Heijden AAWA, Walraven I, van't Riet E, Aspelund T, Lund SH, Elders P, et al. Validation of a model to estimate personalised screening frequency to monitor diabetic retinopathy. *Diabetologia*. 2014 Jul;57(7):1332–8.

[122] Pan SL, Wu HM, Yen AMF, Chen THH. A Markov regression random-effects model for remission of functional disability in patients following a first stroke: a Bayesian approach. *Statistics in Medicine*. 2007 Dec;26(29):5335–53.

[123] Spackman DE, Kadiyala S, Neumann PJ, Veenstra DL, Sullivan SD. Measuring Alzheimer disease progression with transition probabilities: estimates from NACC-UDS. *Current Alzheimer Research*. 2012 Nov;9(9):1050–8.

[124] Garg SK, Marshall G, Chase HP, Jackson WE, Archer P, Crews MJ. The use of the Markov process in describing the natural course of diabetic retinopathy. *Archives of Ophthalmology*. 1990 Sep;108(9):1245–7.

[125] Marshall G, Jones RH. Multi-state models and diabetic retinopathy. *Statistics in Medicine*. 1995 Sep;14(18):1975–83.

[126] Bureau A, Shiboski S, Hughes JP. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine*. 2003 Feb;22(3):441–62.

[127] Kohner EM, Stratton IM, Aldington SJ, Turner RC, Matthews DR. Microaneurysms in the development of diabetic retinopathy (UKPDS 42). UK Prospective Diabetes Study Group. *Diabetologia*. 1999 Sep;42(9):1107–12.

[128] Nunes S, Pires I, Rosa A, Duarte L, Bernardes R, Cunha-Vaz J. Microaneurysm turnover is a biomarker for diabetic retinopathy progression to clinically significant macular edema: findings for type 2 diabetics with nonproliferative retinopathy. Ophthalmologica Journal International d'Ophtalmologie International Journal of Ophthalmology Zeitschrift für Augenheilkunde. 2009 Jan;223(5):292–7.

[129] Jackson C. Multi-state modelling with R : The msm package; 2007.

[130] Titman AC, Sharples LD. A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine*. 2008 May;27(12):2177–95.

[131] Jones S, Edwards RT. Diabetic retinopathy screening: a systematic review of the economic evidence. *Diabetic Medicine : A Journal of the British Diabetic Association*. 2010 Mar;27(3):249–56.

[132] Milton RC, Ganley JP, Lynk RH. Variability in grading diabetic retinopathy from stereo fundus photographs: comparison of physician and lay readers. *The British Journal of Ophthalmology*. 1977 Mar;61(3):192–201.

[133] Beck JR, Pauker SG. The Markov Process in Medical Prognosis. *Medical Decision Making*. 1983 Jan;3(4):419–458.

[134] Box GEP. Use and Abuse of Regression. *Technometrics*. 1966 Nov;8(4):625–629.

[135] Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. *The New England Journal of Medicine*. 1989 Aug;321(6):406–12.

[136] Diao D, Wright JM, Cundiff DK, Gueyffier F. Pharmacotherapy for mild hypertension. *Cochrane Database of Systematic Reviews (Online)*. 2012 Jan;8:CD006742.

[137] Hu P, Zelen M. Planning of randomized early detection trials. *Statistical Methods in Medical Research*. 2004 Dec;13(6):491–506.

[138] Dinnes J, Hewison J, Altman DG, Deeks JJ. The basis for monitoring strategies in clinical guidelines: a case study of prostate-specific antigen for monitoring in prostate cancer. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2012 Mar;184(2):169–77.

[139] Treasure T, Monson K, Fiorentino F, Russell C. Operating to remove recurrent colorectal cancer: have we got it right? *BMJ*. 2014 May;348(2):2085–2085.

[140] NHS Information centre. *QOF DM 25*;. Last accessed: 11/07/2014. https://mqi.ic.nhs.uk/Search.aspx?query=diabetes&ref=1.09.04.21.

[141] May C, Montori VM, Mair FS. We need minimally disruptive medicine. *BMJ (Clinical research ed)*. 2009 Jan;339(aug11_2):b2803.

[142] Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*. 2006;1:1–19.

[143] Haslett J. On the sample variogram and the sample autocovariance for non-stationary time series. *Journal of the Royal Statistical Society: Series D (The Statistician)*. 1997 Dec;46(4):475–484.