Sample Size for Multivariable Prognostic Models

Rachel Claire Jinks

This dissertation is submitted for the degree of PhD

University College London
&
MRC Clinical Trials Unit

I, Rachel Claire Jinks, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Prognosis is one of the central principles of medical practice; useful prognostic models are vital if clinicians wish to predict patient outcomes with any success. However, prognostic studies are often performed retrospectively, which can result in poorly validated models that do not become valuable clinical tools. One obstacle to planning prospective studies is the lack of sample size calculations for developing or validating multivariable models. The often used 5 or 10 events per variable (EPV) rule (Peduzzi and Concato, 1995) can result in small sample sizes which may lead to overfitting and optimism. This thesis investigates the issue of sample size in prognostic modelling, and develops calculations and recommendations which may improve prognostic study design.

In order to develop multivariable prediction models, their prognostic value must be measurable and comparable. This thesis focuses on time-to-event data analysed with the Cox proportional hazards model, for which there are many proposed measures of prognostic ability. A measure of discrimination, the D statistic (Royston and Sauerbrei, 2004), is chosen for use in this work, as it has an appealing interpretation and direct relationship with a measure of explained variation.

Real datasets are used to investigate how estimates of D vary with number of events. Seeking a better alternative to EPV rules, two sample size calculations are developed and tested for use where a target value of D is estimated: one based on significance testing and one on confidence interval width. The calculations are illustrated using real datasets; in general the sample sizes required are quite large.

Finally, the usability of the new calculations is considered. To use the sample size calculations, researchers must estimate a target value of D, but this can be difficult if no previous study is available. To aid this, published D values from prognostic studies are collated into a 'library', which could be used to obtain plausible values of D to use in the calculations. To expand the library further an empirical conversion is developed to transform values of the more widely-used C-index (Harrell et al., 1984) to D.

Table of Contents

1 Prognostic research: an introduction **16**

1.1 Prognostic factors 16

1.1.1 Using prognostic factors 16

1.1.2 What makes a good prognostic factor? 18

1.1.3 Predictive factors 19

1.2 Multivariable prognostic models 20

1.2.1 Developing a prognostic model 20

1.3 Scope of the thesis 24

2 A review of sample size in prognostic studies **25**

2.1 Introduction 25

2.2 Current sample size calculations for prognostic studies 26

2.2.1 Binary outcome 26

2.2.2 Time-to-event outcome: single variable of interest 27

2.2.3 Time-to-event outcome: multivariate case 32

2.3 Sample size and prognostic model performance 34

2.3.1 R^2 based measures of prognostic ability 34

2.3.2 Non R^2 based measures of prognostic ability 38

2.4 Discussion 41

3 Investigating D **43**

3.1 Introduction 43

3.2 How is D affected by sample size? 43

3.2.1 Considerations for investigation 44

3.2.2 Methods 47

3.2.3 Results . . . 48
3.2.4 Results for D_{orig} . . . 50
3.2.5 Results for optimism . . . 53
3.2.6 Results for D_{opt} . . . 57
3.2.7 Conclusion . . . 61
3.3 Investigation of $SE(D)$ . . . 65
3.3.1 Aims . . . 65
3.3.2 Methods . . . 65
3.3.3 Results . . . 67
3.3.4 Conclusion . . . 72
3.4 Distribution and magnitude of D . . . 73
3.4.1 Results . . . 73
3.4.2 Conclusion . . . 74
3.5 Discussion . . . 75

4 A new structural parameter, λ 77
4.1 Introduction . . . 77
4.2 Definition of λ . . . 77
4.3 λ proportionality assumption . . . 78
4.4 Testing proportionality assumption for λ: real data . . . 79
4.4.1 Methods . . . 79
4.4.2 Results . . . 80
4.4.3 Conclusion . . . 82
4.5 Testing proportionality assumption of λ: simulation . . . 83
4.5.1 Methods . . . 83
4.5.2 Results . . . 83
4.5.3 Conclusion . . . 87
4.6 Estimating λ through simulation or bootstrap . . . 87
4.7 Relationship between λ and D . . . 89
4.7.1 Aim . . . 89
4.7.2 Methods . . . 89
4.7.3 Results . . . 90

4.7.4 Performance in simulated data with administrative censoring . . . 92
4.7.5 Performance in real data . . . 94
4.8 Discussion . . . 95

5 Sample size calculation in validation studies using D 97
5.1 Introduction . . . 97
5.2 Sample size calculation based on significance test . . . 98
5.2.1 Example . . . 100
5.2.2 Effect of parameters on calculation Sig-1 . . . 100
5.2.3 Minimum δ for various datasets . . . 102
5.3 Simulation study: significance based calculation Sig-1 . . . 103
5.3.1 Simulating datasets with exact numbers of events . . . 103
5.3.2 Method . . . 103
5.3.3 Results . . . 105
5.3.4 Conclusion . . . 107
5.4 Sample size calculation based on CI for D . . . 108
5.4.1 Example . . . 109
5.4.2 Effect of parameters on calculation CI-1 . . . 109
5.5 Simulation study: CI based calculation CI-1 . . . 110
5.5.1 Method . . . 110
5.5.2 Results . . . 111
5.5.3 How does censoring affect sample size? . . . 111
5.5.4 How does D affect sample size? . . . 113
5.5.5 Conclusion . . . 114
5.6 Discussion . . . 115

6 Sample size calculations using a point estimate of D 117
6.1 Introduction . . . 117
6.2 Sample size calculation based on significance test . . . 118
6.2.1 Example . . . 119
6.2.2 Effect of parameters on calculation Sig-2 . . . 120
6.3 Simulation study: significance based calculation Sig-2 . . . 121
6.3.1 Method . . . 121

6.3.2 Results . . . 122
6.3.3 Conclusion . . . 123
6.4 Sample size calculation based on CI for D . . . 124
6.4.1 Example . . . 124
6.4.2 Effect of parameters on calculation CI-2 . . . 124
6.5 Simulation study for CI based calculation CI-2 . . . 126
6.5.1 Method . . . 126
6.5.2 Results . . . 126
6.5.3 Conclusion . . . 127
6.6 Discussion . . . 128

7 Sample size examples **129**
7.1 Effect of dataset parameters on sample size . . . 131
7.2 Effect of calculation choice on precision . . . 135
7.3 Precision as a percentage of D . . . 137
7.3.1 Sig-2 . . . 138
7.3.2 CI-2 . . . 140
7.4 Discussion . . . 142
7.4.1 Which calculations should be used? . . . 142
7.4.2 Precision as proportion of D . . . 143

8 Systematic review of D values **145**
8.1 Introduction . . . 145
8.2 Methods . . . 146
8.3 Literature search for papers reporting D or R^2_D . . . 146
8.3.1 Aim . . . 146
8.3.2 Search strategy . . . 146
8.3.3 Results . . . 146
8.4 Converting Harrell's c-index to D . . . 148
8.4.1 Proposed conversion (White) . . . 148
8.4.2 Empirical relationship between D and c . . . 154
8.5 Literature search for papers reporting c . . . 158
8.5.1 Aim . . . 158

8.5.2 Search strategy . . . 158
8.5.3 Results . . . 158
8.6 Literature searches: summary of combined results . . . 159
8.6.1 Risk models in healthy subjects . . . 160
8.6.2 Prognostic models in patients with disease . . . 160
8.7 Discussion . . . 161

9 *D* Library **162**
9.1 Introduction . . . 162
9.2 Risk models in healthy subjects . . . 162
9.2.1 Incidence of cardiovascular disease . . . 163
9.2.2 Incidence of other diseases . . . 164
9.3 Prognostic models . . . 165
9.3.1 Cardiovascular disease . . . 165
9.3.2 Cancer . . . 165
9.3.3 Other diseases . . . 169
9.4 Discussion . . . 170
9.4.1 Differences in *D* within the same disease area . . . 170
9.4.2 Other uses of *D* and the *D* library . . . 173

10 Discussion **174**
10.1 Summary of thesis . . . 174
10.1.1 Review of available sample size guidance . . . 174
10.1.2 Investigation of some properties of *D* . . . 175
10.1.3 Development of sample size calculations . . . 176
10.1.4 Application of sample size calculations . . . 178
10.1.5 Improving utility of sample size calculations . . . 179
10.2 Recommendations . . . 180
10.2.1 Sample size recommendations . . . 180
10.2.2 Obtaining estimates of *D* in the model validation context . . . 182
10.2.3 Post-study recommendations . . . 183
10.3 Remarks on *D* . . . 184
10.4 Further research . . . 185

10.5 Final conclusions 187

A Datasets used in the thesis **188**

A.1 Datasets used in Chapters 3, 4, 7, and 8 188

B Models fitted to real datasets **193**

B.1 Datasets used in Chapters 3, 4, 7, and 8 193

B.2 SEER datasets used in Chapter 8 195

C Stata code for bootstrap procedure

200

D Sample size: further examples **203**

E *D* library **207**

E.1 Risk models in healthy subjects 208

E.1.1 Incidence of cardiovascular events 208

E.1.2 Incidence of diabetes 211

E.1.3 Incidence of bone fracture 211

E.1.4 Chronic kidney disease 211

E.1.5 Predicting side effects of statins 211

E.1.6 Colorectal / colon cancer 212

E.1.7 Dental caries 212

E.2 Prognostic models 212

E.2.1 Cardiac patients 212

E.2.2 Cancer 214

E.2.3 HIV 223

E.2.4 Liver disease 223

E.2.5 Respiratory disease 223

E.2.6 Chagas disease 224

E.2.7 Raynaud's phenomenon 224

E.2.8 Epilepsy 224

E.2.9 Leg ulcer 224

List of Figures

2.1 Relationship between D and R^2_D, for proportional hazards survival model 41

3.1 FBC: Scatter plots of D^{orig}_{boot} vs D^{boot}_{boot} for $p = 0.05$, for each chosen n 50
3.2 Example plots of various D values vs n for $p = 0.05$ 50
3.3 Matrix of plots of D_{orig} vs EPV, 8 non-SEER datasets 51
3.4 Matrix of plots of D_{orig} vs events, 8 non-SEER datasets 52
3.5 D_{orig} vs p for full datasets ($n = N$) 52
3.6 D_{orig} vs p for RBC datasets 52
3.7 Optimism vs EPV and vs events 53
3.8 Optimism vs EPV for all datasets, $p = 0.05$ 54
3.9 Optimism vs events for all datasets, $p = 0.05$ 55
3.10 Optimism vs EPV for all datasets, $p = 1.0$ 55
3.11 Optimism vs p for full datasets ($n = N$) 55
3.12 Optimism vs p for RBC datasets 56
3.13 D_{opt} vs EPV for each dataset, $p = 0.05$ 57
3.14 D_{opt} vs events for each dataset, $p = 0.05$ 57
3.15 D_{opt} vs EPV for each dataset, $p = 1.00$ 58
3.16 D_{opt} vs events for each dataset, $p = 1.00$ 58
3.17 Scatter plot of the % of $D_{opt,N}$ which D_{opt} attains when $n < N$, vs EPV . . 60
3.18 Scatter plot of the % of $D_{opt,N}$ which D_{opt} attains when $n < N$, vs events . 60
3.19 D_{opt} vs p for full datasets ($n = N$) 61
3.20 Absolute bias in $SE(D)_{cox}$ and $SE(D)_{boot}$ 71
3.21 Sampling distribution of D from simulated data 74
3.22 Sampling distribution of D from bootstrapped real datasets 74

4.1 Distribution of λ vs dataset size, for 6 real datasets 81

4.2 $SE(\lambda)$ as proportion of λ vs number of events, for 6 real datasets 82
4.3 Distribution of λ for simulated datasets 86
4.4 $SE(\lambda)$ as proportion of λ vs number of events, for simulated datasets . . . 86
4.5 Plot of λ vs D for five censoring rates 90
4.6 Original model: predicted λ overlaid with sim'n study results 91
4.7 Original model: predicted λ overlaid with sim'n study results, $D \leq 3.2$ . . 91
4.8 Alternative model: predicted λ overlaid with sim'n study results, $D \leq 3.2$ 92
4.9 Alternative model: predicted λ overlaid with sim'n study results 93
4.10 Observed λ vs D, administrative censoring 94
4.11 λ_{pred} vs λ_{true} for real datasets; models selected by MFP with various p . . 95

5.1 Sig-1 example: events required vs δ 101
5.2 Sig-1 example: events required vs size of first study 101
5.3 CI-1 example: events required vs w: various λ from 1 to 10 109

6.1 Sig-2: events required vs δ, for various D 120
6.2 Sig-2: events and patients required vs δ 121
6.3 CI-2: events required vs w 125
6.4 CI-2: events and patients required vs w 126

7.1 Events required vs δ: comparison of calculations Sig-1 and Sig-2 134
7.2 Sample sizes required from Sig-2 for δ considered as a percentage of D . . 139
7.3 Events vs D for composite Sig-2 calculation: $\delta = 0.15$ or 10% of D 139
7.4 Anticipated and actual precision vs D from study with 1827 events 140
7.5 Sample sizes required from CI-2 for w considered as a percentage of D . . 141

8.1 Histogram of 101 D values from first literature search 148
8.2 Predicted c from equation 8.2 vs D & observed c with random censoring . 151
8.3 Predicted c from equation 8.2 vs D & observed c with admin. censoring . 153
8.4 c vs D, using White's conversion, overlaid with D and c from real examples 153
8.5 Predicted c from equation 8.2 vs D overlaid with obs'd c from various sources 154
8.6 294 datapoints used to create model for predicting D from Harrell's c . . . 156
8.7 Histogram of 331 values of c obtained from second literature search. . . . 159
8.8 Histograms of values of D for different patient groups 160

9.1 Forest-type plot of D for CV and CHD events in healthy subjects. 164
9.2 Forest-type plot of D for CV events in healthy subjects 165
9.3 Forest-type plot of D for CV events in patients with existing CV condition 166
9.4 Forest-type plot of D for cancer papers with endpoint OS 168
9.5 Forest-type plot of D for cancer papers with endpoint CSS 168
9.6 Forest-type plot of D for cancer papers with endpoint PFS 169

10.1 Flowchart to aid decision making about which sample size to use 182

List of Tables

2.1 Sample sizes required for menopausal status, Hsieh and Lavori (2000) . . 31

2.2 Sample sizes required for Gleason stage-grade, Hsieh and Lavori (2000) . 31

3.1 Results of simulation investigation into $SE(D)$: $\beta = 0.5$, true $D \simeq 0.8$ . . . 68

3.2 Results of simulation investigation into $SE(D)$: $\beta = 1.0$, true $D \simeq 1.6$ . . . 69

3.3 Results of simulation investigation into $SE(D)$: $\beta = 2.0$, true $D \simeq 3.2$ . . . 70

3.4 Mean absolute and relative bias of $SE(D)_{cox}$ 72

3.5 Mean absolute and relative bias of $SE(D)_{boot}$ 72

3.6 Mean, bias and *se* of bias of D from simulated datasets with N=150,2000 . 75

4.1 Results of investigation of λ in real datasets 81

4.2 Results of simulation investigation into λ: $\beta = 0.5$ 84

4.3 Results of simulation investigation into λ: $\beta = 1.0$ 85

4.4 Results of simulation investigation into λ: $\beta = 2.0$ 85

4.5 Comparison of true and predicted λ for 26 real datasets 95

5.1 Minimum δ detectable in validation studies, for various real datasets . . . 102

5.2 Simulation study results for significance-based calculation Sig-1 106

5.3 Simulation study results for CI based calculation CI-1 111

5.4 Simulation study results for CI-1: misspecified censoring, $e_1 = 1500$ 113

5.5 Results of simulation study for CI-1: misspecified D 115

6.1 Simulation study results for significance based calculation Sig-2 123

6.2 Simulation study results for CI based calculation CI-2 127

7.1 Example sample size calculations based on parameters of real datasets . . 132

7.2 Precision in estimate of D obtained from studies of differing size 136

7.3 Sample sizes required by Sig-2 and CI-2 for composite precisions 141

8.1 Quantities reported in the 34 papers reporting D or R_D^2 147
8.2 Conversion tables for c, D using White's transformation 151
8.3 Conversion tables for c, D using our empirical transformation 157
8.4 Quantities reported in the 77 papers reporting c 159

A.1 Datasets used throughout thesis 189

D.1 Sample size calculations based on APC study 203
D.2 Sample size calculations based on GLI study 204
D.3 Sample size calculations based on LEG study 204
D.4 Sample size calculations based on LVA study 204
D.5 Sample size calculations based on MYE study 205
D.6 Sample size calculations based on PBC study 205
D.7 Sample size calculations based on RBC study 205
D.8 Sample size calculations based on SEER DE study 206
D.9 Sample size calculations based on SEER NM study 206
D.10 Sample size calculations based on STE study 206

Acknowledgements

I would like to thank the following people without whom this work would either not have happened or would have been a lot more stressful.

My supervisors: first and foremost Patrick Royston for his support and ideas all the way through my research and for keeping my self-confidence up; Max Parmar for his advice and interest in my work, in particular making sure I always kept my eye on the bigger picture; and Gareth Ambler for help with a few tricky problems and for his useful comments on my thesis draft.

My colleagues at CTU for making it such a nice place to work, in particular Tim Morris who has been a rich source of advice on LaTeX and all things typographical, and provided very helpful comments after reading my thesis draft.

My family: Dad, without whose genes I would surely not be a statistician, thanks for those (I think...); Mum, who makes sure I have a life outside work; and my brother, who can I can always rely on to make me laugh, even when everything is going wrong.

Last but certainly not least, my husband Pete, whose unwavering support and enthusiasm for everything I do is amazing. He has been truly awesome over the last three-and-a-half years and I owe him my final and most heartfelt thank you for all his love and encouragement.

Chapter 1

Prognostic research: an introduction

1.1 Prognostic factors

Prognosis is one of the central principles of medical practice. Understanding the likely course of a disease or condition is vital if clinicians are to treat patients with confidence or any degree of success. No two patients with the same diagnosis are exactly alike, and the differences between them – e.g. age, sex, disease stage, genetics – may have important effects on the course their disease will take. Such characteristics are called 'prognostic factors', and this phrase is usually taken to mean a factor which influences outcome *regardless of treatment*.

1.1.1 Using prognostic factors

Prognostic factors enable the estimation of patient prognosis, which has many applications.

Provides evidence for making treatment decisions

Prognostic factors are used routinely to guide clinical decisions. They help clinicians gauge the likely benefit of a treatment or course of action on an individual basis, which makes it easier to weigh up the risk-benefit ratio for a specific patient. One example is axillary lymph node status in breast cancer patients (Cianfrocca and Goldstein, 2004). While the proportional effect of therapy on hazard is the same for patients with positive or negative node status, their absolute risk is different to start with. If a treatment reduces the risk of death by 20%, then a patient with a positive node status and an underlying

risk (hazard) of death of 30% (say) would have their risk decreased to 24%; while a node-negative patient with a risk of death of 10% would have their risk decreased to 8%. A clinician may decide that the small absolute reduction in risk to the second patient is not worth the side effects and other inconveniences they would experience when undergoing the treatment.

Informs patients about the likely outcome of their disease

When a patient receives the news that they have a particular disease or condition, the first question in their mind is usually 'What is going to happen to me?'. Knowing the likely course of their disease can make it easier to come to terms with the news, and assist with patient counselling (Simon and Altman, 1994). There are practical as well as emotional issues to deal with at this time, particularly when patients have a terminal diagnosis.

Enables fairer comparisons of health care systems

It is becoming more common for hospitals or Trusts to be ranked in some way and the resulting 'league tables' are often publicly available. Patients awaiting surgery are also being given more choice in terms of their surgeon, and again statistics on outcome are available to help them make their decision. However, patient demographics and other factors can be very different from one hospital to the next and these differences must be taken into account if a comparison is to be fair (Moons et al., 2009).

Guides recruitment of patients to clinical trials

Existing known prognostic factors can be used to help define eligibility criteria for clinical trials. This ensures that an appropriate subset of patients are chosen for the research; for example researchers may be interested in recruiting a cohort of low or high risk patients for their study (Moons et al., 2009).

Allows stratification in randomised trials

Knowledge of prognostic factors means that they can be used as stratification factors in randomised trials (Altman and Lyman, 1998). Without stratification there is always the risk that an important factor may be imbalanced between treatment arms. This may make

the analysis more complex and results difficult to interpret, and at worst could render the trial useless.

Allows between-trial comparison and better meta-analyses

In order for trials to be compared, or combined in a good quality meta-analysis, prognostic characteristics of the patients may be quantified and accounted for (Simon and Altman, 1994). In the same way, analyses for non-randomised trials can be adjusted for important factors to facilitate a better comparison of treatment groups (Altman and Lyman, 1998).

Gives insight into pathogenesis and disease mechanisms

Prognostic factors may provide a starting point for research which could go on to expose important disease mechanisms or the pathogenesis of the condition (Simon and Altman, 1994).

1.1.2 What makes a good prognostic factor?

Several authors have listed the properties of an ideal prognostic factor; or properties which a potential prognostic factor should fulfil before being accepted into clinical practice (Gasparini et al., 1993; Simon and Altman, 1994; Henderson and Patek, 1998; Hermanek, 1999). These may be summarised as follows:

Feasible and reliable Determination of the factor must be standardised, reproducible and widely available. Results should be available within an acceptable timeline. Cost should be reasonable, and relative to the information gained from the factor. There should be some method of quality control to ensure that results remain accurate and that inter- and intra- observer variability is acceptably low.

Value added The factor should add substantial new prognostic information over and above what is already available from existing factors. However, a factor which does not add extra information may be of interest if it presents other advantages such as cost savings or a less invasive method of determination.

Discriminatory The factor must identify specific groups of patients with a better or worse outcome.

Interpretable The factor and its implications for treatment should be easily interpreted by clinicians.

Well described The nature of the factor should be known. It must be clear whether the factor is prognostic for a wide range of patients or just for specific subgroups.

Beneficial The differences in outcome that the factor defines should be of benefit to the patient.

Proven The factor must have been shown to be significant and independent of existing or suspected prognostic factors in more than one (independent) dataset.

1.1.3 Predictive factors

As well as signposting likely medical outcomes, certain factors may also indicate how likely a patient is to respond to a particular therapy. These are known as predictive factors and can be very useful in decision making; they are most often biomarkers such as gene mutations or proteins. For example, a patient with breast cancer may be treated with the monoclonal antibody trastuzumab only if their tumour overexpresses the HER2 gene; if it does not, this particular drug will have no positive effect and another therapy would be chosen. Thus HER2 status is a predictive factor for treatment with trastuzumab. However HER2 was first identified as a prognostic factor, as tumours which overexpress HER2 are generally more aggressive (Cianfrocca and Goldstein, 2004), and only later was it used as a starting point for a new treatment.

Like HER2, virtually all predictive factors are also prognostic; but this is not necessarily always the case, however exceptions in the literature are few and far between. Karapetis et al. (2008) reported that although the mutation status of the K-Ras gene was highly predictive of colorectal tumour response to the drug cetuximab, it did not appear to be prognostic when best supportive care alone was given. Many prognostic factors are not predictive of treatment effect; one example is axillary lymph node status in breast cancer patients (Cianfrocca and Goldstein, 2004). McShane et al. (2005) summarise the

practical difference between prognostic and predictive factors thus: 'Prognostic markers ... may be used as decision aids in determining whether a patient should receive adjuvant chemotherapy or how aggressive that therapy should be. Predictive markers are generally used to make more specific choices between treatment options.' However, the terms *predictive* and *prognostic* are often confused.

In basic terms, groups defined by a factor which is prognostic only, have a difference in survival and this (relative) difference between the groups remains the same regardless of treatment. Groups defined by a prognostic factor which is also predictive for a particular treatment show a difference in survival without the treatment of interest, and a larger difference if they receive treatment.

All the considerations involved in prognostic factor research also apply to predictive factor research. However, the latter requires additional work, as detecting interactions between variables adds complexity and many extra statistical considerations. Thus, issues specific to predictive factor research are not discussed further in this thesis.

1.2 Multivariable prognostic models

For most applications, a single predictor is not sufficiently precise; rather a multivariable approach to prognosis is required. Multivariable prognostic research enables the development of tools which give predictions based on multiple important factors; these may variously be called prognostic models, prediction models, prediction rules or risk scores (Moons et al., 2009). Multivariable research also means that potential new prognostic factors are investigated more thoroughly, as it allows the additional value of the factor, above and beyond that of existing variables, to be established (Moons et al., 2009). Most of the uses of prognostic factors described previously can be accomplished to a much greater degree when a multivariable prognostic model is used. As such, most of this thesis concentrates on the situation where a multivariable prognostic model is required, rather than a single prognostic factor.

1.2.1 Developing a prognostic model

In order to develop a prognostic model, it must be discovered to what extent each potential factor influences the outcome of interest, taking into account all the other factors of

interest as well. Some important issues to be considered when developing a prognostic model are discussed briefly.

Study design

The majority of prognostic studies are retrospective, simply because results are obtained much more quickly and cheaply by using existing data; especially where pathological samples are concerned. In their 2010 paper, Mallett et al. found that 68% of prognostic studies using time-to-event data in their review were retrospective. Altman (2009) conducted a search for and review of publications which presented or validated prognostic models for patients with operable breast cancer, and found that of the 61 papers reviewed, 79% were retrospective studies. However there are disadvantages to such studies. Missing data is almost invariably a problem, and there is little that researchers can do to mitigate this. The usual assumption that data are missing at random may be implausible in such datasets, biasing results (Altman and Lyman, 1998). This is particularly true with stored samples, as McGuire (1991) reports that tumour banks usually contain a disproportionate amount of samples from larger tumours, which would certainly introduce bias. Also, existing datasets may contain far more candidate variables than are really required (Royston et al., 2009), which can lead to multiple testing problems and a temptation to 'dredge' the data.

The best way to study prognosis is in a prospective study, which 'enables optimal measurement of predictors and outcome' (Moons et al., 2009). It may be convenient to make a prognostic study part of a randomised trial of treatment; however there are still issues to be considered in this situation. If the trial treatment proves to be effective, this must be taken into account in the analysis. A more serious potential problem is that strict trial eligibility criteria may mean the resulting model is not widely generalisable (Moons et al., 2009).

Sample size

This is always an important issue for clinical studies; however little research has been performed which pertains specifically to the sample size requirements of multivariable prognostic studies. In a review of publications developing and / or validating breast cancer models, Altman (2009) found that none of the 61 papers found justified the sample

size used; and indeed, for many it was impossible to even discern the number of patients or events contributing to the final model. Mallett et al. (2010) found that although 96% of studies in their review of survival models reported the number of patients included in analyses, only 70% reported the number of events – which is the key quantity for time-to-event data. In the same review, 77% of the studies included did not give any justification for the sample size used.

Sample size in prognostic studies, specifically for multivariable prognostic models, is the central theme of this thesis. Chapter 2 outlines the work that has been done so far in this area, and Chapter 3 proposes and describes the results of an investigation into sample size in multivariable prognostic model-building with time-to-event data. The remaining chapters extend this theme in order to provide sample size calculations for use in prognostic studies, and practical tools to try and ensure that these calculations are usable in real situations.

Candidate predictors

The predictors that are to be potentially included in the prognostic model should be decided on before data collection starts (or in the case of retrospective studies, before analysis). These are likely to be a mixture of already known and newly proposed prognostic factors, and may include patient baseline characteristics as well as factors relating to disease and treatment. All should be available at the time the model is to be used (Moons et al., 2009); for example, factors relating to surgery would not be appropriate in a model to be used in the neoadjuvant setting. If possible, any treatment received by patients in the study should be standardised (or randomised) and included as a prognostic factor (Simon and Altman, 1994). However, standardisation is often not possible in an observational setting and this can lead to bias (Moons et al., 2009).

Selection of variables

Once the set of candidate predictors has been chosen and data collected, the problem of how to choose a final model rears its head. One option is to use all candidate predictors in the final model, as sidestepping the issue in this way avoids some problems associated with selection of variables, for example selection bias which contributes to overfitting (Harrell, 2001). However using a full model does place greater importance on choos-

ing candidate predictors in the first place which may add complexity to study planning (Royston et al., 2009).

A popular alternative which is straightforward to implement is an automatic selection procedure using a fixed significance level, however these methods are known to result in selection bias and optimism (a result of overfitting) (Royston et al., 2009). Optimism and the effect of significance level in prognostic models developed using automatic variable selection are investigated in Chapter 3 of this thesis.

Validation

Having produced a prognostic model, its performance must be assessed. It should of course be assessed in the original dataset used to develop the model (internal validation); however it also must be shown to predict outcome well for other groups of patients if it is to be considered for general use (external validation) (Altman et al., 2009). There are various reasons that a model may perform poorly in a validation dataset, as described and illustrated in Altman et al. (2009). These include problems or errors in the model development process (e.g. small sample size, dichotomisation of continuous variables), the validation dataset being different in important ways from the development dataset (e.g. markedly different patient demographics or healthcare settings, changes in the way variables were measured), or an important predictor being excluded from the model (e.g. if a model was developed in patients with a narrow age range, the need to include a term for age in the model may not have been observed).

External validation of prognostic models is still relatively rare (Altman et al., 2009). In their review of prognostic models in operable breast cancer, Altman (2009) reported that of the 61 studies reviewed, only 19 included validation studies; and of these just 3 models were validated on external data. A similar rate of validation was found by Perel et al. (2006) who performed a review of models in traumatic brain injury. They found that of the 66 papers identified, only 25 validated the model under consideration, and only 7 performed external validation.

In this thesis we consider both the model development and validation situations.

1.3 Scope of the thesis

The desire of researchers across clinical areas to model patient prognosis seems to have somewhat outpaced the methodology required to ensure that prognostic studies are performed well and analysed carefully, and that the resulting multivariable models are valid and clinically useful. This thesis outlines the current sample size guidelines available to prognostic researchers using the Cox proportional hazards model and describes an investigation of the relationship between a measure of a model's prognostic ability and sample size, using both real data and simulated data. Sample size calculations are developed, tested and presented for use in two main scenarios in prognostic modelling: firstly where it is desired to validate an estimate of prognostic ability from a previous study, and secondly where only a target estimate of prognostic ability is available. Finally, these calculations are assessed in real data and some tools developed to increase their usefulness in real research.

Chapter 2

A review of sample size in prognostic studies

2.1 Introduction

Sample size is of vital importance when planning clinical research. If too few patients are included in a study, analysis results will have wide confidence intervals and low statistical power and precision. Including too many patients increases precision and power but is resource-costly and may not be feasible in reality. Balance is needed to ensure that a study collects enough data to give statistically valid and clinically useful results, whilst making efficient use of resources and completing in a timely fashion.

Before starting a prospective clinical study, a sample size calculation is performed virtually without fail. In the case of randomised controlled trials, a carefully considered, formal sample size calculation is usually a condition of funding and often required for subsequent publications, for example by journals that endorse the CONSORT statement (Moher et al., 2001). For retrospective studies, formal sample size calculations are available (based on the analysis methods used; for example, calculations for logistic regression are explored by (Demidenko, 2007)), but are not performed as frequently. Such studies are often based on whatever suitable existing data can be easily obtained, therefore sample sizes are haphazard and may be too low.

Sample size calculations for prognostic studies, which are usually retrospective, are not routinely available, meaning these studies may often be underpowered. Existing formulae can be used in some particular situations, but for most analyses of prognostic data,

particularly time-to-event data, little guidance is available to researchers. Mallett et al. (2010) found that in a systematic review of 47 published articles aiming to develop prognostic models in time-to-event data, all but one of the studies performed on retrospective data (n=32) did not provide any justification at all for the sample size used. In a review of publications developing and / or validating models in operable breast cancer, none of the 61 papers found justified the sample size (Altman, 2009).

To gain a better idea of the guidance and recommendations which are available, a literature search for papers dealing with sample size calculations in prognostic studies was performed. A keyword search was initially attempted; however the wide variety of terms associated with prognostic modelling, and their use in a range of different topics meant that this was not a fruitful strategy. Instead, some relevant papers were identified, and then citation searches were used to identify further related papers. The papers found in this search form the first part of this chapter.

In the second part, the possibility of a sample size recommendation based on prognostic ability is explored. Again, a literature review was initiated to explore proposed measures of prognostic ability; two previous detailed reviews of this area were found which informs much of this section (Schemper and Stare, 1996; Choodari-Oskooei, 2008). Finally, one measure of prognostic ability is selected to form the basis of a novel recommendation.

2.2 Current sample size calculations for prognostic studies

2.2.1 Binary outcome

Where logistic regression is to be used in analysis, there are various sample size formulae available, but no consensus on which is the best approach (Demidenko, 2007). There are methods applicable to situations where just one factor is of interest, and where multiple factors are to be investigated, for example using a variance inflation factor to allow for additional variables (Hsieh et al., 1998).

The situation is more complicated when a time-to-event outcome is used, as is most common in cancer research; such an outcome will be concentrated on in this thesis.

2.2.2 Time-to-event outcome: single variable of interest

Sample size formulae have long been available for randomised group comparisons using survival analysis, commonly based on the log-rank test (Lakatos and Lan, 1992; Freedman, 1982) or the Cox proportional hazards (PH) model (Schoenfeld, 1983). However these formulae are not necessarily valid when a prognostic factor is the effect of interest, as such a factor is expected to be correlated with other covariates, whereas a randomised treatment should not be. For example, Bernardo et al. (2000) showed that the power of the unadjusted log-rank test is overestimated when the binary prognostic factor of interest is correlated with other covariates. Three papers, all published in the same year, independently found the same correction to Schoenfeld's (1983) sample size formula for this situation (Schmoor et al., 2000; Bernardo et al., 2000; Hsieh and Lavori, 2000).

Schmoor et al. (2000) extended Schoenfeld's (1983) formula to the situation where the Cox PH analysis is adjusted for a correlated factor. They framed this work in the context of studying the 'prognostic relevance' of one factor (X_1) in the presence of another (X_2). The effects of X_1 and X_2 can be analysed using the Cox PH model:

$$h(t|X_1, X_2) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2),$$

where $h_0(t)$ is the unspecified baseline hazard function, and β_1 and β_2 are the regression coefficients for X_1 and X_2. In the article it is assumed that X_1 and X_2 are binary and that $P(X_1 = 1) = p$ and $P(X_2 = 1) = q$. The hazard ratio between the groups defined by X_1 is denoted by $\theta = \exp(\beta_1)$ and that between the groups defined by X_2 is $\eta = \exp(\beta_2)$. It is assumed that the effect of X_1 will be tested by a two-sided test based on the partial likelihood of the Cox model (Schoenfeld, 1983), with significance level α and power $1 - \beta$ to detect an effect of $\exp(\beta_1) = \theta_1$. If X_1 and X_2 are independent, the total number of patients required was shown by Schoenfeld (1983) to be

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\log \theta_1)^2 \psi (1-p) p}, \tag{2.1}$$

where $1 - \psi$ is the probability of a censored observation (so ψ is the proportion of events in the dataset) and z_γ is the γ-quantile of the standard normal distribution.

Schmoor et al. (2000) then considered the case where X_1 and X_2 (both binary) are not independent, having correlation $\rho = \text{corr}(X_1, X_2)$. In their calculations they used the same approximation as Schoenfeld (1983); namely that the asymptotic variance of the maximum partial likelihood estimator (MPLE) of β_1 under $H_1 : \beta_1 = \log(\theta_1)$ may be approximated by the asymptotic variance under $H_0 : \beta_1 = 0$. They also assumed that the probability of censoring is approximately equal under H_0 and H_1. This led to the following sample size calculation:

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\log \theta_1)^2 \psi (1-p) p} \left(\frac{1}{1-\rho^2} \right). \tag{2.2}$$

The first part of Formula 2.2 contains Schoenfeld's calculation 2.1, and Schmoor et al. termed the bracketed second part involving ρ the 'variance inflation factor'. In a simulation study, it was found that this approximate formula was quite accurate (Schmoor et al., 2000). There was little deviation from the desired power while parameters were of 'moderate size', either when increasing the effect size of X_2 or when increasing ρ; however quantitative results of the simulation study were not presented in the paper.

Bernardo et al. (2000) also worked on this problem but in the context of non-randomised studies, modelling survival times using the exponential distribution. Their solution has wider application as more than one secondary variable can be in the model, and these do not have to be binary. They used general likelihood theory to derive the variance of $\widehat{\beta}_1$, and express this as a function of the coefficient of determination gained from regressing X_1 on the other covariates. Using the same notation as previously, the calculation they obtained can be rearranged to mirror Schmoor et al.'s calculation 2.2:

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\log \theta_1)^2 \psi (1-p) p} \left(\frac{1}{1-R^2_{X_1|X_2}} \right), \tag{2.3}$$

where X_1 is the variable of primary interest and X_2 is now the vector of additional covariates (more details can be found in Bernardo et al. (2000)). $R^2_{X_1|X_2}$ is the coefficient of determination obtained by regressing X_1 on X_2. It is clear that calculation 2.2 is the special case of calculation 2.3 when there is just one additional covariate.

Hsieh and Lavori (2000) also proposed the same calculation for the situation where the binary primary factor of interest is correlated with other covariates. They argued

that if the vector of secondary covariates X_2 explains some of the variance in the primary covariate X_1, then the conditional variance of $X_1 | X_2$ will be less than the marginal variance of X_1 by a factor of $1 - R^2_{X_1|X_2}$. Thus to preserve the desired power, the standard Cox PH sample size calculation 2.1 must be multiplied by the variance inflation factor $1/(1 - R^2_{X_1|X_2})$. Clearly this gives the same final calculation as 2.2 and 2.3.

In the same paper, the authors presented another sample size calculation, again for the Cox PH model, for use when the covariate of interest X_1 is not binary but rather continuous, with a linear effect:

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\log \Delta)^2 \psi \sigma^2}. \tag{2.4}$$

Here, $\log \Delta$ is the log hazard ratio associated with a one unit change in X_1, and σ^2 is the variance of X_1. As for the binary case, the same variance inflation factor $1/(1 - R^2_{X_1|X_2})$ can be used to take into account correlated patient covariates. If the X_1 is binary rather than continuous, then its variance $\sigma^2 = p(1-p)$ and the equation reverts to Schoenfeld's formula 2.1.

The above calculations were developed using various assumptions and approximations. In contrast, Schoenfeld and Borenstein (2005) used the asymptotic distribution of the Wald test statistic to develop an exact algorithm for calculating power for Cox PH and logistic regression models. The authors used simulation to assess the performance of the algorithm in the Cox PH model under various scenarios, and found that it generally compared favourably to simulation unless censoring levels were high (such as 90%). The main practical benefit of using the exact algorithm to determine power rather than simulation, as reported by the authors, was that the algorithm required less computing time than simulations; however the times reported for both methods were not lengthy. Despite mentioning Schmoor et al's and Bernando et al's work, these calculations were not compared to the exact algorithm, rather the paper concentrates mainly on the logistic regression case.

We have shown that sample size calculations exist for use in situations where just one variable is of primary interest, but other correlated variables need to be taken into account in the analysis. This is common in prognostic research, for example when a new factor of uncertain prognostic value is discovered, and also in epidemiology when a risk

factor is investigated in the presence of confounders. In this case the main research aim may be to estimate the additional predictive value of the new factor; other known factors need to be accounted for but it is not of interest to estimate any quantities relating to them. In this scenario the sample size calculations 2.2 and 2.3 would be suitable for a binary primary factor, and 2.4 with the variance inflation factor if the primary factor was continuous.

Examples

Two brief examples are given to illustrate sample size calculation for the Cox model when the primary variable of interest is correlated with other variables, and to gain a better idea of how this correlation affects sample size. The above sample size formula 2.4 (Hsieh and Lavori, 2000) is implemented in Stata by the command `stpower`, which was used to perform the following calculations. Real datasets are used to obtain clinically realistic values for correlation. This is similar to what would be done in real life: an existing dataset in the disease of interest would be used to give an idea of the likely parameters for a new study or trial.

Breast Cancer Example The first dataset used originates from a German trial which recruited patients with primary node positive breast cancer (Sauerbrei and Royston, 1999). The dataset used here includes 7 prognostic variables and a variable for hormonal treatment (for details see dataset FBC in Appendix B). In this example the variable of primary interest is chosen to be menopausal status, a binary variable which is split approximately 40% / 60% pre- / post-menopausal. Linear regression of this variable on the other 7 gives $R^2 = 0.60$. The following assumptions are made in the calculations: $\alpha = 0.05$ (two-sided), power is 80%, standard deviation of the primary variable is 0.5 (as is approximately correct for a 40/60 binary variable), and for simplicity there is no censoring nor loss to follow up. The sample size required to detect hazard ratios from 1.1 to 1.5 in steps of 0.1 are given in Table 2.1: firstly for the situation where correlation between variables is ignored ($R^2 = 0$) and then where correlation is $R^2 = 0.6$, as in the breast cancer dataset.

As shown by Hsieh and Lavori, the sample size required when no correlation is assumed is multiplied by $1/(1 - R^2)$ to account for an $R^2 > 0$. In this case, with $R^2 = 0.6$,

	Events required	
	$R^2 = 0$	$R^2 = 0.6$
HR=1.1	3457	8641
HR=1.2	945	2362
HR=1.3	457	1141
HR=1.4	278	694
HR=1.5	191	478

Table 2.1: Sample sizes (events) required to detect HR for menopausal status variable according to Hsieh and Lavori (2000)

	Events required	
	$R^2 = 0$	$R^2 = 0.5$
HR=1.02	5004	10008
HR=1.04	1276	2552
HR=1.06	578	1156
HR=1.08	332	663
HR=1.10	217	433

Table 2.2: Sample sizes (events) required to detect HR for Gleason stage-grade variable according to Hsieh and Lavori (2000)

the multiplier is 2.5. A lower R^2 of 0.2 results in a multiplier of 1.25, and a higher R^2 of 0.8 results in a multiplier of 5.

Prostate Cancer Example In our second example, we use data from a randomised trial in advanced prostate cancer (Byar and Green, 1980), which includes 13 prognostic variables plus a treatment variable. The variable of interest is Gleason stage–grade category, which takes discrete values from 5 to 15 in this dataset. In this example we will treat it as continuous and use equation 2.4 to calculate sample size. Linear regression of this variable on the other 14 gives $R^2 = 0.5$. The same assumptions are made as in the previous example, however a realistic standard deviation for the primary variable is calculated from the dataset, which gives $\sigma = 2$. This time, hazard ratios are based on a one unit increase in the variable of interest. Table 2.2 shows the sample size required to detect hazard ratios from 1.02 to 1.10 in steps of 0.02, for the situation where correlation between variables is ignored and then where correlation is $R^2 = 0.5$.

This time the multiplier is 2 as $R^2 = 0.5$. Although the hazard ratios chosen are small, the high standard deviation of the continuous variable of interest makes the sample sizes required lower than for the binary example.

2.2.3 Time-to-event outcome: multivariate case

Another – perhaps more common – scenario in prognostic research is when researchers wish to produce a multivariable prognostic model. In this situation, all the variables in the model are of equal importance. Since there is not one single variable of primary importance, the previously mentioned formulae do not apply and a different solution must be sought.

Standard sample size formulae

With multiple variables of equal interest, development of 'standard' sample size formulae is problematic. Such formulae are usually based on effect size, but in the situation where an arbitrary number of variables are of interest, which effect is being measured? The magnitude of each covariate individually? Some combined measure of magnitude? The formulae shown in the previous section could possibly be extended to the case where there are two variables of primary interest to be tested, with a composite null hypothesis, but beyond this the covariance matrices would make the algebra required intractable and specification of the alternative hypothesis difficult.

Events per variable calculations

There are various recommendations in the literature about how many patients or outcome events are required to estimate regression coefficients in a multivariable model with reasonable accuracy. The most often cited recommendation is the rule of '10 events per variable (EPV)' which originated from two simulation studies (Concato et al., 1995; Peduzzi et al., 1995). In these papers, exponential survival times were simulated for 673 patients from a real randomised trial with 252 deaths and 7 variables (36 EPV), and then the number of deaths were varied to reduce the EPV. From this they considered whether there was a minimum EPV needed for tests based on Cox PH analysis to have the required power, confidence intervals the right coverage, and reasonably unbiased coefficient estimation. They found that choosing a single minimum value for EPV was difficult but that results from studies having fewer than 10 EPV should be 'cautiously interpreted'.

A later simulation study by Vittinghoff and McCulloch (2007) found that in 'a range of circumstances' having less than 10 EPV still provided acceptable confidence interval coverage and bias when using Cox regression. Unlike the Concato and Peduzzi (1995) papers, this study did not directly consider the statistical power of analyses nor the variability of the estimates. The final line of this paper states that 'systematic discounting of results...from any model with 5-9 EPV does not appear to be justified'. The authors do not recommend aiming for less than 10 EPV when planning analyses and agree with Peduzzi et al. (1995) that results from studies with low EPV should be interpreted with caution. However it is perhaps inevitable that this paper has been cited to justify low sample sizes, for example in (Mahambrey et al., 2009) and (Putman et al., 2009). Mallett et al. (2010) found in their review of papers reporting development of prognostic models in time-to-event data, that of the 28 papers reporting sufficient information to calculate EPV, 14 had fewer than 10 EPV.

Looking at this issue from a different angle, Smith, Harrell, and Muhlbaier (1992b) used simulation to assess the error in survival predictions with increasing numbers of model covariates. Datasets of 250 and 750 subjects (64 and 185 events respectively) were drawn from an exponential distribution such that the average 5-year survival was 75%. Cox models were fitted to the simulated data, with between 1 and 29 uniformly distributed covariates. The authors found that in both the 64 and 185 event datasets, 5-year survival predictions from the Cox models became increasingly biased upwards as the EPV decreased. In both datasets, the average error was below 10% when EPV$>$10, and below 5% when EPV$>$20. For 'sick' subjects – those at high risk of death – higher EPVs were required: EPV$>$20 was required to reduce the expected error to 10%.

It should be noted that when calculating EPV all the candidate variables should be counted, even if they are not included in the final model. This is of particular importance when using a stepwise model selection method, as the variables initially considered for inclusion may be far greater in number than those in the final model. However, defining what exactly constitutes a candidate variable may not be straightforward, especially in retrospective studies where the list of variable available for analysis may be extensive. It can also be difficult to count candidate variables when model variables are to be transformed; for example, if fractional polynomials or splines are to be used to model continuous variables, a single variable may be represented in the model by multiple power

terms. Another issue arises with categorical variables: a categorical variable taking 5 possible values is equivalent to 4 dummy binary variables. Does such a variable contribute one or four to the denominator of the EPV calculation? Some statisticians prefer to consider events per parameter (EPP) (Hosmer and Lemeshow, 2000) but again, this phrase is not clearly defined and indeed EPV and EPP are often used interchangeably, which adds to the confusion.

2.3 Sample size and prognostic model performance

When developing a prognostic model it is likely that individual covariate effects are not of major interest. Instead the main aim is likely to be measuring the ability of the model to predict outcomes for future patients, or to discriminate between groups of patients. Copas (1983) says that '...a good predictor may include variables which are "not significant", exclude others which are, and may involve coefficients which are systematically biased'. Thus basing sample size decisions on the significance of model coefficients alone may not result in the best prognostic model.

Currently there do not seem to be any sample size calculations or recommendations based on the prognostic ability of a model, rather than the significance of its coefficients. Developing such tools would be of great use considering how much research – especially in oncology – attempts to produce multivariable prognostic models which are clinically useful.

Before trying to develop a recommendation based on the prognostic ability of a survival model, it must be decided how best this ability can be assessed. This section reviews proposed measures, which can broadly be divided into those measures which are based on the statistical quantity R^2, and those which are not. Potentially useful measures are described in more detail.

2.3.1 R^2 based measures of prognostic ability

The quantity R^2 in a model describes how much of the variation in the dependent variable is explained by the independent variables. Thus it is is a measure of how well the model may predict outcome in future cases. In normal linear regression the single quantity R^2 measures the explained variation, explained randomness and predictive accuracy

of the model, but in models of time-to-event data these three quantities do not coincide. This leads naturally to the categorisation of this family of measures into three classes.

Schemper and Stare (1996) outlined and reviewed the various R^2 based measures proposed for the Cox PH model. They did not find a measure that fulfilled all their criteria for a good measure; in particular, most of the measures were affected by censoring. Since 1996 new measures have been developed and in his PhD thesis Choodari-Oskooei (2008) comprehensively reviewed and investigated the currently proposed measures, again against a list of properties. The properties classed as essential were independence from censoring, independence from sample size, and monotonicity of the measure's magnitude in terms of the magnitude of parameters and number of variables in the model. Desirable properties were robustness to outliers, generalisability to different types of survival model and availability of confidence intervals, partial R^2, and adjusted R^2. This review informs the following three sections.

Measures of explained variation

One interpretation of R^2 is that it is the proportion of variation in the dependent variable which is explained by the model. The higher the proportion explained, the better the predictive ability of the model. The measures in this class differ in how they measure the variation in outcome; more detail can be found in Choodari-Oskooei (2008) and subsequent paper Choodari-Oskooei et al. (2011).

Five measures in this class were found to be potentially useful based on work done prior to the review. These were Kent and O'Quigley's (1988) measure R^2_{PM} (see below for details), O'Quigley and Flandre's (1994) R^2_{OQF} (which utilises Schoenfeld residuals), O'Quigley and Xu's (2001) R^2_{XuOQ} (a further development of R^2_{OQF}), Royston and Sauerbrei's (2004) R^2_D (see below for details), and Royston's (2006a) $R^2_{Royston}$ (a modification of O'Quigley's (2005) ρ^2_k, a measure of explained randomness described below).

A conclusion of Choodari-Oskooei's thesis and his subsequent paper (Choodari-Oskooei et al., 2011) was that this class of measures was most recommended for use because they are easily interpretable. For example, a value of 0.3 means that the prognostic variables explain 30% of the variation in the outcome on the log hazard scale. The other classes do not have such intuitive explanations. Further to this, R^2_{PM} and R^2_D were specifically recommended as they mostly fulfilled the essential criteria set out, and among

other benefits noted by Choodari-Oskooei (2008), may be used with flexible parametric models (Royston and Parmar, 2002).

As a measure of explained variation, the simplest way to describe R^2 is:

$$R^2 = \frac{\text{variation in outcome explained by covariates}}{\text{total variation in outcome}},$$

where the total variation in outcome is made up of the variation explained by the covariates, and the remaining unexplained variation. Kent and O'Quigley's (1988) measure R^2_{PM} for the Cox PH model uses the variance of the prognostic index $\beta' x$, and approximates the unexplained variation by the variance in the error term of the log-Weibull model ($\frac{\pi^2}{6}$):

$$R^2_{PM} = \frac{Var(\beta' x)}{Var(\beta' x) + \frac{\pi^2}{6}}.$$

One drawback to R^2_{PM} is that it cannot be used in this form as a tool for external model validation (that is, validation on data not used to develop the model), since it does not take into account the fit of the model to the outcome data. It is also mildly affected by extreme and outlier observations in the data; but appears not to be affected by the level of censoring (Choodari-Oskooei et al., 2011).

Royston and Sauerbrei's (2004) R^2_D is based on R^2_{PM} but uses their measure of prognostic separation of survival curves D to create a new measure for the Cox PH model:

$$R^2_D = \frac{D^2/\kappa^2}{D^2/\kappa^2 + \sigma^2}.$$

For more details on D see section 2.3.2.

R^2_D can be used with various survival models by defining the scaling parameter σ^2 differently; in the Cox PH model $\sigma^2 = \frac{\pi^2}{6}$ (Royston and Sauerbrei, 2004), again the variance in the error term of the log-Weibull model. The constant $\kappa = \sqrt{8/\pi}$ is used to give a direct interpretation to D (see Section 2.3.2 for more information). R^2_D is not affected by censoring if the model prognostic index (PI) is normally distributed (see section 2.3.2); but if this assumption is violated, censoring can have a large effect on its value (Choodari-Oskooei et al., 2011).

Measures of explained randomness

This class exploits the relationship between R^2 and the concept of information as a measure of uncertainty (see Choodari-Oskooei (2008) for a more detailed explanation of information in this context).

Three measures in this class were potentially recommendable based on previous work. These were Kent and O'Quigley's (1988) measures ρ^2_W and $\rho^2_{W,A}$ (based on model likelihoods), Xu and O'Quigley's (1999) ρ^2_{XuOQ} (similar to ρ^2_W and $\rho^2_{W,A}$ but with an alternative information gain), and O'Quigley et al.'s (2005) ρ^2_k. Just one of these (ρ^2_W) fulfilled all the essential criteria, but it was not recommended for use because it is difficult to interpret, may be complex to calculate and lacks generalisability to survival models other than Cox PH (Choodari-Oskooei, 2008).

Measures of predictive accuracy

This interpretation of R^2 measures how close model predictions are to observed outcomes. In a survival model context, these predictions are in terms of a subject's survival probability over time, rather than length of survival.

Only two members of this class were found to be possibly suitable on initial review: Graf et al.'s (1999) measure R^2_G and Schemper and Henderson's (2000) V_{SchH}. Both calculate predictive accuracy of a model by taking a weighted average of the prediction error at every event time. Choodari-Oskooei (2008) found these measures both lacking due to their dependency on length of study follow up.

Other R^2 based measures

There are some R^2 based measures which do not fall into any of the three categories above. Choodari-Oskooei (2008) found the most promising of these to be Schemper and Kaider's R^2_{SchK}, which imputes censored survival times and then uses nonparametric correlation to measure association. Upon further investigation it was found to fulfill all of the essential criteria on the checklist, but its complexity and lack of a clear interpretation made other options more favourable.

2.3.2 Non R^2 based measures of prognostic ability

There are several measures not based on R^2 which quantify the prognostic ability of a model. These are primarily measures of discrimination; they consider how well a prognostic model can distinguish between observed outcomes. In the case of survival analysis, this is between the patients who do and do not experience the event in question. These measures fall into two classes. Firstly, concordance statistics, which measure the agreement between observed outcomes and predicted risk, and secondly, those which measure how the observed risk varies from the lowest to highest predicted risks.

Concordance statistics

Harrell et al.'s (1984) c-index is a rank-based test which measures agreement between pairs of subjects, in terms of their outcome and predicted risk. A pair is concordant if the subject with the higher predicted risk experienced the event of interest before the other subject. It considers all possible pairs where the shorter follow-up time ends in failure, and c is then the proportion of concordant pairs. Thus c can be interpreted as the probability that for a random pair of patients, the one with the higher prognostic index will experience the event first. $c = 0.5$ indicates that the model predicts no better than would be expected by chance, and $c = 1$ means perfect concordance. $c < 0.5$ implies that patients with lower model risk are more likely to experience the event than those with higher risk; thus the model still has predictive value, but in the opposite direction to that expected. Being based on rank, the c-index is not model dependent. In his book 'Regression Modeling Strategies', Harrell (2001) reports that 'the c-index is relatively unaffected by the amount of censoring'; however Gonen and Heller (2005) found that its value seemed to increase slightly with the proportion of censoring.

To improve upon the c-index for time-to-event data, Gonen and Heller (2005) derived an analytical expression for concordance probability within the Cox model specifically. They termed their result, based on the partial likelihood estimator of β, the 'concordance probability estimator', $K_n(\widehat{\beta})$. Because the effect of censoring on $\widehat{\beta}$ is negligible, their statistic $K_n(\widehat{\beta})$ is robust to censoring.

One problem with concordance statistics is that their scale may not be intuitive to researchers and clinicians. Also, on a practical level, $K_n(\widehat{\beta})$ is a non-smooth function of $\widehat{\beta}$

and so must be approximated by smoothing or numerical differentiation before applying to data. For these reasons other measures of prognostic ability are preferred; however, the c-index is commonly reported for prognostic models in the literature.

Measures of difference in observed risk across the prognostic index

These measures attempt to quantify the separation in observed survival curves between subgroups of patients with differing predicted risks. This should be intuitively easy to understand for researchers with understanding of Kaplan-Meier survival curves.

To measure this separation, Sauerbrei et al. (1997) proposed the measure SEP , which is the geometric mean of the absolute relative risks in each strata, weighted by the number of patients in the strata. It was further explored by Graf et al. (1999) who described SEP as being 'constructed to assess by which amount survival within risk strata differs on average from survival in the entire population'. There are some problems with SEP, noted by Royston and Sauerbrei (2004): it requires the prognostic index to be split into risk groups prior to calculation, it does not take into account the ordering of risk categories, it is always positive regardless of the predictive usefulness of the model, and it has no easily calculated standard error. Some of these problems are particularly important when validating a model in another dataset and as a result of this, Royston and Sauerbrei (2004) worked to develop an improved measure of prognostic separation, called D.

D was developed in the Cox model framework and is based on risk ordering, in that the first step in deriving D is to order individuals' prognostic indices from lowest to highest risk. Thus D can be calculated whether the prognostic tool outputs a continuous prognostic index, prognostic groups, or is even a subjective rule. However, it is assumed that the prognostic index resulting from the model is Normally distributed (although this is an approximation in the case of a non-continuous prognostic index). The full derivation of D is described in Royston and Sauerbrei (2004); an abridged version follows below.

The standard Cox model can be written

$$\ln h(t_i|\mathbf{x}_i) = \ln h_0(t_i) + h_i,$$

where $h_i = \boldsymbol{\beta}'\mathbf{x}_i$, the prognostic index of the ith subject. The h_i are ranked and then replaced with the corresponding standard Normal order scores; these scores are then

divided by $\kappa = \sqrt{8/\pi} \simeq 1.60$. Cox regression is performed on these scaled Normal scores and the log hazard ratio resulting is

$$D = \kappa\sigma^*,$$

where under the assumption that the h_i are normally distributed, σ^* is an estimate of the standard deviation of the prognostic index values. The motivation for scaling by κ is that it gives D an intuitively appealing interpretation as the log hazard ratio between two equally sized prognostic groups, formed by dichotomising the prognostic index at its median. This is because the mean of a standard half-Normal distribution is $\frac{1}{2}\kappa$, and again relies on normality of the prognostic index; for a full explanation see Royston and Sauerbrei (2004).

D can theoretically take any value in the range $(-\infty, \infty)$, but in real situations it is likely to be much closer to zero. $D = 0$ implies that the selected model has zero predictive ability, and $D < 0$ may arise when a model fitted to one dataset is validated on another, indicating that the original model was overfitted. D's interpretation as a log hazard ratio means that it can be translated to a hazard ratio between the equally sized prognostic groups; so a D of 1 corresponds to a hazard ratio of $e^1 = 2.7$ and $D = 2$ to $e^2 = 7.4$. This allows researchers familiar with hazard ratios of treatments (for example) to have some idea of the increase in risk across the prognostic index of the model.

We have previously described R^2_D (Royston and Sauerbrei, 2004), a measure of explained variation based on D, and the two quantities have a one-to-one relationship:

$$R^2_D = \frac{D^2/\kappa^2}{\sigma^2 + D^2/\kappa^2}.$$

The value of σ^2 depends on the survival model being used; it is 1 for standard Normal distribution with probit link, $\frac{\pi^2}{3}$ for standard logistic distribution with proportional odds, and $\frac{\pi^2}{6}$ for standard extreme value distribution with proportional hazards (Royston and Sauerbrei, 2004). Figure 2.1, adapted from a graph in Royston and Sauerbrei (2004), shows the relationship between R^2_D and D for the Cox model ($\sigma^2 = \frac{\pi^2}{6}$). This relationship is important as most researchers will be more familiar with the 0–100% range of R^2.

As well as its interpretability and applicability to many types of prognostic model, D has many other properties which make it suitable for practical use. These include robust-

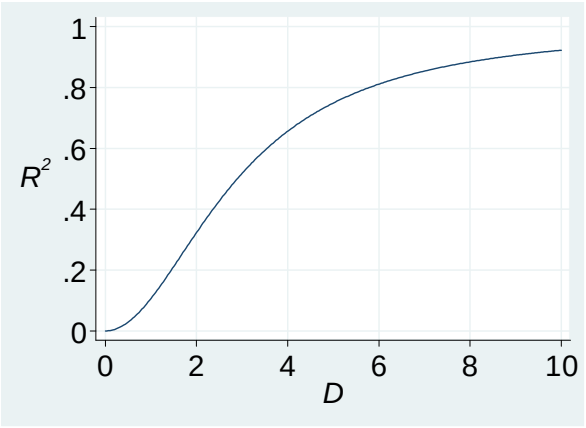

Figure 2.1: Relationship between D and R^2_D, for proportional hazards survival model

ness to outliers, sensitivity to risk ordering, independence from censoring (provided the prognostic model has been correctly specified and the PI is normally distributed), and an easily calculated standard error (Royston and Sauerbrei, 2004). Also, since it takes into account the fit of the model to the outcome data, it can be used in a model validation context; a vital part of a good prognostic study.

2.4 Discussion

In this chapter we have explored sample size calculations for prognostic research when a time-to-event outcome is used. The situation where one variable of primary interest is correlated with other secondary variables already appears to be covered by existing calculations, as seen in Section 2.2.2. When a multivariable prognostic model is of interest, the only sample size guidance currently available comes in the form of events-per-variable recommendations, which generally suggest having at least 10 events for each candidate predictor variable. However this rule of thumb is based on just a few simulation studies and is focused on the estimation of individual variables rather than the performance of the model as a whole.

Since the aim of forming a prognostic model is to be able to predict outcomes accurately, it would be useful for a sample size recommendation to be based on the predictive ability of a model, rather than the effect size of model covariates. As a first step towards this end we reviewed proposed measures of prognostic ability in survival models and found R^2_{PM} and R^2_D or D to be the most promising. Due to their utility in the model

validation context, D and R^2_D were felt to be slightly superior and so D was chosen for use in the development of a sample size recommendation for multivariable prognostic models. Later in the thesis, in Chapter 8, we will return to another of the measures of prognostic value discussed here, Harrell's c-index. Because it is so commonly reported in the medical literature, we consider a conversion from c to D, in order that the value of D in various different disease areas may be estimated.

In the next chapter we investigate the properties of our chosen quantity D further, particularly looking at how it varies with sample size, its standard error, and its sampling distribution.

Chapter 3

Investigating D

3.1 Introduction

Chapter 2 discussed the lack of existing sample size recommendations for studies aiming to develop a multivariable prognostic model. In particular there are no recommendations based on the prognostic ability of a model, rather than the effect size of one key covariate. In order to develop such recommendations, Royston and Sauerbrei's D was chosen as the preferred measure of prognostic ability.

In this chapter we consider various properties of D in order to facilitate the development of a sample size calculation. Firstly, we investigate how D behaves as sample size or events-per-variable (EPV) changes, using real data. Secondly, we look at how best to calculate the standard error of D. Finally, we consider the distribution of D.

3.2 How is D affected by sample size?

The aim of this investigation is to discover how D is affected by sample size, using real datasets. Breaking this down further, we wish to determine how the estimate of D changes as the sample size (in terms of both number of events and EPV) increases; in particular, to find out whether the convergence of estimates to the full sample value of D happens at a common sample size across different datasets. We also aim to look at how the optimism present in D, caused by model overfitting, changes with increasing sample size. A secondary aim is to investigate how D, and the optimism present in D, changes with the p value used in the model selection procedure.

3.2.1 Considerations for investigation

The basic plan of the investigation is to take a real clinical dataset consisting of failure time and censoring indicator for each patient, along with various covariates. Subsamples of the dataset of decreasing size will be randomly chosen and the best model selected using a reproducible method. D will then be calculated using the Stata command `str2d` (Royston, 2006b), and the various Ds for different sized subsamples compared.

There are several aspects of this plan that need further consideration before we design this investigation. Firstly, we must account for the optimism that will be inherent in estimates of D, since we are fitting a model and assessing its goodness-of-fit on the same dataset. Secondly, a model selection procedure needs to be chosen. Finally, consideration should be given to the subsample sizes used.

Optimism

Optimism is a consequence of model overfitting, and of using the same dataset to both estimate model parameters and evaluate model performance. One cause of overfitting is selection bias, which occurs when variables are selected for inclusion in a model by their statistical significance, such as in a stepwise procedure. In this situation, variables are more likely to be selected if they have a large estimated effect size in the dataset, regardless of their true effect. This causes an upward bias in the size of estimated predictor effects in the model (Copas and Long, 1991). Selection bias does not occur if the terms (variables and any transformations) to be included in the model are pre-specified. The problem is lessened when predictors truly have large effects – as they will almost always be selected for the model – and when sample size is increased. A closely related issue, parameter uncertainty, can occur regardless of model selection technique, because parameters are estimated in the model with uncertainty (Steyerberg, 2008). This adds a component of variance to the variance of the prognostic index across patients, and can result in overestimation at the extremes of the linear predictor, expressed as low predictions being generally too low and high predictions too high. It is related to regression to the mean.

The result of these issues may be a model containing too many variables, some with optimistic effect estimates. Some of the model variables are likely to be spurious and

contribute explanations of noise rather than true relationships in the data. Thus if model performance is assessed in the *same* dataset, it is likely to be over-estimated, presenting an optimistic view of how good the model is. It will probably perform less well in an independent dataset from the same population, because some (or much) of its apparent ability in the first dataset was due to explaining noise and other idiosyncrasies. The difference between the statistic as calculated in the first dataset and as calculated in the second is the optimism inherent in the first estimate.

This means that we must take steps to estimate the optimism present in D if we are to avoid overfitted models and inflated estimates of predictive ability. The easiest way to do this would be to split the dataset to form training and validation sets, develop the model in the training set and then evaluate D in both (Harrell, 2001). However, in this investigation we wish to see how D varies across a wide range of sample sizes, so reducing the size of the available datasets to start with is not desirable. Another possibility is to use the quantity D_{adj} which was introduced by Royston and Sauerbrei in their 2004 paper on D. D_{adj} is based on an adjusted R^2 which was developed to account for the known positive bias of R^2 in a linear regression model with Normal errors, and Royston and Sauerbrei found that indeed D_{adj} showed low bias in their tests using simulated and real data. D_{adj} accounts for the bias due to parameter uncertainty, so adjusts for one source of optimism, but does not adjust for the optimism caused by data dependent modelling techniques. We plan to use a reproducible method such as backwards elimination or stepwise selection to select models in this investigation, so this will mean an additional source of optimism which D_{adj} will not adjust for. Instead we will use the method described by Harrell et al. (1996) to estimate the optimism in D.

Harrell's method is based on Efron's (1983) refined bootstrap method, and can be used for any index of model performance where overfitting is a concern. The method is broadly as follows, but is described in more detail in Section 3.2.2. After estimating D in the original dataset, draw a sample with replacement of the same size n. Select a model and calculate D based on the bootstrap sample. Then predict from the *same* model using the original dataset and re-calculate D. The difference between these two values (D in the bootstrap sample $-$ D in the original data) is an estimate of the optimism inherent in the original D calculated. Repeating this for 100 or more bootstrap samples gives an

averaged optimism which can then be used to correct the original estimate for overfitting, and also provides an estimate of the standard error of the optimism.

Model selection procedure

In order to use this bootstrap method to estimate optimism, the same model selection procedure must be used with each bootstrap sample. Thus we must use an automatic procedure so it can be exactly replicated. Royston and Sauerbrei's (2008) multivariable fractional polynomials (MFP) method will be used, as implemented by the Stata command `mfp`. This combines backward elimination of variables, which is generally preferred to forward selection or a stepwise method (Harrell, 2001), with the use of fractional polynomials to flexibly model continuous variables.

The selection procedure will be repeated with various p values for variable selection, to observe how D varies as the strictness of the model inclusion criteria changes. The p values chosen were 0.01, 0.05, 0.157, 0.50 and 1.00. This list encompasses 'traditional' p values (0.01, 0.05), the p value corresponding to Akaike's information criterion for selecting a single variable (0.157) and the full model (1.00). $p = 0.50$ allows deletion of some variables without being too stringent, and bridges the gap between 0.157 and 1.00.

In general, higher p values result in larger models, and hence we would expect to see increasing values of D (since in general larger models mean better prediction) and increasing optimism (due to the increased likelihood of overfitting) as p rises.

Sub-dataset size

The sub-datasets chosen for each dataset should be a compromise between getting a good spread of sample sizes to base analyses on, while ensuring that the planned protocol will run in a reasonable time frame. Steps of at least 3 EPV, and not more than 10 EPV between sub-datasets should ensure we get good coverage across the range from the minimum to the complete dataset. This means that the values chosen will vary between datasets. Another consideration is that the modelling procedure may fail for very small samples. Starting with a lowest EPV of around 5 should minimise this (Vittinghoff and McCulloch, 2007). For simplicity, the sample size n will be chosen rather than the number of events; this means that the censoring proportion will not be the same across all the subsamples.

3.2.2 Methods

The protocol for investigation of D is as follows.

1. Choose integers $n_1, n_2, ..., n_x$ such that $0 < n_1 < n_2 < ... < n_x < N$, where N is the number of records in the full dataset.

2. For each chosen n_i, randomly select a sub-dataset of size n_i.

3. Use the MFP method to select the best model for the chosen sub-dataset.

4. Calculate D for this model, in this sub-dataset. Call this quantity D_{orig}.

5. Bootstrap 100 samples from the sub-dataset. For each sample:

   (a) Use the MFP method with the same list of candidate variables to select the best model for the bootstrap sample. Call this model A.

   (b) Calculate D for model A, in this bootstrap sample. Call this quantity D^{boot}_{boot}.

   (c) Calculate D for model A, in the original sub-dataset from step 2. Call this quantity D^{orig}_{boot}.

6. Calculate the average estimated optimism op over the bootstrap samples.

$$op = \frac{1}{100}\sum_{j=1}^{100}(D^{boot}_{boot(j)} - D^{orig}_{boot(j)}),$$

   where $D^{X}_{boot(j)}$is the D^{X}_{boot} from the jth bootstrap sample.

7. Subtract the estimated optimism from the original estimate of D to get the optimism-adjusted D, called D_{opt}.

$$D_{opt} = D_{orig} - op$$

Steps (2) to (7) will be repeated for each chosen subset size n_i, and each p for variable selection within the MFP procedure, p =0.01, 0.05, 0.157, 0.50 and 1.00. Stata code is provided in Appendix C for the .ado file written to perform the bootstrap procedure used in steps (3) to (7).

Datasets

In order to draw generalisable conclusions from this investigation, especially since we only select one subdataset of each sample size for each dataset, we need to perform the procedure on various different datasets. The datasets used in this chapter are FBC (breast cancer), APC (prostate cancer), GLI (glioma), RBC (breast cancer), MYE (myeloma), PBC2 (primary biliary cirrhosis), LVA (lung cancer), KCA (renal cancer), SEER (breast cancer; 9 separate datasets). Full details of all these datasets can be found in Appendix A, and the Stata `mfp` command lines used for each one in Appendix B.

For this chapter, three additional datasets RBC5, RBC10 and RBC 15 were formed by adding 5, 10 and 15 uniformly distributed uncorrelated random variables respectively to the RBC datasets to make three new datasets. These are used to investigate the effect of additional noise variables on optimism and D.

Due to the size of the SEER datasets only two p values were used when working with these data, $p = 0.05$ and $p = 1.00$.

Practical issues

The investigation and analyses were performed in Stata 10 (StataCorp, 2000). Results are structured so as to separate the results for D_{orig}, optimism and D_{opt}. Within each of these three sections, the results relating to the relationship of the quantity with EPV and with p for variable selection are described. Selected results are also displayed graphically to aid understanding and interpretation. On the whole these graphs are simple plots of one quantity against EPV, number of events, or p. Note that we define the number of variables in the EPV calculation as a simple count of variables in the `mfp` command line (which is given for all datasets in Appendix B).

3.2.3 Results

The first section gives example graphs from the investigation to further illustrate the procedure outlined in Section 3.2.2. Sections 3.2.4, 3.2.5 and 3.2.6 give results pertaining to D_{orig}, optimism, and D_{opt} respectively and each is divided into 3 sub-sections. These cover the relationship between the statistic and sample size, the relationship between the statistic and p for model selection, and the differences seen across datasets. The results

concerning sample size are drawn from all 17 datasets considered. The results concerning p are drawn from all 17 datasets, however the 9 SEER datasets contribute less as only $p = 0.05$ and $p = 1.00$ were used for these.

Example output

For each combination of dataset and p, two graphs were created, which were inspected to check that the procedure was working as expected. An example of each of these is given to aid understanding of the steps of the investigation outlined in Section 3.2.2.

The first of these is a panel of scatter plots, one for each sub-dataset size, giving the output from the bootstrap process. An example is given in Figure 3.1. These show that the values of D_{boot}^{orig} (D of model estimated in bootstrap sample, fitted to original dataset) have a narrower range than those of D_{boot}^{boot} (D of model estimated in and fitted to bootstrap sample); this would be expected since the latter quantity contains more optimism and so will range higher. The plots also show that outlying low values of D_{boot}^{orig} are seen in some instances. This may occur when one or more particularly influential datapoints are repeated in the bootstrap sample, which can cause the resulting model to fit to this idiosyncrasy in the data. Once this model is fitted back on the original sample, it loses most of its predictive ability, as the original dataset does not contain this cluster of influential points so the model is a poor fit. This results in a very low D_{boot}^{orig}. Such outliers may affect the mean D_{boot}^{orig} which in turn affects the estimates of optimism and D_{opt}. However upon inspection of these plots, it was felt that none of the datasets considered in this investigation appeared to have a severe problem with outliers, so we did not attempt to mitigate their effects.

The second type of graph is illustrated in Figure 3.2. This is a connected scatter plot showing how the final values of D_{orig}, D_{boot}^{boot}, D_{boot}^{orig}, D_{opt}, optimism, and standard error of D_{opt} vary as n increases (averaged over the 100 bootstrap replications). Figure 3.2 shows examples of these graphs for for two datasets (FBC & RBC) and one p value (0.05). The patterns in the various D and optimism can be seen, as well as the decreasing standard error in D_{opt} as the subsample size increases.

As mentioned above these graphs are numerous, being available for each dataset and p combination. Thus they are not presented further in this chapter; instead graphs are

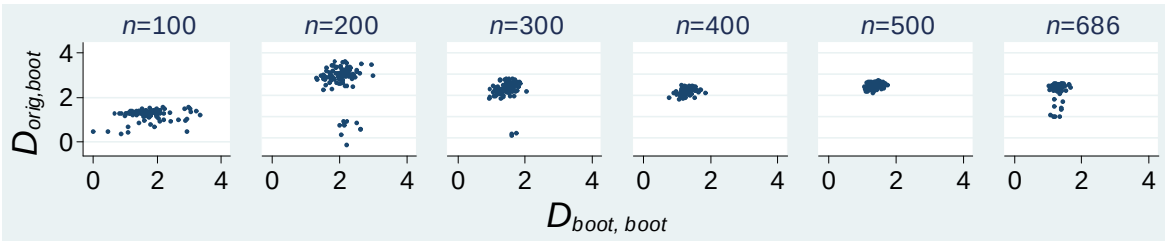

Figure 3.1: FBC: Scatter plots of D_{boot}^{orig} vs D_{boot}^{boot} for $p = 0.05$, for each chosen n

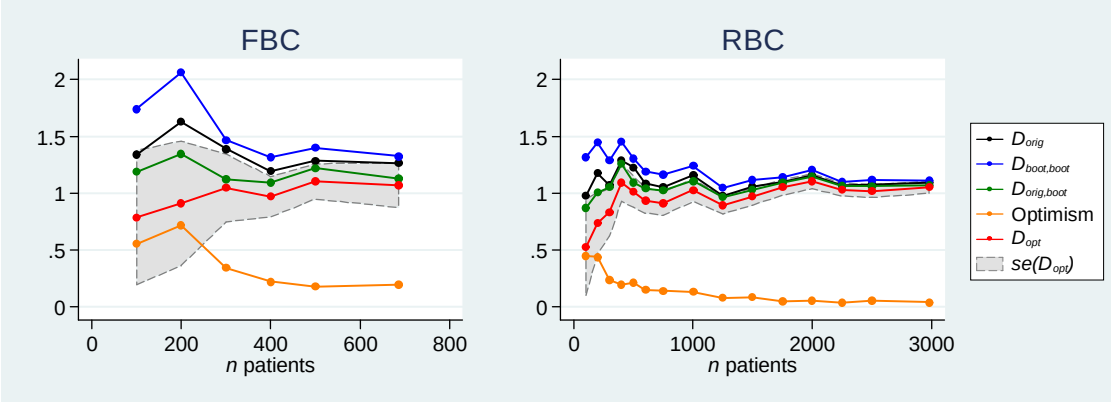

Figure 3.2: D_{orig}, D_{boot}^{boot}, D_{boot}^{orig}, D_{opt}, optimism, and $se(D_{opt})$ vs n for FBC & RBC, $p = 0.05$

presented which aggregate results across one or more of EPV, p, or dataset. The next three sections report results and patterns across datasets.

3.2.4 Results for D_{orig}

Results are presented below to illustrate the relationships between D_{orig} and sample size, D_{orig} and p for variable selection, and to show how D_{orig} varies across the datasets used.

D_{orig} over changing sample size

Figure 3.3 shows the profile of D_{orig} across EPV for each (non-SEER) dataset and p-value. There is quite high variability amongst values of D_{orig} up to around 10 EPV, with values tending to start high and then decrease. The profile seems to be flatter for lower p. From about 20 EPV, D_{orig} remains fairly stable. However four of the datasets have a maximum EPV of around 20 (or less), so it is difficult to see any firm patterns. The standard error of D_{orig} is also shown on these graphs (for $p = 0.05$ only) to illustrate how the uncer-

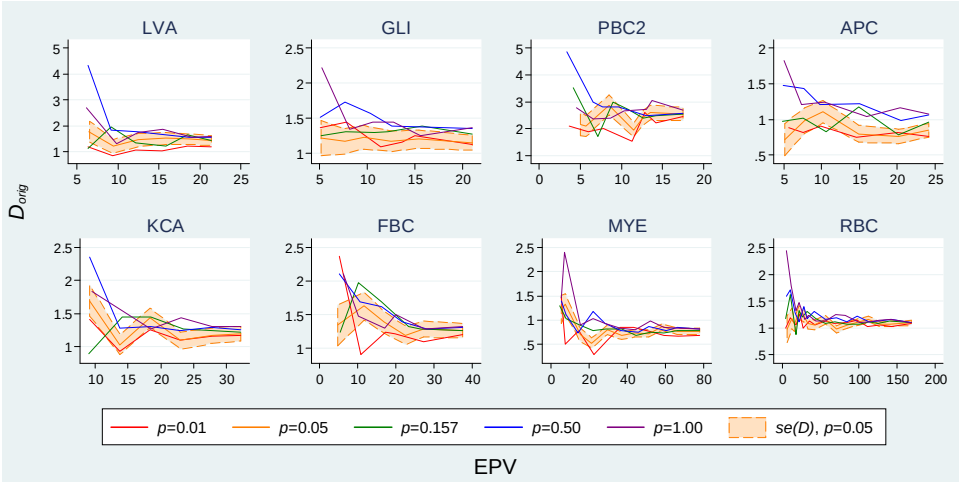

Figure 3.3: Matrix of plots of D_{orig} vs EPV; $se(D_{orig})$ at $p = 0.05$ shown. Each graph contains results for one dataset, for each p value (note differing X and Y axis scales)

tainty around the estimate varies. As expected from Figure 3.1, there is a decrease in the standard error with increasing number of events.

Figure 3.4 plots D_{orig} vs number of events for each (non-SEER) dataset and p-value. Obviously patterns are similar to those seen in Figure 3.3; it is difficult to pinpoint a particular number of events above which values of D_{orig} are stable.

D_{orig} over changing p

In considering the relationship between D_{orig} and p, we look primarily at the $n = N$ case, where results should be most stable. When $n = N$, D_{orig} slightly increases with p (Figure 3.5). In fact the profile of D_{orig} over increasing p is non-decreasing between p values, for all 8 datasets. The profile appears to be flatter for larger datasets (MYE, RBC, SEER).

D_{orig} over changing p in RBC noise datasets

To see whether the relationship between D_{orig} and p changes with increasing number of noise variables, Figure 3.6 plots D_{orig} vs p for the four RBC datasets. The same model is chosen for all four datasets when $p = 0.01$ and $p = 0.05$, and for the three noise datasets when $p = 0.157$. However for $p = 0.5$ and $p = 1.0$, different models are chosen and in this example the D_{orig} appears to increase more for the datasets that have more noise variables added. The scale of change in D_{orig} is however quite small across the four datasets; at $p = 1.0$ the difference between the highest and lowest D_{orig} is just 0.06.

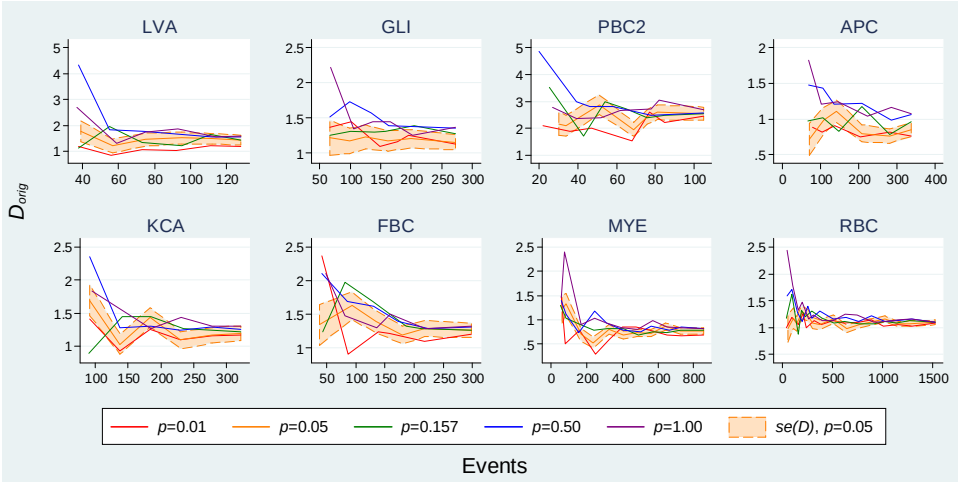

Figure 3.4: Matrix of plots of D_{orig} vs events; $se(D_{orig})$ at $p = 0.05$ shown. Each graph contains results for one dataset, for each p value (note differing X and Y axis scales)

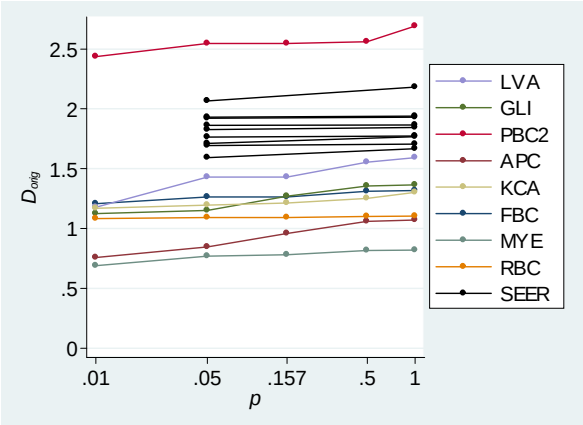

Figure 3.5: D_{orig} vs p (log scale) for full datasets ($n = N$)

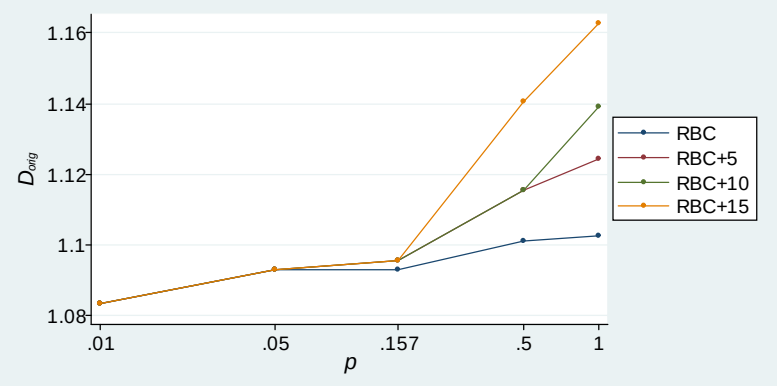

Figure 3.6: D_{orig} vs p for original RBC dataset, and for dataset with 5, 10 and 15 noise variables added

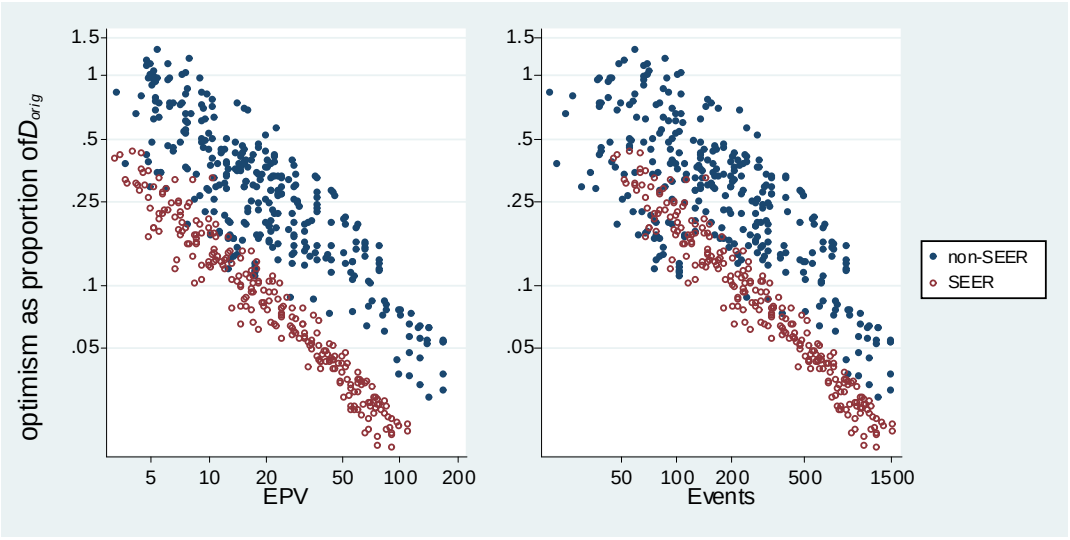

Figure 3.7: Optimism vs EPV (left) and vs events (right). Each point corresponds to one dataset / n / p combination, with SEER datasets identified. Note log scale for both axes.

D_{orig} across datasets

It can be clearly seen from Figure 3.5 that D_{orig} varies in magnitude across the datasets used. The cirrhosis dataset PBC2 has by far the highest D_{orig} at around 2.5 ($R^2_D = 60\%$); the SEER datasets vary from 1.59-2.07 (38%-51%) and the others from about 0.7-1.5 (10%-35%) when $n = N$. This wide range of explained variation reflects the different diseases represented and the varying amounts of prognostic information in the covariates. Even within a particular disease, D will be affected by the case mix in the study; a more heterogenous group means more scope for differentiating patients and thus greater potential for a higher value of D.

3.2.5 Results for optimism

Optimism over changing sample size

Figure 3.7 shows the proportion of optimism in D_{orig} for every combination of dataset, n and p, plotted against EPV and number of events respectively. The pattern is clear: optimism decreases as EPV and number of events increases. The SEER and non-SEER datasets appear to form two bands, with the SEER datasets showing markedly lower optimism at the same EPV. The optimism in the SEER datasets drops below 10% once EPV>20 (approximately), or once there are more than 200 events. For the non-SEER

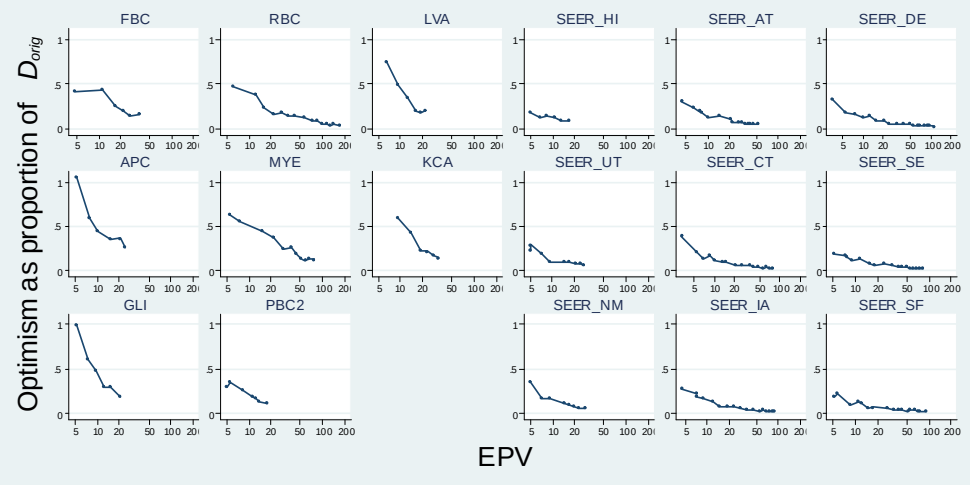

Figure 3.8: Optimism vs EPV for all datasets, $p = 0.05$. Note x axis is log scale.

datasets EPV>70 or more than 500 events seem to be required for 10% optimism; although there is more variation for these datasets.

Figure 3.8 contains graphs of optimism vs EPV separately for each dataset (for $p = 0.05$ only) and shows that the same pattern is present over all datasets. Figure 3.10 shows the same graphs for $p = 1.00$; the pattern clearly still holds when no selection is used. Optimism seems to show a similar profile against EPV for all the datasets investigated: it decreases fairly sharply from the minimum EPV to 20-30 EPV, and then continues to decrease more gently as EPV increases further. For datasets with high EPV (RBC, the larger SEER datasets), the optimism continues to decrease even as EPV passes 50 or 100; albeit in much smaller increments. Figures 3.9 shows optimism vs number of events for each dataset, for $p = 0.05$. Similar patterns are seen as for EPV.

Optimism over changing p

When n is much smaller than N, optimism varies widely with increasing p. As n approaches N these fluctuations lessen and a pattern appears which is seen across the 8 datasets used. In general, when $n = N$ (where optimism is lowest and most stable), optimism increases with p (Figure 3.11). There were individual deviations from this general pattern between datasets; a couple showed decreasing optimism from one p to a higher p, or sharper increases over the lower p values.

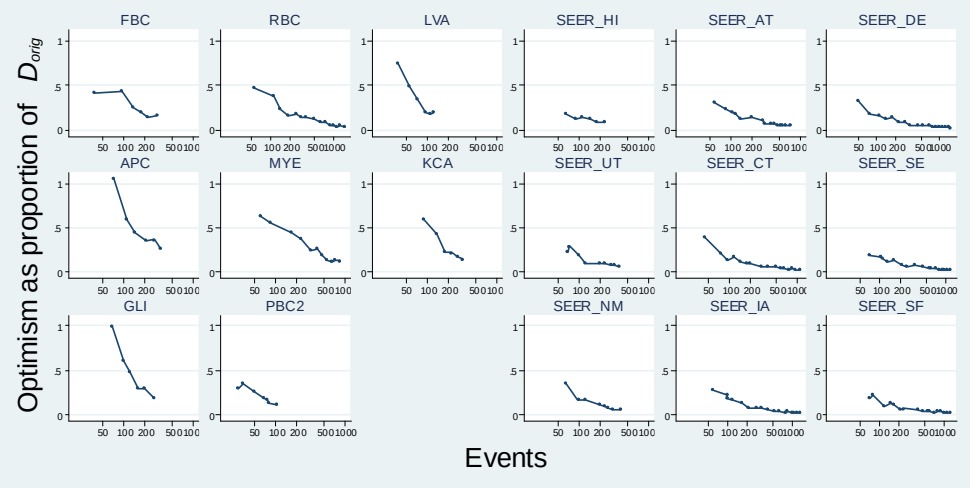

Figure 3.9: Optimism vs events for all datasets, $p = 0.05$. Note x axis is log scale.

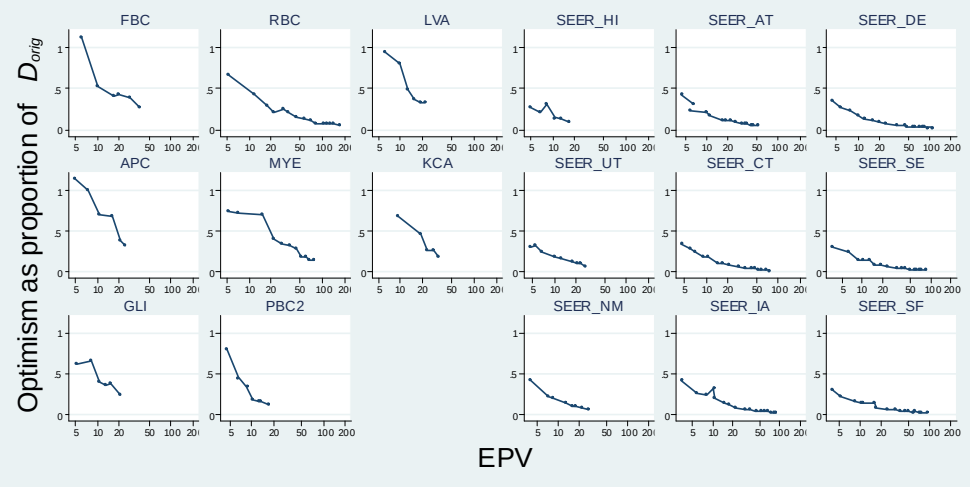

Figure 3.10: Optimism vs EPV for all datasets, $p = 1.0$. Note x axis is log scale.

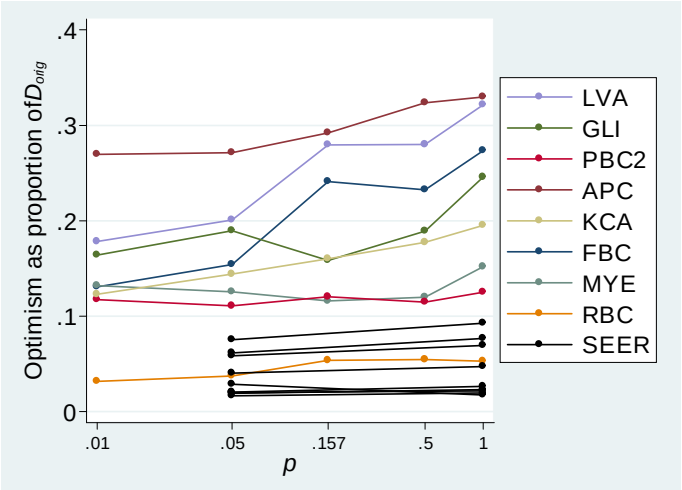

Figure 3.11: Optimism vs p (log scale) for full datasets ($n = N$)

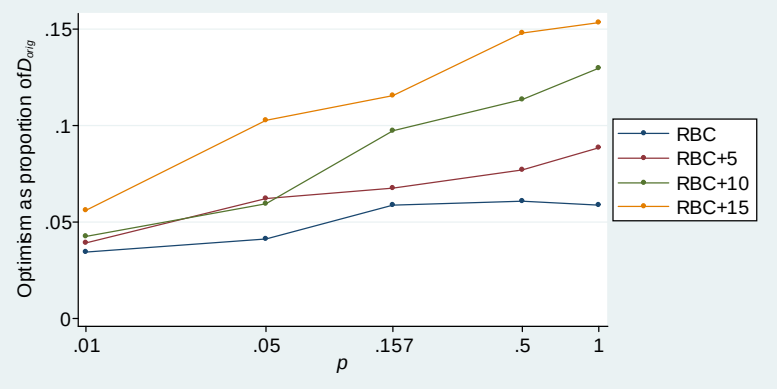

Figure 3.12: Optimism vs p for original RBC dataset, and for dataset with 5, 10 and 15 noise variables added

Optimism over changing p in RBC noise datasets

To see whether the relationship between optimism and p changes with increasing number of noise variables in the dataset, Figure 3.12 plots optimism vs p for the original RBC dataset and the 3 with additional noise variables added. It shows that the relationship between optimism and p does appear to be stronger for the datasets with added noise variables, and that the amount of optimism increases roughly in line with the number of noise variables added. However in absolute terms, the amount of optimism is small for this dataset and so are the increases seen.

Optimism across datasets

The amount of optimism present (as a proportion of D_{orig}) when $n = N$ does vary quite widely between datasets, even accounting for the different maximum EPV in each dataset. This can be clearly seen in Figures 3.8 and 3.10.

Optimism is much lower in the SEER datasets than other datasets. This is clearly seen in Figure 3.7, where the SEER datasets are identified. The SEER points appear to split off from the other datasets, most clearly from about 30 EPV onwards, but have lower optimism throughout the whole range of EPV. For example, at 5 EPV the SEER datasets have less than 40% optimism, and often less than 20%. This compares with more than 50% and sometimes 100% for the other datasets at this EPV.

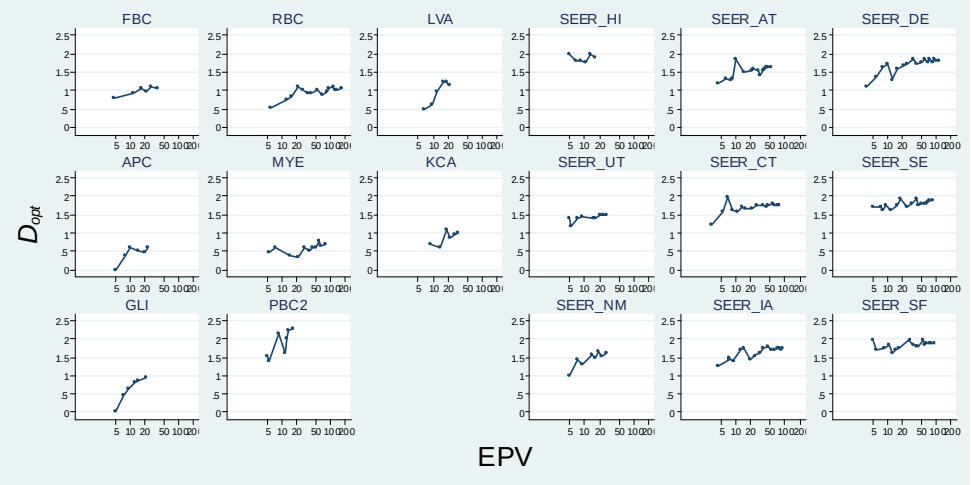

Figure 3.13: D_{opt} vs EPV for each dataset, $p = 0.05$. Note x axis is log scale.

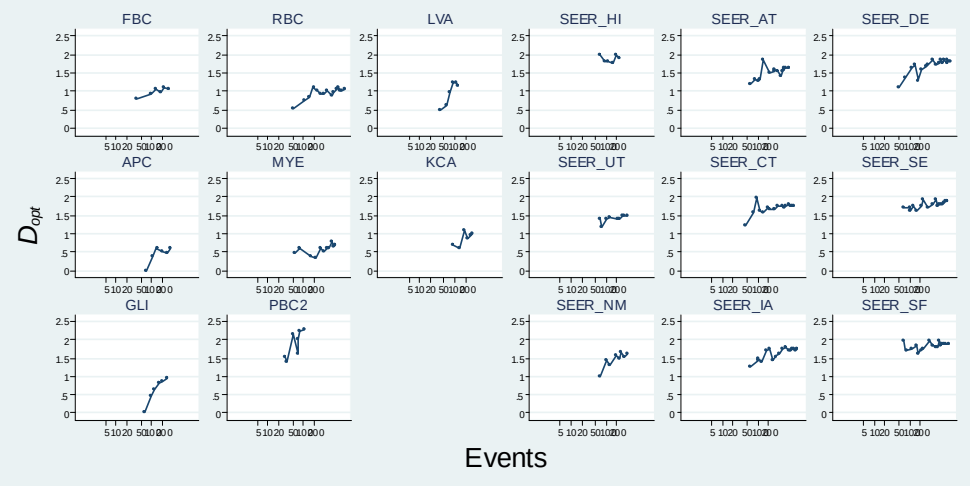

Figure 3.14: D_{opt} vs events for each dataset, $p = 0.05$. Note x axis is log scale.

3.2.6 Results for D_{opt}

D_{opt} over changing sample size

We have already seen that D_{orig} decreases with increasing EPV for most datasets, sharply at first and then more gently. Optimism also showed a strong inverse relationship with EPV across all datasets. Since $D_{opt} = D_{orig} - optimism$, we might expect D_{opt} to show a relatively flat profile against EPV.

As can be seen in Figures 3.13, 3.15, D_{opt} can be quite variable when EPV is low and this variability decreases as EPV increases. Despite these fluctuations there is a clear pattern: D_{opt} increases with EPV. The level of increase varies across the datasets; for

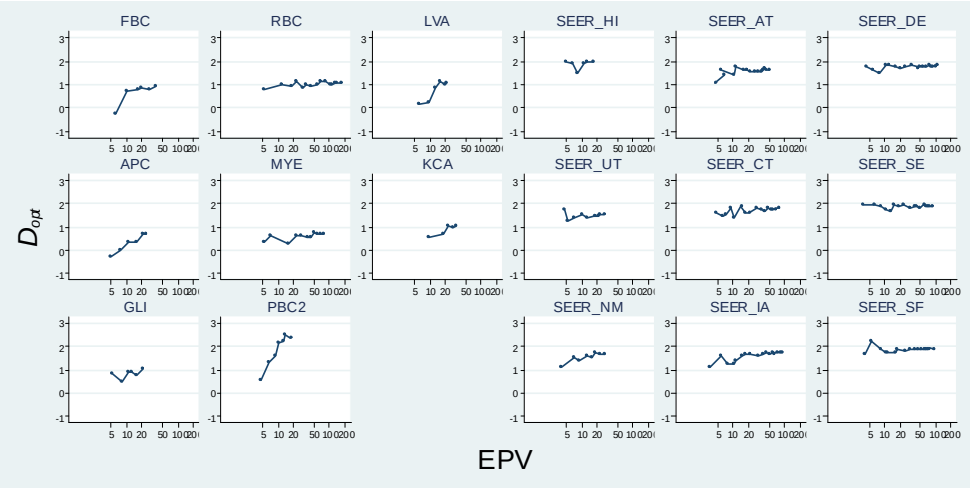

Figure 3.15: D_{opt} vs EPV for each dataset, $p = 1.00$. Note x axis is log scale.

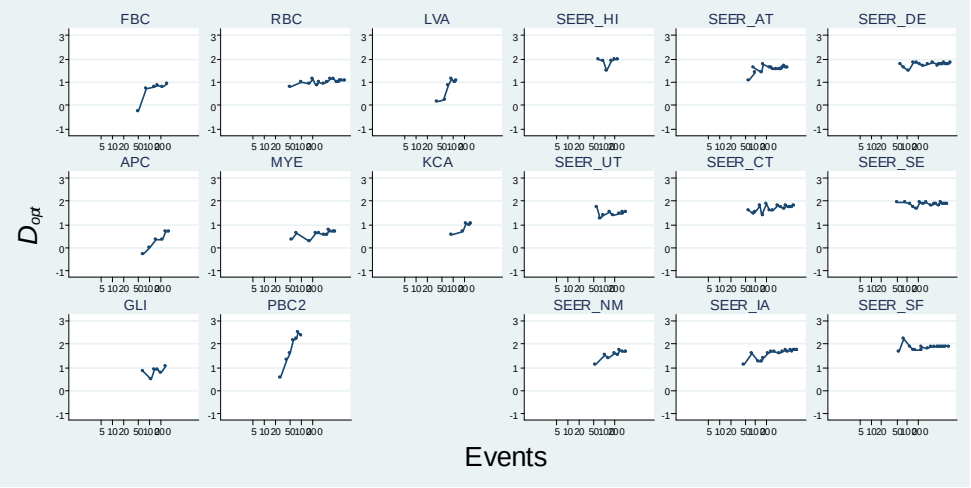

Figure 3.16: D_{opt} vs events for each dataset, $p = 1.00$. Note x axis is log scale.

datasets with low optimism at low EPV (for example the SEER datasets), the increase is less noticeable in the graph. In general, the increase occurs up to about 25 EPV. From about 25 or 30 EPV, further increases in D_{opt} are not easily distinguished from random fluctuations, which may still be as large as 20% of the value of the statistic. Once 50 EPV is reached, D_{opt} continues to fluctuate about a 'middle' value to a greater or lesser degree, however the variability is reduced, with fluctuations mostly $<$10% of the value of D_{opt}.

Figures 3.14, and 3.16 plot D_{opt} against number of events. The same patterns are seen as with EPV but there is no common number of events across all the dataset which could be said to be sufficient for a reliable estimate of D_{opt}.

In order to quantify the convergence of D_{opt} as n approaches N, we combined results for all datasets, n and p into one plot. For simplicity denote $D_{opt,N}$ as the D_{opt} for a particular dataset and p when $n = N$ (i.e. in the full dataset). Figure 3.17 shows the size of the D_{opt} relative to $D_{opt,N}$ for the same dataset and p value as EPV increases, with one point for each dataset / n / p combination (excluding the full datasets). When considering this plot, it must be remembered that it contains results from 17 heterogenous datasets. Going from 5 to 20 EPV improves the accuracy of the estimated D_{opt} greatly as from 20 EPV the vast majority of estimates are within 50% of $D_{opt,N}$. From 30 EPV almost all are within 25%. A smoothed line fitted to the scatter of points suggests that if a dataset has less than 30 EPV, it is likely that D_{opt} from the dataset is underestimating the 'true' D_{opt}. From about 30 EPV upwards, it is as likely to be overestimating as underestimating.

Figure 3.18 plots the same y-axis as Figure 3.17 against number of events. In this plot, the majority of datasets with more than 200 events were within 25% of $D_{opt,N}$. Datasets with 300 events or fewer were more likely to be underestimating the 'true' D_{opt} than overestimating.

D_{opt} over changing p

We have seen that D_{orig} increases slightly with p, with this increase lessened for larger datasets. Optimism as a percentage of D_{orig} also appears to increase slightly with p. Since $D_{opt} = D_{orig} - optimism$, we might expect not to see a strong relationship between D_{opt} and p.

Figure 3.19 shows that this is indeed the case, as the profile of D_{opt} against p varies across the datasets chosen, so no firm conclusions can be drawn. The highest value of

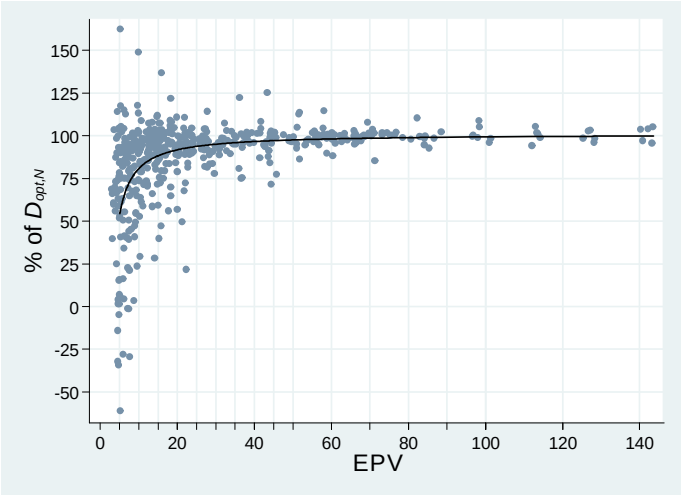

Figure 3.17: Scatter plot of the % of $D_{opt,N}$ which each D_{opt} attains when $n < N$, vs EPV. Each dot represents one dataset / n / p combination. Line is fractional polynomial smoother.

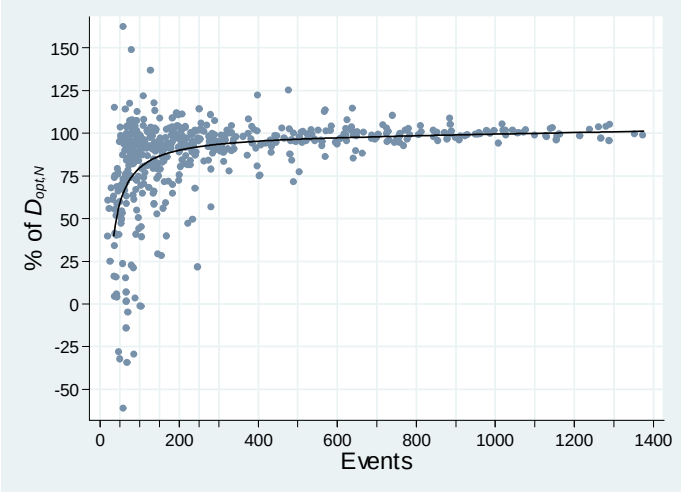

Figure 3.18: Scatter plot of the % of $D_{opt,N}$ which each D_{opt} attains when $n < N$, vs number of events. Each dot represents one dataset / n / p combination. Line is fractional polynomial smoother.

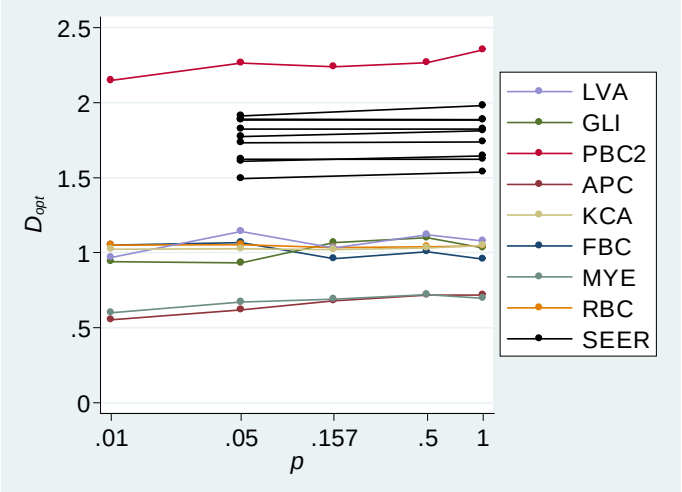

Figure 3.19: D_{opt} vs p (log scale) for full datasets ($n = N$)

D_{opt} occurs variously at $p = 0.05$, 0.50 and 1.00 for the 8 datasets. The larger datasets seem to have a flatter profile for D_{opt} against p than smaller ones, as was seen for D_{orig}.

D_{opt} across datasets

D_{opt} varies quite substantially between the 17 datasets used. As might be expected, the 9 SEER datasets had quite similar values for D_{opt}, with the D_{opt} for the four largest datasets varying between 1.73 and 1.89 (R^2_D 42% – 46%). The dataset with the highest D_{opt} was the PBC2 database, with $D_{opt} = 2.3$ (104 events); this corresponds to $R^2_D = 56\%$. This was much higher than any other dataset, the next highest non-SEER datasets were FBC and RBC at $D_{opt} \simeq 1$ ($R^2_D = 19\%$). APC and MYE datasets showed the lowest D_{opt}, both at $\simeq 0.7$ ($R^2_D = 10\%$).

3.2.7 Conclusion

This investigation has elucidated some relationships between D and sample size which have not been previously studied. These are described and commented on in this section.

D_{orig}

The wide range of values of D_{orig} seen across the datasets investigated here – equivalent to levels of explained variation between 10% and 60% – likely reflects the different diseases represented and the varying amounts of prognostic information in the covariates.

Optimism is still included in D_{orig} and again this will have an effect on its magnitude in a particular dataset.

We found that D_{orig} decreased as sample size increased. Overfitting is worse in small datasets and this is likely why D_{orig} is so high when sample size is very low. The smaller n is, the better a fitted model will be at predicting outcomes in *that* dataset, leading to a high D_{orig}. As the sample size increases the models fitted don't have the same predictive accuracy, so D_{orig} decreases.

D_{orig} appeared to increase slightly with p, but the slope of the increase seemed to be flatter for datasets with higher number of events. An increase of D_{orig} with p might be expected, since the models selected get larger as p increases, and more model terms mean better prediction. Thus the potential of D_{orig} to increase with p may depend on the number of potential covariates available; however, we did not see a greater increase in D_{orig} with p for the RBC datasets with added noise variables. It may be that there is an effect but more than 15 noise variables are needed to see it; unfortunately with limited computing power we cannot investigate this further.

The fact that D_{orig} increased less with p in datasets with higher number of events is again due to the fact that a model will have better apparent predictive ability in a small dataset than a large one. If there are few events, adding a couple of terms to the model could potentially make a big difference to the model's predictive ability (e.g. going from 3 terms to 5 terms in a model used to predict 20 events will appear to greatly assist prediction). If there are many events, adding a couple of terms to the model does not make as much difference to the model's predictive ability (e.g. going from 3 terms to 5 terms in a model used to predict 200 events).

Optimism

Our investigation has highlighted the importance of accounting for optimism when developing a prognostic model. The method we used to estimate optimism (Harrell et al., 1996) found that with 25–30 EPV, the median optimism in D_{orig} was 18% and the maximum was 40%. Even with 50 EPV, 20% optimism was seen in some datasets. We have not validated Harrell et al.'s method here, but in their work Royston and Sauerbrei (2004) found that it may overestimate optimism, causing D_{opt} to be overcorrected (too low). Despite this, it is still clear that prognostic ability is likely to be overestimated to some extent

if overfitting and parameter uncertainty are not accounted for, even in some datasets with high EPV.

We found that the SEER datasets showed consistently lower optimism in estimates (expressed as a proportion of D) than others. While this may be explained by the fact that the values of D in the SEER datasets were higher than in most of the other datasets used, we speculate that it may also be partly because the SEER datasets are designed to contain only variables which are known to have some prognostic value. Thus in a model they explain true relationships in the data rather than noise, which should result in lower optimism.

A clear pattern of optimism decreasing with increasing number of events was observed, and such a relationship is expected; since the smaller the sample size, the better a fitted model will be at predicting outcomes in that dataset. However, this prediction ability is mostly spurious and the model is likely to be severely overfitted, leading to high optimism.

In the full dataset, optimism showed an increase with increasing p. To explain this pattern we must consider the two causes of optimism. We have already seen that D_{orig} increases slightly with p as the models selected contain more covariates. However, the new 'extra' terms now included have lower statistical significance and so are more likely to explain noise than true relationships in the data – increasing overfitting. Furthermore, parameter uncertainty increases as more variables enter the model, again contributing to higher optimism. Under these explanations, we might expect to see more of an increase in optimism across p for datasets containing more variables. Our investigation of the datasets with added noise variables did support this hypothesis but we cannot draw any firm conclusions.

D_{opt}

Below 30 EPV, estimates of D_{opt} were quite variable, but showed a clear increase with sample size. Above around 30 EPV, the estimated value of D_{opt} did not increase much further but fluctuated around what what appears to be a stable value. The fluctuations lessened in magnitude as EPV increased further. We are unsure of the reason for the positive relationship between D_{opt} and sample size in the 0–30 EPV range but it may be related to the finding of Royston and Sauerbrei (2004) that Harrell et al.'s (1996) method

for measuring optimism may overestimate the quantity; or an artefact from using the bootstrap method in small samples. Another possibility arises if we consider the idea of the 'true' (but unknown) model using the parameters selected, which must give the highest possible D_{opt}. As the size of the dataset increases our model selected approaches the true model and thus D_{opt} is seen to increase.

From the datasets studied in this chapter we have gathered some evidence which may be useful in eventually determining some specific recommendations. Among our heterogenous group of datasets, we observed that if we have 30 EPV or more; or 200 events or more, the vast majority of estimated D_{opt} values were within 25% of the stable value of D_{opt} for their dataset. Up to about 30 EPV, or $\sim$300 events, more estimated values of D_{opt} were lower than the stable D_{opt}, than were higher.

We also considered the effect that the p value used for model selection in the MFP procedure has on D_{opt}. This investigation did not find a strong relationship between D_{opt} and p. Again this emphasises the importance of eliminating optimism from estimates; since as already mentioned D_{orig} increases with p, which might lead researchers looking for the highest possible predictive ability to always choose the full model.

3.3 Investigation of $SE(D)$

Some of the data produced in Section 3.2 suggested that the method used in Royston and Sauerbrei (2004) (and implemented in Stata command `str2d`) to estimate the standard error of D may be underestimating this quantity. We need to ensure that our estimate of the standard error is accurate enough before we go on to try and develop sample size calculations. To investigate this further we carry out a simulation study to compare various possible estimators of $SE(D)$ in a systematic way.

3.3.1 Aims

The aim of this investigation is to explore the best way to estimate $SE(D)$, using simulated data. Specifically, we wish to determine more conclusively whether the method in Royston and Sauerbrei (2004) and used by the Stata command `str2d` (producing a quantity we term $SE(D)_{cox}$) is underestimating the true value of $SE(D)$. If we do find that this is the case, we hope to gain some idea of the magnitude of the underestimation and whether any data or model parameters influence this. We also wish to investigate an alternative estimator of $SE(D)$, a bootstrap estimator which we term $SE(D)_{boot}$.

3.3.2 Methods

Broadly, a survival dataset and normally distributed model PI is simulated; D and $SE(D)_{cox}$ are calculated for the model using `str2d` and $SE(D)_{boot}$ is calculated using bootstrapping. This is repeated 500 times and the empirical standard error generated: the standard deviation of the 500 values of D from the repetitions. We will call D's empirical standard error $SD(D)$.

Simulations of survival data are performed in Chapters 5 and 6 as well as this chapter, and here we outline in detail the method used to produce independent identically distributed (iid) datasets for all of these studies.

Generation of simulated time-to-event data in this thesis

Generation of survival times The method of simulation used is that described by Bender et al. (2005), who described how a random variable with U[0,1] distribution (called

U) can be transformed to survival times T_s of the Cox model by using the equation

$$T_s = H_0^{-1}[-\log(U)\exp(-\beta' X)], \tag{3.1}$$

where H_0 is the baseline (cumulative) hazard function (the precise form of which depends on the desired distribution of survival times), X is the vector of covariates and β is the vector of regression coefficients (log hazard ratios). In all simulations performed in this thesis, we wish to use an exponential distribution for survival times and thus set the baseline hazard to be a scalar θ; thus the cumulative baseline hazard function is $H_0(t) = \theta t$ and its inverse $H_0^{-1}(t) = \theta^{-1}t$. Specifically, we set $\theta = 0.002$ for most simulations. Since simulating a full multivariable vector is complex both computationally and in terms of interpretation, we chose to instead use a surrogate variable X for the multivariable index, so β and X were effectively scalar. The surrogate variable X was simulated as normally distributed: $X \sim N(0, 1)$, so that the resulting prognostic index βX was also normal.

Generation of censoring times In many of our simulations we wish to consider censored data. We obtained random non-informative right-censoring by using the equation 3.1 to simulate a censoring time (T_c) for each record, again these are exponentially distributed (note that T_c were not dependent on x). Records where $T_c < T_s$ were considered censored at time T_c; records where $T_s < T_c$ were considered failures at time T_s. The desired censoring proportion was achieved by changing h_0; the baseline hazard required depends on β and was determined through an iterative process.

Detailed protocol for investigation of $SE(D)$

1. Generate dataset with N records and approximate censoring proportion *cens* using the method outlined above. Note this means the number of events e is not fixed.

2. Calculate D and $SE(D)_{cox}$ for the model PI using `str2d`. Note, D is not adjusted for optimism in this simulation or any other simulations presented in this thesis.

3. Bootstrap 500 samples from the simulated dataset. For each sample, calculate D for the model PI in the bootstrap sample.

4. Calculate and record the standard deviation of D from the 500 bootstrap samples; this is $SE(D)_{boot}$.

5. Repeat steps (1) to (4) 500 times.

6. Calculate the empirical standard error $SD(D)$: the standard deviation of D over the 500 simulations.

These steps will be repeated for various values of N (150, 300, 600, 1000, 2000), censoring proportion *cens* (0%, 40%, 80%) and β (0.5, 1.0, 2.0). Note that 'manual' bootstrapping as implemented in step (3) is not a requirement of this method; we could have used the bootstrap variance option available in `str2d`.

3.3.3 Results

Tables 3.1, 3.2 and 3.3 contain the results of this investigation for β=0.5, 1.0 and 2.0 respectively. Note table column headers: N is the number of patients, $\bar{e}$ is the average number of events in the 500 datasets simulated, $\overline{D}$ is the average value of D over the 500 simulations, $SD(D)$ refers to the standard deviation of D over the 500 simulations, $SE(D)_{cox}$ refers to the mean of the 500 values of $SE(D)_{cox}$ produced by the `str2d` command, and $SE(D)_{boot}$ refers to the mean of the 500 values of $SE(D)_{boot}$. The biases presented for the estimators of $SE(D)$ are calculated relative to $SD(D)$, these were calculated using the Stata command `simsum` (White, 2010).

These tables show clearly that $SE(D)_{cox}$ is almost always negatively biased (in 44 out of the 45 combinations of parameters in Tables 3.1, 3.2 and 3.3 it is lower than $SD(D)$) which means the variance of D is being consistently underestimated, as suspected. The magnitude of the underestimation increases markedly with β. In almost all cases $SE(D)_{boot}$ is less biased than $SE(D)_{cox}$ and the biases are both positive and negative (although slightly more often negative). The magnitude of bias in $SE(D)_{boot}$ does not appear to change with β in relative terms, although there is a slight increase in absolute terms with β.

						$SE(D)_{cox}$				$SE(D)_{boot}$			
β	*cens*	N	$\bar{e}$	$\overline{D}$	$SD(D)$	mean	abs. bias		rel. bias	mean	abs. bias		rel. bias
0.5	0	150	150	0.80	0.1619	0.1479	-0.0140	♦	-8.7%	0.1541	-0.0078	♦	-4.8%
		300	300	0.80	0.1096	0.1031	-0.0065	♦	-5.9%	0.1081	-0.0015	♦	-1.3%
		600	600	0.80	0.0757	0.0724	-0.0034	♦	-4.4%	0.0759	0.0001	♦	+0.2%
		1000	1000	0.80	0.0577	0.0560	-0.0017	♦	-3.0%	0.0585	0.0008	♦	+1.5%
		2000	2000	0.80	0.0427	0.0395	-0.0032	★	-7.6%	0.0413	-0.0014	♦	-3.3%
	40	150	89.8	0.80	0.1875	0.1851	-0.0024	•	-1.3%	0.1937	0.0062	•	+3.3%
		300	179.9	0.80	0.1408	0.1297	-0.0110	♦	-7.8%	0.1340	-0.0068	♦	-4.8%
		600	359.9	0.80	0.0934	0.0908	-0.0026	♦	-2.8%	0.0937	0.0003	♦	+0.4%
		1000	601.3	0.80	0.0698	0.0699	0.0001	♦	+0.2%	0.0719	0.0021	♦	+3.1%
		2000	1200.7	0.80	0.0531	0.0493	-0.0038	★	-7.2%	0.0510	-0.0022	♦	-4.1%
	80	150	29.3	0.84	0.3335	0.3213	-0.0122	•	-3.7%	0.3246	-0.0088	•	-2.6%
		300	59.3	0.80	0.2193	0.2182	-0.0011	•	-0.5%	0.2220	0.0027	•	+1.3%
		600	119.3	0.80	0.1608	0.1522	-0.0086	♦	-5.3%	0.1537	-0.0070	♦	-4.4%
		1000	198.6	0.80	0.1196	0.1176	-0.0020	♦	-1.7%	0.1195	-0.0002	♦	-0.1%
		2000	396.9	0.80	0.0821	0.0827	0.0006	♦	+0.7%	0.0838	0.0017	♦	+2.1%

Legend for standard error (*se*) of absolute bias: • $\Rightarrow 0.001 \leq se < 0.01$. ♦ $\Rightarrow 0.0001 \leq se < 0.001$. ★ $\Rightarrow se < 0.0001$

Table 3.1: Results of simulation investigation into $SE(D)$: $\beta = 0.5$, true $D \simeq 0.8$

						$SE(D)_{cox}$				$SE(D)_{boot}$			
β	*cens*	N	$\bar{e}$	D	$SD(D)$	mean	abs. bias		rel. bias	mean	abs. bias		rel. bias
1.0	0	150	150	1.60	0.2101	0.1770	-0.0332	♦	-15.8%	0.1993	-0.0109	•	-5.2%
		300	300	1.60	0.1386	0.1233	-0.0153	♦	-11.0%	0.1391	0.0004	♦	+0.3%
		600	600	1.60	0.0937	0.0866	-0.0071	♦	-7.6%	0.0981	0.0044	♦	+4.7%
		1000	1000	1.59	0.0737	0.0667	-0.0070	♦	-9.5%	0.0755	0.0018	♦	+2.5%
		2000	2000	1.60	0.0567	0.0472	-0.0094	★	-16.6%	0.0535	-0.0032	♦	-5.6%
	40	150	89.1	1.59	0.2438	0.2140	-0.0298	♦	-12.2%	0.2371	-0.0067	•	-2.7%
		300	178.2	1.59	0.1623	0.1496	-0.0127	♦	-7.8%	0.1650	0.0028	♦	+1.7%
		600	356.9	1.61	0.1155	0.1050	-0.0105	♦	-9.1%	0.1161	0.0006	♦	+0.5%
		1000	596.2	1.60	0.0877	0.0809	-0.0068	♦	-7.7%	0.0891	0.0014	♦	+1.7%
		2000	1191.5	1.60	0.0658	0.0570	-0.0088	♦	-13.4%	0.0628	-0.0030	♦	-4.5%
	80	150	29.8	1.62	0.4154	0.3502	-0.0652	•	-15.7%	0.3839	-0.0315	•	-7.6%
		300	59.5	1.61	0.2630	0.2387	-0.0243	♦	-9.2%	0.2550	-0.0081	♦	-3.1%
		600	119.4	1.60	0.1769	0.1662	-0.0107	♦	-6.1%	0.1767	-0.0002	♦	-0.1%
		1000	199.3	1.61	0.1304	0.1281	-0.0023	♦	-1.8%	0.1354	0.0050	♦	+3.8%
		2000	397.7	1.60	0.0960	0.0901	-0.0058	♦	-6.1%	0.0959	-0.0001	♦	-0.1%

Legend for standard error (*se*) of absolute bias: • $\Rightarrow 0.001 \leq se < 0.01$. ♦ $\Rightarrow 0.0001 \leq se < 0.001$. ★ $\Rightarrow se < 0.0001$

Table 3.2: Results of simulation investigation into $SE(D)$: $\beta = 1.0$, true $D \simeq 1.6$

						$SE(D)_{cox}$				$SE(D)_{boot}$			
β	*cens*	N	$\bar{e}$	D	$SD(D)$	mean	abs. bias		rel. bias	mean	abs. bias		rel. bias
2.0	0	150	150	3.16	0.3248	0.2533	-0.0714	⧫	-22.0%	0.3150	-0.0097	•	-3.0%
		300	300	3.17	0.2218	0.1763	-0.0455	⧫	-20.5%	0.2203	-0.0016	⧫	-0.7%
		600	600	3.18	0.1562	0.1238	-0.0324	⧫	-20.8%	0.1550	-0.0011	⧫	-0.7%
		1000	1000	3.18	0.1253	0.0957	-0.0296	⧫	-23.6%	0.1198	-0.0055	⧫	-4.4%
		2000	2000	3.19	0.0816	0.0677	-0.0139	⧫	-17.0%	0.0849	0.0033	⧫	+4.0%
	40	150	89.1	3.16	0.3752	0.2999	-0.0753	•	-20.1%	0.3695	-0.0057	•	-1.5%
		300	178.2	3.18	0.2678	0.2089	-0.0589	⧫	-22.0%	0.2590	-0.0089	•	-3.3%
		600	356.9	3.19	0.1831	0.1471	-0.0360	⧫	-19.7%	0.1818	-0.0013	⧫	-0.7%
		1000	596.2	3.18	0.1377	0.1129	-0.0247	⧫	-18.0%	0.1395	0.0018	⧫	+1.3%
		2000	1191.5	3.19	0.0997	0.0798	-0.0199	⧫	-20.0%	0.0991	-0.0006	⧫	-0.6%
	80	150	29.8	3.24	0.5860	0.4730	-0.1130	•	-19.3%	0.5872	0.0012	•	+0.2%
		300	59.5	3.22	0.3997	0.3203	-0.0794	•	-19.9%	0.3858	-0.0139	•	-3.5%
		600	119.4	3.18	0.2657	0.2209	-0.0448	⧫	-16.9%	0.2652	-0.0006	•	-0.2%
		1000	199.3	3.19	0.2060	0.1711	-0.0349	⧫	-16.9%	0.2044	-0.0016	⧫	-0.8%
		2000	397.7	3.19	0.1559	0.1200	-0.0358	⧫	-23.0%	0.1446	-0.0112	⧫	-7.2%

Legend for standard error (*se*) of absolute bias: • $\Rightarrow 0.001 \leq se < 0.01$. ⧫ $\Rightarrow 0.0001 \leq se < 0.001$. ★ $\Rightarrow se < 0.0001$

Table 3.3: Results of simulation investigation into $SE(D)$: $\beta = 2.0$, true $D \simeq 3.2$

These results are summarized in Figure 3.20, which shows the absolute bias in $SE(D)_{cox}$ and $SE(D)_{boot}$ over the different values of β, censoring and sample size. The increasing bias in $SE(D)_{cox}$ as D increases – and lack of corresponding increase in the bias of $SE(D)_{boot}$ – is clear to see. Figure 3.20 also shows the decrease in bias with increasing sample size in both estimands; although it is not always the case that the largest sample size considered here ($n = 2000$) has the lowest bias. It also shows that $SE(D)_{boot}$ may still be slightly negatively biased for large D.

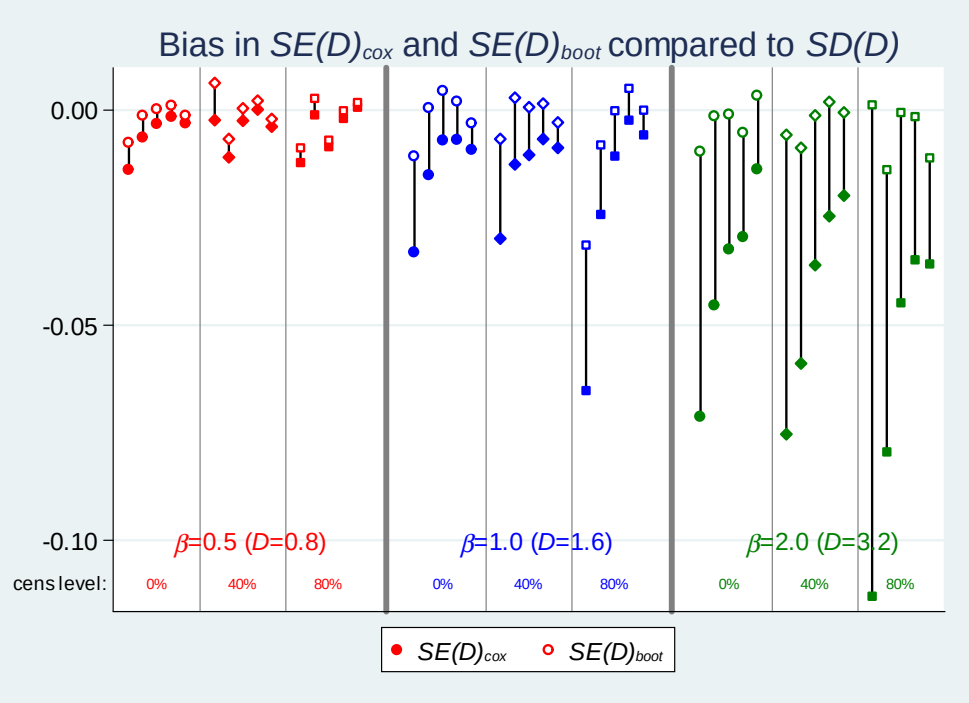

Figure 3.20: Absolute bias in $SE(D)_{cox}$ and $SE(D)_{boot}$. Hollow markers represent bias of $SE(D)_{boot}$ and solid markers $SE(D)_{cox}$

Table 3.4 gives the mean absolute and relative bias for each combination of censoring and β (across all sample sizes). It shows that for low β (and hence D), the difference between $SE(D)_{cox}$ and $SD(D)$ is fairly low. For $D = 0.8$, all mean absolute biases are <0.006. The bias increases as D increases; with $D = 3.2$ the value of $SE(D)_{cox}$ is only 80% of the value of $SD(D)$.

Absolute bias				Relative bias			
	cens				*cens*		
β (True D)	0%	40%	80%	β (True D)	0%	40%	80%
0.5 (0.8)	-0.0058	-0.0039	-0.0047	0.5 (0.8)	0.941	0.962	0.979
1.0 (1.6)	-0.0144	-0.0137	-0.0217	1.0 (1.6)	0.879	0.900	0.922
2.0 (3.2)	-0.0386	-0.0430	-0.0616	2.0 (3.2)	0.792	0.801	0.808

Table 3.4: Mean absolute and relative bias of $SE(D)_{cox}$

Table 3.5 shows the mean absolute and relative bias between $SE(D)_{boot}$ and $SD(D)$ for each combination of censoring and β. The absolute bias is much smaller than seen with $SE(D)_{cox}$ in Table 3.4, and the relative bias is much better too: $SE(D)_{boot}$ is no worse than 97.7% of $SD(D)$ even for $D = 3.2$, compared to 80% for $SE(D)_{cox}$.

Absolute bias				Relative bias			
	cens				*cens*		
β (True D)	0%	40%	80%	β (True D)	0%	40%	80%
0.5 (0.8)	-0.0019	0.0000	-0.0023	0.5 (0.8)	0.985	0.996	0.992
1.0 (1.6)	-0.0015	-0.0010	-0.0070	1.0 (1.6)	0.993	0.993	0.986
2.0 (3.2)	-0.0029	-0.0029	-0.0052	2.0 (3.2)	0.991	0.990	0.977

Table 3.5: Mean absolute and relative bias of $SE(D)_{boot}$

3.3.4 Conclusion

The simulation study reported in this section shows that there is an important difference between the $SE(D)_{cox}$ calculated by the `str2d` command and the 'best' estimate of $SD(D)$ from 500 repetitions. $SE(D)_{cox}$ almost consistently underestimates $SD(D)$ but the magnitude of the underestimation varies: it is smaller with lower D. Bootstrapping $SE(D)$ appears to correct the problem well, with estimates consistently much closer to the $SD(D)$ and not much affected by the magnitude of D. Although here the bootstrapping was done manually, the `str2d` command does include an option to calculate bootstrap standard error.

3.4 Distribution and magnitude of D

Before beginning to develop any formal sample size calculations it is important to know the sampling distribution of D and confirm its magnitude in relation to β. In their (2004) paper on D, Royston and Sauerbrei suggested that with a normally distributed prognostic index (PI) $X \sim N(0,1)$, since $var(\beta X) = \beta^2$, β estimates the standard deviation of the PI values, and thus $D = \beta\kappa = \beta\sqrt{8/\pi}$. We wish to confirm this.

Both these questions (distribution of D and relationship of D to β) can be investigated using simulation data obtained in Section 3.3.

We also conduct a small bootstrap study to investigate the sampling distribution of D in various real datasets. The method for this bootstrap study is:

1. Select a 'best' model from the full dataset using MFP, with model selection α=0.05. Save this model (variables & coefficients).

2. Fit the saved model to the dataset and calculate D.

3. Bootstrap 500 samples from the dataset. For each sample fit the saved model to the bootstrap sample and calculate D for this model in the bootstrap sample.

These steps are repeated for a number of real datasets (all described in Appendix B).

3.4.1 Results

Simulation study

Figure 3.21 shows the distribution of the 2000 values of D found in Section 3.3 for every combination of *cens* and β. In all cases values are approximately normally distributed and appear to peak at $D \simeq \beta\kappa$. The exception to this is for the smallest datasets (N =150 and 300) with 80% censoring, where there is a slight positive skew to the distribution. This is likely due to the very small number of events in these datasets (e =30 and 60) producing overfitting in a few simulated datasets. As would be expected, the spread of D values is wider for lower values of N, and for higher proportions of censoring, as the number of events decreases.

The mean D from Tables 3.1, 3.2 and 3.3 for the smallest ($N = 150$) and largest ($N = 2000$) datasets simulated are given in Table 3.6 along with the bias (compared to $\beta\kappa$)

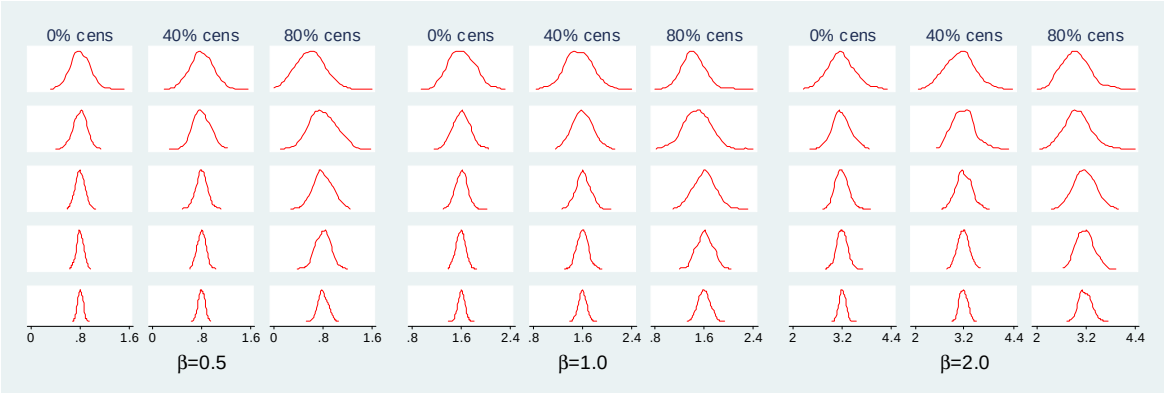

Figure 3.21: Sampling distribution of D from simulated data. From top to bottom, rows correspond to $N = 150, 300, 600, 1000$ and 2000.

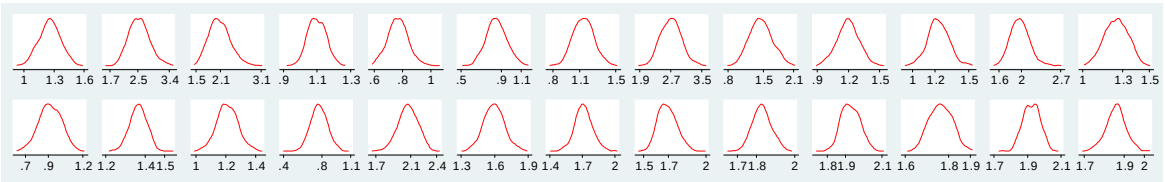

Figure 3.22: Sampling distribution of D for 26 real datasets (500 bootstraps)

and standard error of this bias. This shows that the mean of D over the 2000 simulated datasets does indeed closely approximate $\beta\kappa$ for all combinations of β and censoring and that there is no bias to be concerned about.

Real data study

Figure 3.22 shows the sampling distribution for the real datasets; they are all quite normally distributed.

3.4.2 Conclusion

D appears to be normally distributed in both simulated and real datasets. For our simulations with model PI simulated as $X \sim N(0,1)$, with log hazard ratio β, $D \simeq \beta\kappa$.

The spread of D across identically and independently distributed (iid) simulated datasets, and across bootstraps of the real datasets, is quite high when N is small. We should be aware that by chance a dataset being used for a study could be in the tails of this theoretical distribution and so the estimate of D too high or too low.

β	$\beta\kappa$ (expected D)	censoring	N	D mean	bias	*se* of bias
0.5	0.7978	0%	150	0.8038	0.0059	0.0072
			2000	0.8023	0.0044	0.0019
		40%	150	0.7999	0.0020	0.0149
			2000	0.7990	0.0011	0.0024
		80%	150	0.8375	0.0396	0.0149
			2000	0.8007	0.0028	0.0037
1.0	1.5958	0%	150	1.6049	0.0091	0.0094
			2000	1.6008	0.0050	0.0025
		40%	150	1.5896	-0.0612	0.0109
			2000	1.5950	-0.0007	0.0029
		80%	150	1.6203	0.0246	0.0186
			2000	1.5972	0.0014	0.0043
2.0	3.1915	0%	150	3.1600	-0.0315	0.0145
			2000	3.1932	0.0017	0.0036
		40%	150	3.1632	-0.0283	0.0168
			2000	3.1854	-0.0062	0.0045
		80%	150	3.2411	0.0495	0.0262
			2000	3.1940	0.0025	0.0070

Table 3.6: Mean, bias and *se* of bias of D from simulated datasets with N=150,2000: by β and censoring proportion

3.5 Discussion

The three investigations of D presented in this chapter were designed as ground work to help us develop sample size guidelines based on D.

In Section 3.2 we considered the behaviour of D as sample size changes, both in terms of number of events and EPV. This investigation showed that a dataset with 10 EPV – widely viewed as sufficient for multivariable modelling – is not likely to produce reliable estimates of model prognostic value (in terms of D). We found that such a small dataset may easily result in an estimate of D containing as much as 50% optimism. We conclude that at least 20 or 30 EPV, or at least 200-300 events, is a preferable rough rule in order to minimise optimism and have a better chance of estimating D with reasonable accuracy. However, another conclusion from this investigation is that a blanket rule of thumb based on number of events, or EPV, is too coarse to give a reliable, efficient estimate of D in every situation. A formal calculation will be better able to take into account the nuances of different datasets and so this is what we aim to develop in the next chapters.

Also in this investigation, we considered the optimism present in an estimate of D. We found that Harrell et al.'s (1996) bootstrap method was quite time consuming to implement here as the MFP model must be reselected from scratch in each bootstrap dataset.

As such, for each combination of dataset, n, and p we only selected one subdataset, and only used 100 bootstraps, but the whole investigation still took months rather than weeks or day to complete. Using the technique on a single dataset should be feasible in most cases, however, and could be a useful tool to temper expectations when a high D is reported from an automatically selected model, especially if the dataset is quite small. Royston and Sauerbrei (2004) suggests that this method may overestimate optimism in D , but we have not investigated this concern here.

Section 3.3 showed that the method used by the Stata command `str2d` to estimate the standard error of D consistently underestimates this quantity, and the underestimation is worse for higher values of D. We found that estimating the standard error using a bootstrap gives a much more accurate result with any errors being both positive and negative. As the bootstrap method is straightforward to implement and quick to run on a single dataset, we would recommend it for use when an accurate estimate of $SE(D)$ is required. Since it is likely that any formal sample size calculation will include the standard error of D as a term, it is important that we are able to calculate it accurately.

Finally, in Section 3.4 we confirmed that D has a normal sampling distribution. We also confirmed that if βX is the normally distributed prognostic index, $D \simeq \beta\kappa\sqrt{var(X)}$, so that in our simulation studies where $X \sim N(0,1)$, the D resulting from a given β is roughly 1.6β. The only small deviations from this are likely to be when a dataset contains very few events, or where β and hence D is close to 0, so that the lower bound of D at zero causes skewing. This information is likely to be important for developing formal sample size calculations, which we move on to do in the next few chapters.

Chapter 4

A new structural parameter, λ

4.1 Introduction

In Chapter 3 we considered the behaviour of D with respect to changing sample size. We were not able to find a single rule of thumb for the sample size required to estimate D with reasonable precision, and so concluded that a more customisable solution would be needed. This motivated us to consider more formal sample size calculations which could be applied across a variety of scenarios, and the development and validation of these calculations are described in detail in Chapters 5 and 6.

During the development of these D-based sample size calculations, we increasingly came to see the importance of a structural parameter which we have termed λ. For a survival dataset and model of interest, λ is the product of the number of events in the dataset and the variance of D (for that particular dataset and model). This parameter seemed to be pivotal in our sample size calculations and so to ensure that Chapters 5 and 6 are as clear as possible to the reader, we will first introduce and define λ, look at some of its properties and also investigate its relationship with D.

4.2 Definition of λ

λ is the product of the number of events in the dataset and the variance of D; that is

$$\lambda = e \cdot var(D).$$

As e increases, $var(D)$ decreases; so we may speculate that λ may not have a wide range of values. Indeed, the assumption of a constant λ for a particular dataset and model is key to developing sample size calculations based on D. We outline and test this assumption in Sections 4.3 and 4.4.

The variance of λ can be written

$$\begin{aligned} var(\lambda) &= var(e \cdot var(D)) \\ &= e^2 \cdot var(var(D)) \end{aligned}$$

and thus its standard error as $SE(\lambda) = e \cdot SE(var(D))$. The nested bootstrap required to estimate $SE(var(D))$ with most accuracy is time consuming, so instead

$$SE(\lambda) = e \cdot SE_{boot}(var(D)_{cox})$$

is used in this chapter, where $var(D)_{cox} = (SE(D)_{cox})^2$, and $SE_{boot}(var(D)_{cox})$ is the bootstrap standard error of $var(D)_{cox}$.

4.3 λ proportionality assumption

The proportionality assumption for λ is as follows. For a given model with a certain 'true' value of D, the ratio of the variances σ_1^2, σ_2^2 of D in two datasets with differing numbers e_1, e_2 of events but sampled from the same distribution of covariates equals the reciprocal of the ratio of the corresponding numbers of events:

$$\frac{\sigma_1^2}{\sigma_2^2} = \frac{e_2}{e_1}.$$

This is reasonable, since the variance of a statistic is inversely related to the information in the data, which in a censored time-to-event sample is represented by the number of events (Volinsky and Raftery, 2000). Under this proportionality assumption we can write

$$\lambda = e_1\sigma_1^2 = e_2\sigma_2^2.$$

Note that we don't expect or require λ to be constant across different datasets and models.

We will now test this assumption with empirical work and simulation.

4.4 Testing proportionality assumption for λ: real data

We have stated the assumption that for a particular model used on data with a particular covariate structure, λ is independent of sample size (number of events). In order to test this assumption and also investigate the sampling error of λ we will carry out two investigations. The first essentially calculates λ for subsamples of varying sizes of a real dataset and compares the results; the second (described in Section 4.5) utilises simulated data.

4.4.1 Methods

For investigation of λ in real datasets, the following procedure was used.

1. Select a 'best' model from the full dataset using MFP, with model selection α=0.05. Save this model (variables & coefficients).

2. Choose integers $e_1, e_2, ..., e_j$ such that $0 < e_1 < e_2 < ... < e_j < e$, where e is the number of events in the full dataset. Let N be the total number of patients in the full dataset. It follows that $N - e$ is the number of censored observations in the full dataset. Each integer number of events chosen $e_1, e_2, ..., e_j$ has a corresponding subdataset size $N_1, N_2, ..., N_j$ such that $e_i/N_i = e/N$ for all $i \in 1, 2, ..., j$.

3. From the full dataset, bootstrap a subdataset of size N_i, stratified on event, so that it contains exactly e_i events and exactly $N_i - e_i$ censored observations. This means that the censoring proportion is the same in each subdataset as in the full dataset.

4. Bootstrap 500 samples from the subdataset. For each sample fit the saved model to the bootstrap sample and calculate D for this model in the bootstrap sample. Also, for each sample record $var(D)_{cox}$, the variance of D using the Cox model method in `str2d`.

5. Calculate and record the standard deviation of D from the 500 bootstrap samples; this is $SE(D)_{boot}$.

6. Calculate and record the standard deviation of $var(D)_{cox}$ from the 500 bootstrap samples; this is $SE_{boot}(var(D)_{cox})$ and is used to estimate the standard error of λ, $SE(\lambda) = e \cdot SE_{boot}(var(D)_{cox})$.

7. Calculate $\lambda = e_i \cdot SE(D)_{boot}$.

8. Repeat steps (3) to (7) 500 times for each value of e_i chosen in step 1 and also the full dataset with N patients and e events.

9. Calculate a mean value of λ and $SE(\lambda)$ over the 500 sub-datasets bootstrapped.

This procedure is repeated for six datasets described in Appendix B (FBC, PBC2, LEG, RBC, MYE, SEER_CT).

4.4.2 Results

Table 4.1 gives pertinent results from this investigation; namely mean D, mean $SE(D)_{boot}$, mean λ and mean $SE(\lambda)$ from the 500 bootstrapped subdatasets. It suggests that λ is higher for smaller sub-datasets, but then decreases and stabilises as the sub-datasets get larger.

Figure 4.1 shows the values of λ obtained from the 500 bootstrap samples for each sample size of each of the six datasets. This shows that there can be wide variation in λ between datasets from the same population; and this variation is greater for smaller bootstrap datasets, as would be expected. These graphs also appear to show a positive skew to the distribution of λ for the smaller subdataset sizes; again this would be expected as this quantity is proportional to a variance. Generally once $e > 100$ the distribution appears fairly symmetrical.

Our estimates of $SE(\lambda)$ show that as expected, the sampling error of λ reduces as the dataset size increases. Figure 4.2 shows that to obtain a $SE(\lambda)$ of 10% or less of the magnitude of λ, at least 150 events are needed. For a $SE(\lambda)$ of 5% of the value of λ, at least 500 events are needed, and only very small reductions in $SE(\lambda)$ are obtained by increasing the sample size beyond this.

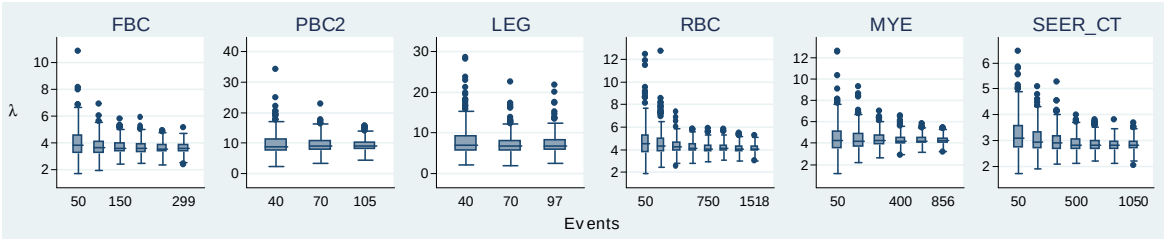

Figure 4.1: Distribution of λ over 500 bootstrapped datasets of each size, for 6 real datasets

Dataset	Patients	Events	EPV	D	$SE(D)_{boot}$	λ	$SE(\lambda)$
FBC	115	50	6.3	1.29	0.278	3.942	0.649
FBC	229	100	12.5	1.25	0.191	3.666	0.408
FBC	344	150	18.8	1.26	0.155	3.631	0.324
FBC	459	200	25.0	1.26	0.134	3.621	0.277
FBC	574	250	31.3	1.26	0.119	3.570	0.244
FBC	686	299	37.4	1.26	0.109	3.578	0.222
LEG	82	40	4.4	2.15	0.431	7.758	1.826
LEG	144	70	7.8	2.08	0.313	7.042	1.164
LEG	200	97	10.8	2.07	0.267	7.113	0.979
MYE	62	50	4.5	0.78	0.293	4.417	0.628
MYE	123	100	9.1	0.78	0.206	4.309	0.412
MYE	247	200	18.2	0.78	0.146	4.311	0.285
MYE	494	400	36.4	0.77	0.103	4.232	0.197
MYE	740	600	54.6	0.77	0.084	4.245	0.161
MYE	1056	856	77.8	0.77	0.070	4.266	0.135
PBC2	82	40	6.7	2.47	0.483	9.593	2.081
PBC2	144	70	11.7	2.55	0.362	9.333	1.393
PBC2	216	105	17.5	2.53	0.292	9.077	1.052
RBC	98	50	5.6	1.14	0.301	4.631	0.674
RBC	196	100	11.1	1.13	0.210	4.453	0.451
RBC	589	300	33.3	1.11	0.119	4.274	0.251
RBC	982	500	55.6	1.11	0.091	4.199	0.192
RBC	1473	750	83.3	1.09	0.074	4.113	0.153
RBC	1964	1000	111.1	1.10	0.064	4.125	0.133
RBC	2456	1250	138.9	1.09	0.057	4.092	0.119
RBC	2982	1518	168.7	1.10	0.052	4.093	0.109
SEER_CT	837	80	5.7	1.85	0.199	3.192	0.429
SEER_CT	1569	150	10.7	1.84	0.141	3.003	0.296
SEER_CT	3138	300	21.4	1.83	0.099	2.953	0.202
SEER_CT	5230	500	35.7	1.83	0.076	2.860	0.153
SEER_CT	7322	700	50.0	1.83	0.064	2.840	0.128
SEER_CT	9414	900	64.3	1.83	0.056	2.810	0.112
SEER_CT	11393	1084	77.4	1.83	0.051	2.820	0.101

Table 4.1: Results of investigation of λ in real datasets

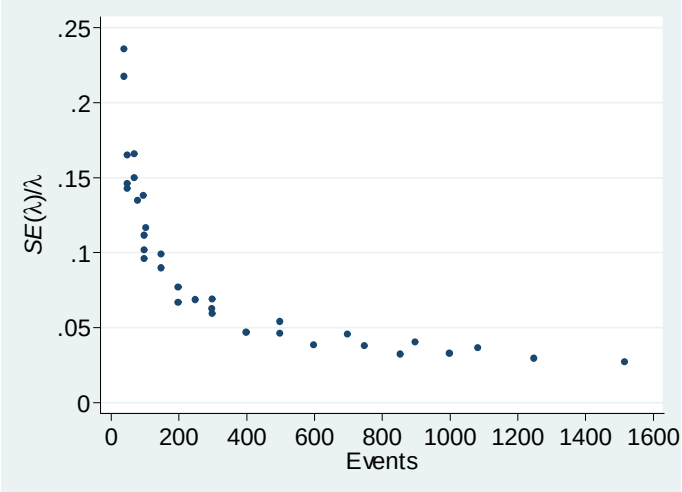

Figure 4.2: $SE(\lambda)$ as proportion of λ vs number of events, for 6 real datasets

4.4.3 Conclusion

This investigation using real data has shown that λ appears to converge to a stable 'true' value as the number of events increases. This suggests that our assumption about the constant nature of λ in a dataset is correct as long as the dataset is above a certain size. Additionally, it has flagged that even in supposedly similar datasets λ can vary quite widely, although this variability reduces as N increases. If more than 150 events are present in the dataset the standard error of λ is likely to be less than 10% of the value of λ; by increasing the sample size to 500 or more events the standard error can be reduced further to 5%. Due to a positive skew it may be slightly more likely that an estimate of λ is too large, rather than too small, if fewer than 100 events are present in the dataset.

4.5 Testing proportionality assumption of λ: simulation

4.5.1 Methods

A simulation study was performed to investigate the proportionality assumption of λ further and hopefully strengthen and generalise the results of Section 4.4. In this study we generate differently sized (but identically distributed) datasets and compare λ for each. Again, $SE(D)_{boot}$ was used to calculate λ. The protocol used was as follows.

1. Generate dataset with N records and approximate censoring proportion *cens* using the method outlined in 3.3.2.
2. Calculate D for the model.
3. Bootstrap 500 samples from the simulated dataset. For each sample, calculate D for the model PI in the bootstrap sample. Also, for each sample record $var(D)_{cox}$, the variance of D using the Cox model method in `str2d`.
4. Calculate and record the standard deviation of D from the 500 bootstrap samples; this is $SE(D)_{boot}$ for the dataset.
5. Calculate and record the standard deviation of $var(D)_{cox}$ from the 500 bootstrap samples; this is $SE_{boot}(var(D)_{cox})$ and is used to estimate the standard error of λ, $SE(\lambda) = e \cdot SE_{boot}(var(D)_{cox})$.
6. Calculate $\lambda = e \cdot (SE(D)_{boot})^2$.
7. Repeat steps (1) to (6) 500 times.
8. Calculate a mean value of λ and $SE(\lambda)$ over the 500 sub-datasets selected.

These steps will be repeated for various values of N (150, 300, 600, 1000, 2000), *cens* (0%, 40%, 80%), and β (0.5, 1.0, 2.0).

4.5.2 Results

Tables 4.2, 4.3 and 4.4 contain the results of this investigation for $\beta = 0.5$, 1.0 and 2.0 respectively. The results agree with the findings in Section 4.4.2 that mean λ is slightly higher for the smallest datasets, but decreases and stabilises as sample size increases.

For the datasets with $\beta = 0.5$ ($D = 0.8$), the magnitude of the difference in λ between the smallest and largest subdatasets is only about 0.3. For the datasets with $\beta = 1.0$ ($D = 1.6$), the difference is slightly greater, at between 0.3-0.8, and for the datasets with $\beta = 2.0$ ($D = 3.2$) it is greater still, at 0.65 when the censoring rate = 0%, but as high as 2.4 when the rate is 80%. For most of the combinations considered, a sample size of 300-500 events seemed adequate to be able to assume a constant value of λ. We also note a decrease in the value of λ with increasing censoring.

Figure 4.3 shows the distribution of values of λ across the 500 identically simulated datasets for each sample size, for $\beta = 1.0$, $\beta = 0.5$ and $\beta = 2.0$. It shows how λ can vary among identically distributed data, and confirms the result in Section 4.4.2 that the distribution of λ appears positively skewed, especially for lower values of n. The different scale of the y axis in the three rows of the graphs should be noted; λ is higher for larger β, so the scales are different.

The estimated standard error of λ was quite high in the smallest datasets but drops quickly as the number of events increases. Similarly to the real datasets, it seems that to obtain a $SE(\lambda)$ of 10% or less of the magnitude of λ, at least 150 events are needed, and for a $SE(\lambda)$ of 5%, 500 events are needed (shown in Figure 4.4).

			Mean over 500 simulations				
β (True D)	Censoring	N	e	D	$SE(D)_{boot}$	λ	$SE(\lambda)$
0.5 (0.8)	0%	150	150	0.80	0.155	3.662	0.335
		300	300	0.80	0.108	3.502	0.232
		600	600	0.81	0.076	3.502	0.166
		1000	1000	0.81	0.059	3.452	0.129
		2000	2000	0.80	0.041	3.433	0.092
	40%	150	90.0	0.81	0.192	3.357	0.421
		300	180.8	0.80	0.133	3.221	0.279
		600	360.8	0.80	0.093	3.144	0.197
		1000	600.2	0.80	0.072	3.158	0.153
		2000	1199.7	0.80	0.051	3.111	0.109
	80%	150	29.6	0.83	0.319	3.021	0.746
		300	59.5	0.80	0.223	2.973	0.447
		600	119.0	0.81	0.155	2.859	0.298
		1000	197.3	0.80	0.119	2.815	0.227
		2000	396.8	0.80	0.084	2.800	0.159

Table 4.2: Results of simulation investigation into λ: $\beta = 0.5$

			Mean over 500 simulations				
β (True D)	Censoring	N	e	D	$SE(D)_{boot}$	λ	$SE(\lambda)$
1.0 (1.6)	0%	150	150	1.61	0.199	6.042	0.565
		300	300	1.60	0.139	5.873	0.385
		600	600	1.60	0.098	5.735	0.262
		1000	1000	1.59	0.076	5.728	0.203
		2000	2000	1.60	0.053	5.707	0.144
	40%	150	88.9	1.60	0.239	5.126	0.655
		300	178.8	1.61	0.166	4.954	0.438
		600	357.5	1.60	0.115	4.783	0.299
		1000	596.7	1.59	0.090	4.796	0.232
		2000	1192.9	1.60	0.063	4.750	0.164
	80%	150	30.0	1.60	0.375	4.252	1.047
		300	59.4	1.62	0.257	3.937	0.626
		600	119.8	1.61	0.179	3.739	0.402
		1000	119.4	1.59	0.136	3.693	0.304
		2000	398.1	1.60	0.095	3.613	0.212

Table 4.3: Results of simulation investigation into λ: $\beta = 1.0$

				Mean over 500 simulations			
β (True D)	Censoring	N	e	D	$SE(D)_{boot}$	λ	$SE(\lambda)$
2.0 (3.2)	0%	150	150	3.16	0.314	14.984	1.524
		300	300	3.17	0.219	14.568	0.998
		600	600	3.18	0.155	14.512	0.688
		1000	1000	3.18	0.120	14.431	0.526
		2000	2000	3.18	0.085	14.327	0.368
	40%	150	89.4	3.18	0.371	12.476	1.757
		300	178.8	3.17	0.260	12.146	1.122
		600	358.4	3.18	0.181	11.853	0.787
		1000	596.3	3.17	0.140	11.659	0.577
		2000	1193.5	3.19	0.099	11.753	0.408
	80%	150	30.2	3.22	0.582	10.424	3.090
		300	60.0	3.23	0.390	9.198	1.577
		600	120.1	3.19	0.264	8.416	0.961
		1000	199.8	3.20	0.205	8.478	0.727
		2000	401.4	3.19	0.144	8.393	0.499

Table 4.4: Results of simulation investigation into λ: $\beta = 2.0$

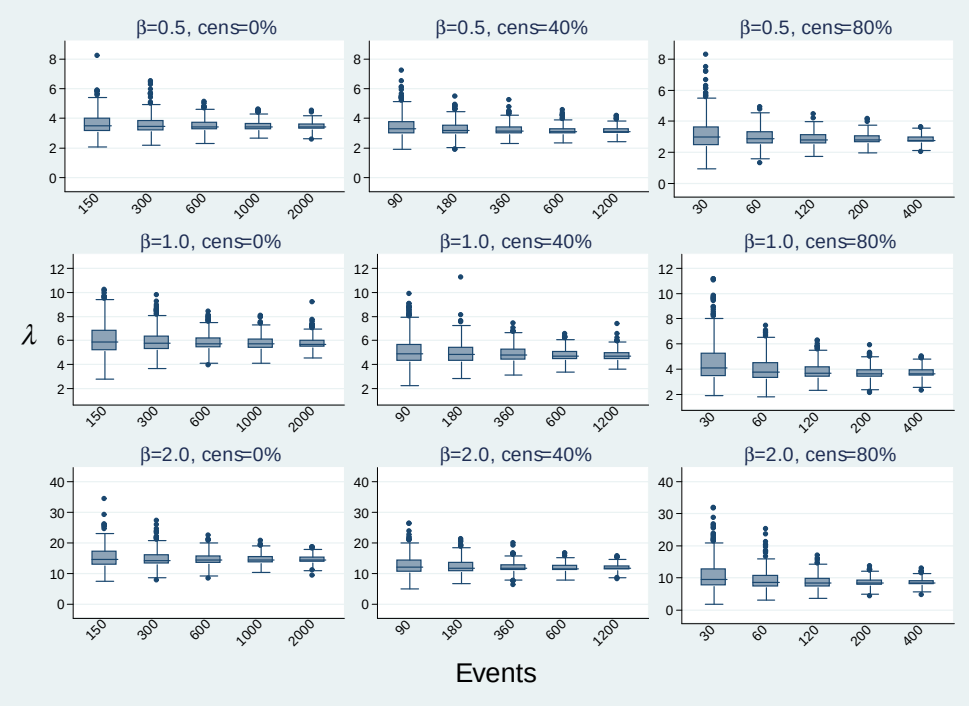

Figure 4.3: Distribution of λ over 500 simulated datasets of each β, size & censoring level

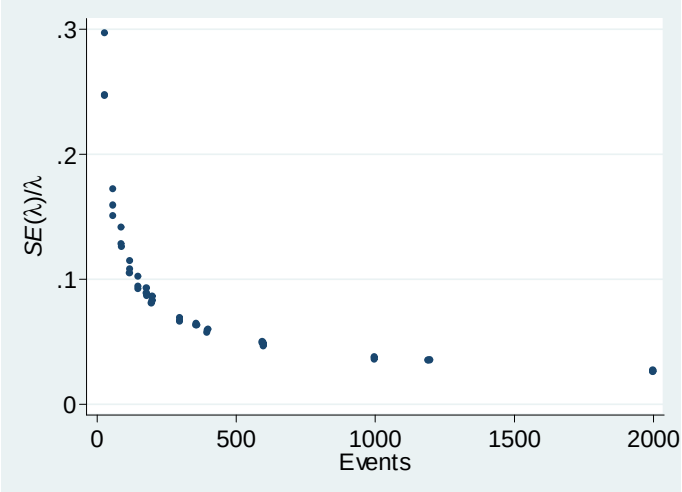

Figure 4.4: $SE(\lambda)$ as proportion of λ vs number of events, for simulated datasets

4.5.3 Conclusion

This more systematic investigation of λ and its relationship with sample size has confirmed the findings of the investigation using real data. λ is higher for smaller datasets and decreases as the number of events increases (with the distribution of data kept the same); at first steeply and then more gently. The lower the D of the dataset and model, the lower the absolute difference between the estimates of λ in the smallest and largest subdatasets. For most situations, 300–500 events seem to be required to obtain a reasonably reliable estimate of λ. Additionally, the value of λ decreases as the proportion of censoring in the dataset increases (everything else being equal). The heterogeneity seen in estimates of λ across identically simulated datasets can be quite large, especially in smaller datasets and this should be borne in mind; two studies of identical (small) size, performed in similar populations, may give quite different estimates of λ. As for the real datasets, a sample size of 150 or more events gives a $SE(\lambda)$ of 10% of the value of λ, and 500 events should give a $SE(\lambda)$ of 5% or less.

4.6 Estimating λ through simulation or bootstrap

The estimates of $SE(\lambda)$ found in the previous two sections allow us to consider how many simulations or bootstraps are required to estimate λ. According to Burton et al. (2006), to estimate a quantity with standard error σ to within a desired accuracy δ at the two-sided α% level,

$$B = \left(\frac{z_{1-\alpha/2}\sigma}{\delta}\right)^2 \tag{4.1}$$

bootstraps are required, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. If we wish to express the level of accuracy as a proportion (p) of the value of the parameter, for λ this translates to

$$B = \left(\frac{z_{1-\alpha/2}SE(\lambda)}{p\lambda}\right)^2.$$

Inputting our estimates of λ from real data in Table 4.1, if we wish our bootstrap estimate to be within 1% of the value of λ (at the two-sided 5% level), then between 30 and 2100 bootstraps would be required for the various combinations of parameters.

Rearranging equation 4.1 tells us that the 500 bootstraps actually used in the study should gave us accuracy of between 0.2% and 2% of the value of λ, which is acceptable.

Assuming that equation 4.1 also holds for simulation studies, for the estimates of λ and $SE(\lambda)$ from Tables 4.2-4.4 we observe that between 25 and 3000 simulations are required for estimates of λ to have 1% accuracy. Again, the 500 simulations used should result in accuracies of between 0.2% and 2.6%, according to equation 4.1.

How many bootstraps for $SE(D)_{boot}$?

We can also use equation 4.1 to calculate a rough rule for the number of bootstraps required to estimate $SE(D)_{boot}$.

We saw in Figures 4.2 and 4.4 that $\frac{SE(\lambda)}{\lambda} \simeq 0.1$ when there are around 150 events in the dataset, and $\frac{SE(\lambda)}{\lambda} \simeq 0.05$ for around 500 events. Since

$$\begin{aligned} \frac{SE(\lambda)}{\lambda} &= \frac{e \cdot SE(var(D))}{e \cdot var(D)} \\ &= \frac{SE(var(D))}{var(D)}, \end{aligned}$$

it holds that that $SE(var(D)) \simeq 0.1 var(D)$ for datasets of $\sim$150 events and $SE(var(D)) \simeq 0.05 var(D)$ for $\sim$500 events. Thus the number of bootstraps required to estimate $var(D)$ to an accuracy of proportion p of its value is

$$\begin{aligned} B &= \left(\frac{z_{1-\alpha/2} SE(var(D)_{boot})}{p \cdot var(D)_{boot}} \right)^2 \\ &= \left(\frac{z_{1-\alpha/2} \cdot 0.1}{p} \right)^2 \end{aligned}$$

for datasets of $\sim$150 events, and for datasets of $\sim$500 events

$$B = \left(\frac{z_{1-\alpha/2} \cdot 0.05}{p} \right)^2.$$

Note that as these equations refer to estimating $var(D)_{boot}$, a desired accuracy level for $SE(D)_{boot}$ must be squared before being inserted into the equation as p.

These equations result in the following. If $\alpha = 0.05$, then for datasets of around $\sim$150 events, 6147 bootstraps are required for 5% accuracy in the estimate of $SE(D)_{boot}$, and 384

bootstraps for 10% accuracy. For datasets of ∼500 events, 1537 bootstraps are required for 5% accuracy and 96 for 10% accuracy.

The 500 bootstraps we have used so far for estimating $SE(D)_{boot}$ give accuracy of 6.6% for a 500 event dataset and 9% for a 150 event dataset. We feel these levels are acceptable. For datasets with less than 100 events, Figures 4.2 and 4.4 suggest that $SE(var(D)) \geq 0.15var(D)$; in this case 500 bootstraps could lead to accuracy of up to $\sim$ 13% of the value of $SE(D)_{boot}$. If this is felt to be too high, 1000 bootstraps will give an accuracy of around 10%, but up to 25000 bootstraps would be required to bring this down to 5%.

4.7 Relationship between λ and D

Since λ is defined by $var(D)$ we expect there to be a relationship between λ and D. As it may be easier to estimate D for a dataset than $var(D)$, such a relationship may give us an easier way to estimate λ. To try and uncover this relationship a simulation study is performed.

4.7.1 Aim

A simulation study is performed to produce data which will be used to investigate and describe the relationship between λ and D. The proposed relationship will then be assessed in simulated datasets with a different censoring pattern and also in real datasets to see how generalisable it is.

4.7.2 Methods

This is similar to the procedure outlined in 4.5.1; but we chose to use $SD(D)$, the empirical standard error of D, in the calculation of λ in order to get more accurate estimates. We also repeated the study 5000 times to try and ensure smooth results which would make any pattern easily visible.

1. Generate dataset with 5000 records and approximate censoring proportion *cens* using the method outlined in 3.3.2.

2. Calculate D for the model.

3. Repeat steps (1) and (2) 5000 times.

4. Calculate the empirical standard error $SD(D)$ from the 5000 estimates of D; calculate $\lambda = e \cdot SD(D)^2$.

These four steps were repeated for censoring rates of 0%, 20%, 40%, 60% and 80%; and closely spaced values of the log hazard ratio β: 0.1, 0.2 and then in steps of 0.2 up to $\beta = 3.4$, which corresponds to $D = 5.4$.

The results of the simulation study will be inspected and we will attempt to describe the relationship numerically using fractional polynomials to regress λ on D.

4.7.3 Results

Figure 4.5 shows graphically the results of this simulation study. There is clearly a strong positive relationship between λ and D, and increasing the proportion of censoring appears to reduce the magnitude of λ.

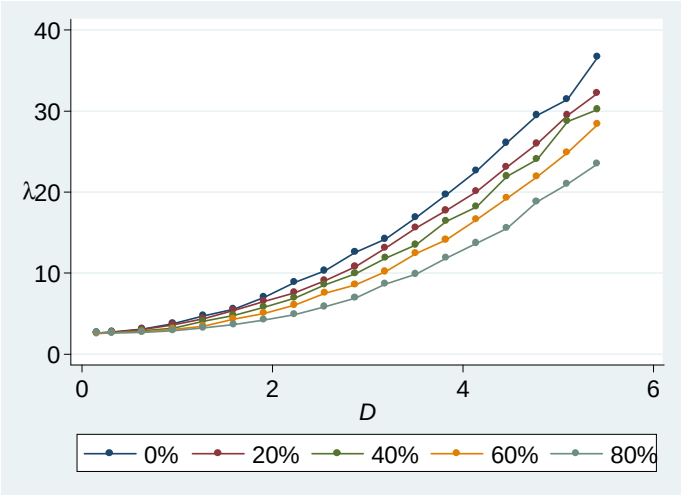

Figure 4.5: Plot of λ vs D from simulation study, for five censoring rates

Regressing λ on D with additional fractional powers over the full range of D tested (0.1-5.4) we found the optimal regression model to be

$$c_0 + c_1 D^{1.9} + c_2 (D \cdot cens)^{1.2},$$

where $c_0 = 2.64$, $c_1 = 1.32$, and $c_2 = -1.98$. This resulted in a $R^2_{adj} = 99.8\%$.

Figure 4.6 shows the accuracy of this model's predictions compared to the results of the simulation study for censoring levels of 0%, 40%, 80%. Although it fits very well over the full range of D, the fit of this model to the lower ranges of D (where real life data is

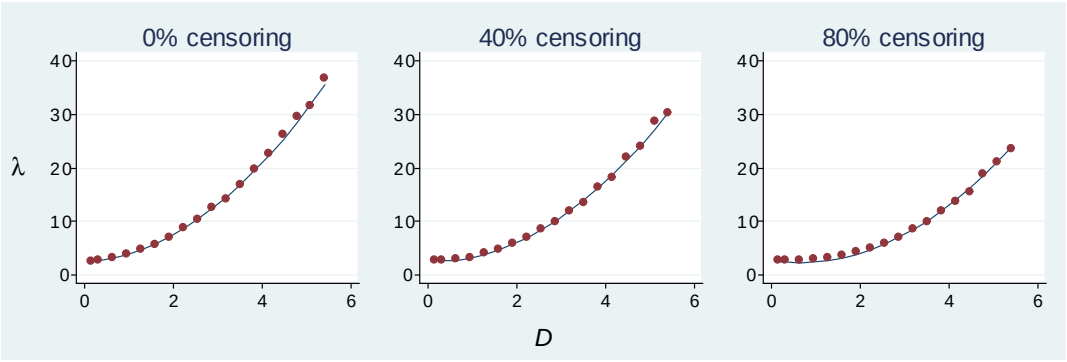

Figure 4.6: Original model: predicted λ vs D overlaid with simulation study results, for 0%, 40% and 80% censoring

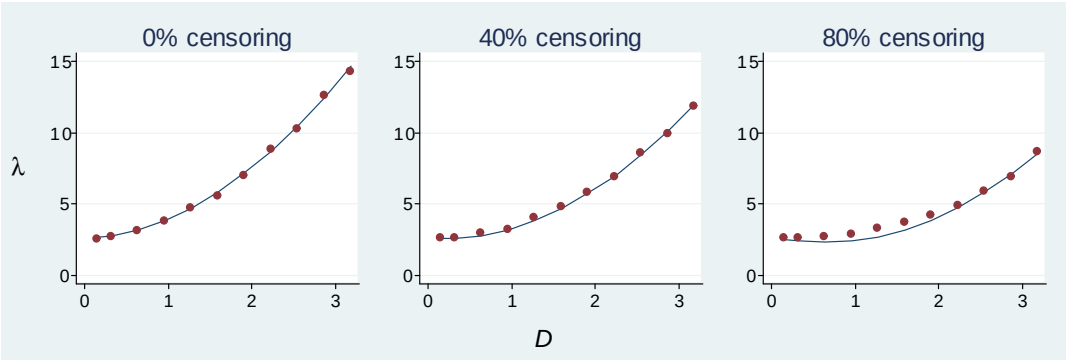

Figure 4.7: Original model: predicted λ vs D overlaid with simulation study results, over range $D \leq 3.2$

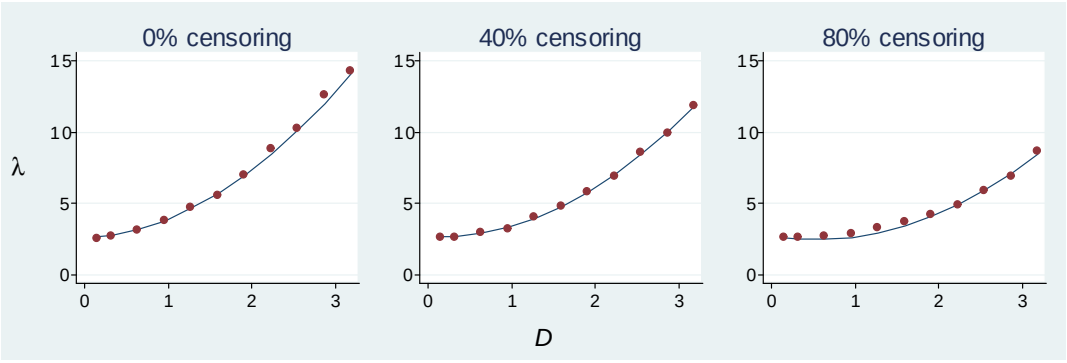

Figure 4.8: Alternative model: predicted λ from equation 4.2 vs D, overlaid with simulation study results, over range $D \leq 3.2$

more likely to lie) is slightly poorer, especially for higher censoring levels. This is shown in Figure 4.7, which shows the plots in Figure 4.6 zoomed in to the range $D \leq 3.2$.

For a better estimate of λ over this more likely range of D, we regressed λ on D over the range $D \leq 3.2$, which corresponds to a log hazard ratio of up to $\beta = 2$. The resulting optimal model was

$$c_0 + c_1 D^{1.9} + c_2 (D \cdot cens)^{1.3}, \tag{4.2}$$

where $c_0 = 2.66$, $c_1 = 1.26$, and $c_2 = -1.65$. This resulted in a better fit over the range $D \leq 3.2$ (as shown in Figure 4.8); $R^2_{adj} = 99.7\%$ for this range. The fit is not quite so good over the full range of D as shown in Figure 4.9; the model appears to underestimate λ for high D, low censoring scenarios; and slightly overestimate λ when there is high D and high censoring. However, the errors in the higher range of D are quite small. In the trade off between a good fit for the lower ranges of D and for the higher ranges of D, we must choose a better fit for the lower ranges as this is where the vast majority of real life datasets will lie.

4.7.4 Performance in simulated data with administrative censoring

To check the generalisability of this model to datasets with different patterns of censoring we generated datasets with administrative censoring only, and compared the observed λ to the value predicted by equation 4.2 above. To simulate datasets of patients with purely administrative censoring we followed this procedure.

1. Simulate N survival times (T_s) by following the procedure in 3.3.2.

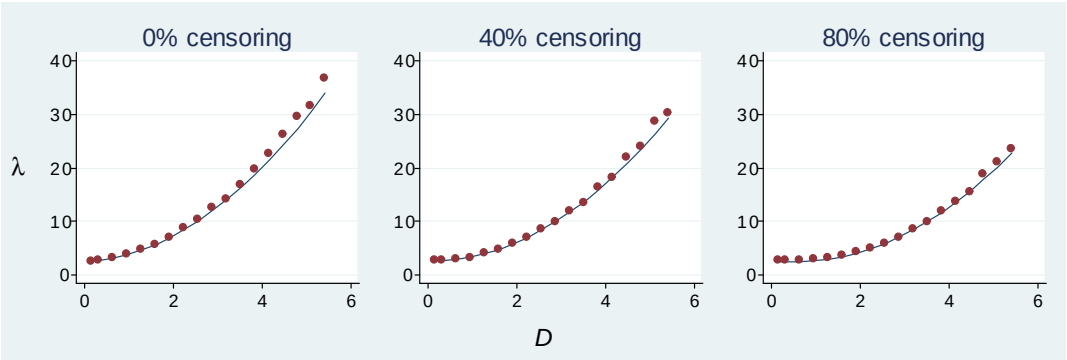

Figure 4.9: Alternative model: predicted λ from equation 4.2 vs D, overlaid with simulation study results

2. Calculate a censoring time T_c for the n^{th} record as follows:

$$T_c = \frac{n}{N}r + f$$

where r was the length of the study recruitment period and f the study follow up period. This assumes that entry of patients to the trial was uniformly staggered over the recruitment period, and that all patients were censored at the end of the follow up period if they had not failed by this time.

3. Records where $T_c < T_s$ were considered censored at time T_c; records where $T_s < T_c$ were considered failures at time T_s. No other censoring was performed.

Here, r was set to 2 years and f to 4 years to reflect a likely study duration (but this particular choice of r and f was essentially arbitrary), and generated datasets of size $N = 2000$. Steps (1) to (3) were repeated 1000 times for each β value of 0.2–2.0 in steps of 0.2, and censoring levels of 40% and 80%. The desired censoring proportion was achieved by changing the baseline hazard of the survival times T_s; the hazard required is dependent on β and was determined through an iterative process.

Figure 4.10 shows the observed λ marked against the line of equation 4.2, for both the full range of D and for $D \leq 3.2$ only. Equation 4.2 fits this data remarkably well, in fact there is little difference in fit between this data and the data the model was developed on. Clinical trials are often used as data sources for prognostic studies, and since censoring in such datasets is mostly administrative, it is reassuring that equation 4.2 works well for simulated data with this pattern of censoring.

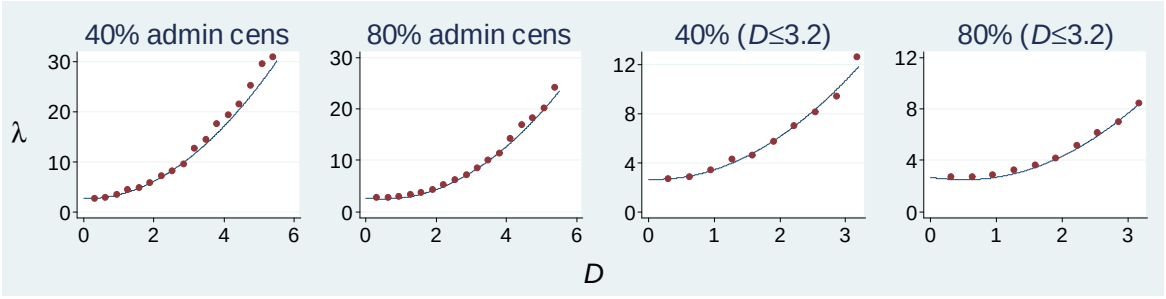

Figure 4.10: Predicted λ from equation 4.2 vs D, overlaid with observed λ from simulation study with administrative censoring

4.7.5 Performance in real data

Having developed and checked this model on simulated data with different censoring patterns, we must assess its performance in real data. To do this we use the 26 datasets already described and select models using MFP with $p = 0.05$ (using the Stata `mfp` command lines given in Appendix B).

Table 4.5 presents the following quantities for each dataset. λ_{true} is the λ calculated in the dataset for the MFP model (using $SE(D)_{boot}$), note that this is not a mean λ over bootstraps, but rather the single value of λ calculated from the original dataset; this is why the values of λ for FBC, PBC2, LEG, RBC, MYE and SEER_CT in Table 4.5 are not the same as those already seen in Table 4.1. $SE(\lambda)$ is also given: for the six datasets used in Section 4.4.2 this quantity is taken from Table 4.1; the same method of estimation described in Section 4.4.1 was also used to estimate $SE(\lambda)$ in the other 20 datasets. λ_{pred} is the λ predicted using equation 4.2, based on the D and *cens* in the dataset.

Table 4.5 appears to show that the estimates of λ predicted by equation 4.2 are reasonable. There is no obvious systemic bias as errors are as often positive as negative. Most predicted values of λ are within 20% of the true value (21 out of 26 datasets), and importantly for this work, most of the predicted λ are within 2 standard errors of the true value (20 out of 26). This means that for most of the datasets in this list, equation 4.2 is likely to give as accurate an estimate of λ as would be obtained from calculating λ from the data itself using $SE(D)$.

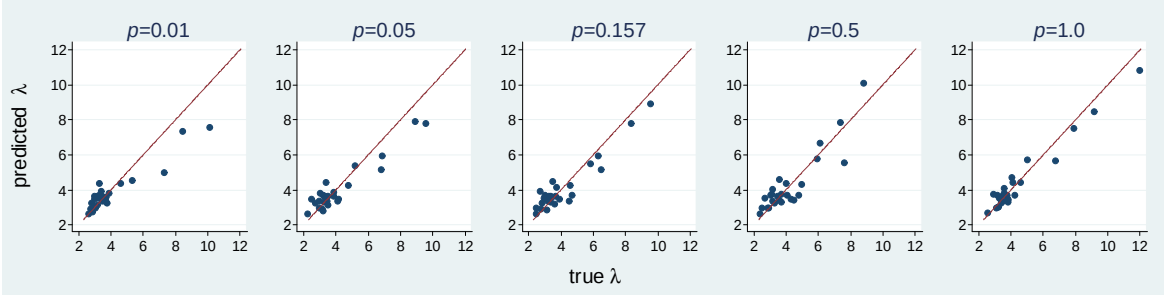

Figure 4.11: λ_{pred} vs λ_{true} for real datasets; models selected by MFP with various p

Dataset	λ_{true} (SE)	λ_{pred}	$\frac{\lambda_{pred}}{\lambda_{true}}$	Dataset	λ_{true} (SE)	λ_{pred}	$\frac{\lambda_{pred}}{\lambda_{true}}$
FBC	3.34 (0.222)	3.56	+7%	SEER_DE	3.54 (0.297)	3.56	+1%
PBC2	9.62 (1.052)	7.75	-19%	APC	3.02 (0.234)	3.32	+10%
LEG	6.92 (0.979)	5.88	-15%	GLI	3.90 (0.313)	3.82	-2%
RBC	4.23 (0.109)	3.42	-19%	PBC	8.99 (1.781)	7.86	-13%
MYE	4.19 (0.135)	3.29	-21%	LVA	6.84 (0.616)	5.07	-26%
SEER_HI	3.07 (2.017)	3.76	+22%	KCA	3.45 (0.303)	4.37	+27%
SEER_UT	3.54 (0.660)	3.05	-14%	FOL	4.80 (0.258)	4.17	-13%
SEER_NM	2.79 (0.661)	3.21	+15%	HOS	5.23 (0.742)	5.34	+2%
SEER_AT	3.30 (0.436)	3.24	-2%	STE	3.23 (0.475)	2.75	-15%
SEER_CT	2.56 (0.101)	3.43	+34%	OVA	3.90 (0.245)	3.59	-8%
SEER_SE	3.25 (0.362)	3.65	+12%	WHI2	3.02 (0.225)	2.88	-5%
SEER_IA	3.53 (0.323)	3.31	-6%	WHI3	3.18 (0.340)	2.86	-10%
SEER_SF	3.35 (0.361)	3.61	+8%	WHI4	2.31 (0.333)	2.56	+11%

Table 4.5: Comparison of true and predicted λ for 26 real datasets

Figure 4.11 shows the results from Table 4.5 in graphical form, as well as equivalent graphs for models selected using $\alpha = 0.01$, 0.157, 0.50 and 1.0. In all five cases the predicted λ is reasonably close to the true λ and no clear patterns are seen across values of p.

It is difficult to know the true reasons for the divergence from our λ model which predicted so well in simulated datasets, however we can speculate on some possible explanations. Any non-normality of the prognostic index (PI) is likely to have an effect, as D is based on the assumption that the model PI is normally distributed.

4.8 Discussion

λ appears to be an important structural parameter which is closely related to D. It is the product of two inversely proportional quantities, the number of events in a dataset and

the variance of D as measured in the dataset. In this chapter we have researched some of the properties of λ.

We have tested the assumption that for a particular model and dataset sampled from the same distribution of covariates, λ is the same regardless of the size of the dataset. This assumption is vital in producing sample size calculations based on D, as we will show in the next chapter. We first tested this in real datasets in Section 4.4 and found that the assumption did hold, as long as the dataset is large enough. When small subsets of the data were used, we found that they tended to show higher values of λ than the full dataset. In Section 4.5 we checked the proportionality assumption of λ again, this time using simulated survival data. Once again we came to the same conclusion: the assumption holds as long as the dataset is large enough.

While testing the proportionality assumption, we also investigated the sampling error of λ, in order to assess whether this quantity can be well estimated. We found that for both real and simulated datasets, randomly selected subdatasets of the same size could show quite variable estimates of λ. This variability was much greater in small samples: datasets with more than 150 events showed a standard error for λ of less than 10% of the value of λ. With more than 500 events, $SE(\lambda)$ was less than 5% of λ.

Finally we used simulated data to develop an equation to predict λ from D and the proportion of censored records in a dataset. This equation predicted λ well in simulated datasets with both random and administrative censoring, and worked fairly well in real datasets too. In particular, the sampling error for λ meant that for most of the real datasets investigated, the prediction equation for λ was likely to give as accurate an estimate of the quantity as would be obtained from calculating λ from the data itself using $SE(D)$.

Chapter 5

Sample size calculation in validation studies using D

5.1 Introduction

In Chapter 2 the importance of performing a sample size calculation prior to starting a prospective study was discussed, as well as the non-availability of such calculations for prognostic studies. We also looked at various measurements of prognostic value in Cox survival models and chose Royston and Sauerbrei's (2004) D measure to form the basis of a novel sample size calculation. In Chapter 3 we considered the behaviour of D with changing sample size and found that there was no common sample size or events per variable level which gave reasonably precise values of D across all the datasets we considered. For this reason we wish to develop a more formal sample size calculation which provides results tailored more to individual study circumstances. To aid development of such a calculation, in Chapter 3 we showed that the sampling distribution of D is approximately normal, both in simulated and real datasets. In Chapter 4 we introduced a parameter, λ, which is the product of the number of events in a dataset and the variance of D, and as such is a relationship of some importance when considering formal sample size calculations.

Having laid this groundwork, in this chapter and the next we move on to formulate possible sample size calculations for two scenarios which may arise in the development or validation of multivariable prognostic models. The first, presented in this chapter, is where the researcher has a model in mind which they wish to validate and access to suit-

able individual patient data from some previous study. Thus estimates of D and $SE(D)$ are available for this model and dataset, and it is desired to validate the estimate of D in new data. The second scenario, covered in Chapter 6, covers several possible research situations but is broadly similar to what usually happens in a clinical trial, where researchers anticipate a particular target value of D and collect data to test this hypothesis.

In this chapter we develop two sample size calculations for the first validation scenario, one based on significance testing and another on confidence interval width. For want of a better expression we term these calculations Sig-1 and CI-1 respectively. We give some examples of their use, and evaluate them using simulated data.

5.2 Sample size calculation based on significance test

We first suggest a method based on significance testing.

First let us introduce some notation. The scenario we consider in this chapter is where estimates of D and $SE(D)$ exist from a previous study using the same model, and researchers wish to validate the estimate of D in a new study. Let D_1 be the value of D in the first study, σ_1^2 the variance of D_1, and e_1 the number of events in the first study. We ask the question as to how to do a sample size calculation for a second, 'validation', study in the same disease. Let D_2 be the D value in the second study, $\sigma_2^2 = var(D_2)$ and e_2 the number of events. Suppose we were willing to tolerate a reduction in D in the second study of δ; so that $D_1 - D_2 \leq \delta$. We want to estimate e_2 so that a true difference of δ is just significant at the one-sided α level with probability (power) $1 - \beta$.

This is a non-inferiority design and δ is the non-inferiority margin, hence the use of one-sided α. Let $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ and $z_{1-\beta} = \Phi^{-1}(1-\beta)$, where $\Phi^{-1}(\bullet)$ is the inverse normal probability function. Then, as in Armitage et al. (2001) (4th printing, p186), we note that positive values of the difference in the estimators $\widetilde{D_1} - \widetilde{D_2} = \widetilde{\delta}$ are significant at the one-sided α level if

$$\widetilde{D_1} - \widetilde{D_2} > z_{1-\alpha}\sqrt{\sigma_1^2 + \sigma_2^2}, \tag{5.1}$$

where Armitage's $\sigma\sqrt{2/n}$ has been substituted with $\sqrt{\sigma_1^2 + \sigma_2^2}$, the standard error of $D_1 - D_2$.

To obtain power greater than $1-\beta$, the right hand side of (5.1) must be less than the point defining a one-sided probability of β when $D_1 - D_2 = \delta$; that is

$$\begin{aligned} z_{1-\alpha}\sqrt{\sigma_1^2+\sigma_2^2} &< \delta - z_{1-\beta}\sqrt{\sigma_1^2+\sigma_2^2} \\ (z_{1-\alpha}+z_{1-\beta})\sqrt{\sigma_1^2+\sigma_2^2} &< \delta \\ \delta &> zz\sqrt{\sigma_1^2+\sigma_2^2}, \end{aligned}$$

where $zz = z_{1-\alpha} + z_{1-\beta}$. Rearranging, we obtain

$$\sigma_2^2 < \left(\frac{\delta}{zz}\right)^2 - \sigma_1^2. \tag{5.2}$$

Before rearranging further to get the number of events e_2, we note that unlike in the classic study design mode, here the value of σ_1^2 is known from the first study. Since σ_2^2 must be positive, the expression places a lower limit on δ:

$$\delta > \sigma_1 zz. \tag{5.3}$$

To convert values in equation 5.2 to numbers of events, we must use the proportionality assumption on λ which was described and tested in Chapter 4. This assumption states that if a given model is fitted to two datasets sampled from the same distribution of covariates but with differing numbers e_1, e_2 of events, with D variances of σ_1^2 and σ_2^2, then

$$e_1\sigma_1^2 = e_2\sigma_2^2 = \lambda.$$

Using this assumption, to calculate e_2 given δ, e_1, σ_1^2 and zz we set σ_2^2 to the limiting value $(\delta/zz)^2 - \sigma_1^2$ in equation 5.2 and write

$$\begin{aligned} e_2 &= e_1\frac{\sigma_1^2}{\sigma_2^2} \\ e_2 &= e_1\frac{\sigma_1^2}{\left(\frac{\delta}{zz}\right)^2 - \sigma_1^2} \\ e_2 &= e_1\left[\left(\frac{\delta}{\sigma_1 zz}\right)^2 - 1\right]^{-1}. \end{aligned} \tag{Sig-1}$$

This final calculation – which we term Sig-1 for reference – is independent of D. Note that calculation Sig-1 can also be written in terms of λ from the previous study, if required:

$$e_2 = \frac{\lambda}{(\delta/zz)^2 - \sigma_1^2}. \tag{5.4}$$

5.2.1 Example

The parameters for this example are taken from the FBC breast cancer dataset (detailed in Appendix A). Suppose that the first study had $e_1 = 299$, and the estimates of D and $SE(D)$ were $D_1 = 1.226$ and $\sigma_1 = 0.105$. According to equation 5.3, the lower limit of δ is 0.307, representing a 25% degradation in D_1. The moral of this is that the difference in D detectable in the validation study with a particular power depends on the size of the original study. If the original study was small, the power to detect a small reduction in D in the validation study will be very low.

In this example, if we desire to detect a difference of $\delta = 0.4$ with significance level $\alpha = 0.05$ and and 90% power ($\beta = 0.1$), then $\left(\frac{\delta}{\sigma_1 zz}\right)^2 = 1.694$ and

$$e_2 > 299 \times [1.694 - 1]^{-1} = 431 \text{ events.}$$

With these assumptions, we would need a validation study with 431 events to have 90% power to detect a degradation of D of 0.4 or more at the one-sided 5% level. Notice that these results do not depend on the value of D in either of the two datasets.

We will use this example to illustrate how sample size varies with δ and e_1.

5.2.2 Effect of parameters on calculation Sig-1

In all sample size calculations, the smaller the effect that it is desired to detect or exclude, the larger the study must be. Thus here, the smaller δ is, the larger the sample size output from calculation Sig-1. In the example in Section 5.2.1, $\lambda = 3.3$. Figure 5.1 shows how the sample size required in this situation varies with δ, if we consider λ and e_1 (and hence σ_1^2) fixed at these values (keeping 90% power and one-sided $\alpha = 0.05$).

The size of the first study will also affect the number of events required in the validation study. As λ is fixed for a particular data structure and model (as shown in Chapter 4), σ_1^2 varies as e_1 changes. Figure 5.2 shows the relationship between the size of the first

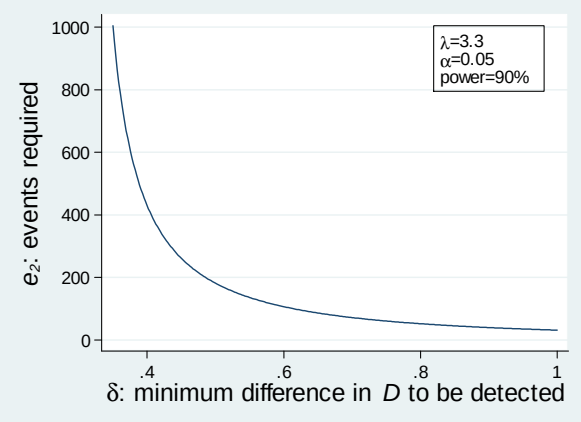

Figure 5.1: Sig-1 example: events required vs δ

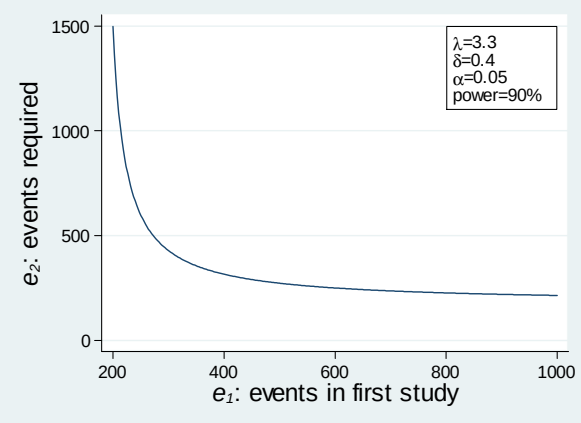

Figure 5.2: Sig-1 example: events required vs size of first study

study and the required size of the validation study, where we fix $\lambda = 3.3$ (as per the example), and $\delta = 0.4$, and keep 90% power and one-sided $\alpha = 0.05$. This suggests that in this situation the original study needs to have at least 400 events if the second study is to be kept to a reasonable size of 300 events or fewer.

Note that the minimum size of the first study is constrained here by the fact that $\delta > \sigma_1 zz$. As we have fixed $\delta = 0.4$ and $\lambda = 3.3$, and $\sigma_1 = \sqrt{\lambda/e_1}$, we have $\delta > zz\sqrt{\lambda/e_1}$ which rearranges to

$$e_1 > \lambda \left(\frac{zz}{\delta}\right)^2$$

which in this example means that $e_1 > 176.5$, thus if the original study has fewer than 177 events, we must choose a larger δ to be detected.

Dataset	Disease	Events	D	$se(D)$	minimum δ	minimum δ as % of D
APC	Prostate cancer	338	0.85	0.095	0.28	33%
FBC	Breast cancer	299	1.26	0.106	0.31	25%
FOL	Follicular lymphoma	573	1.23	0.091	0.27	22%
GLI	Malignant glioma	273	1.15	0.120	0.35	30%
HOS	Cardiovascular disease	215	1.98	0.156	0.46	23%
KCA	Kidney cancer	322	1.20	0.103	0.30	25%
LEG	Leg ulcer	97	2.07	0.267	0.78	38%
LVA	Lung cancer	128	1.43	0.231	0.68	48%
MYE	Myeloma	856	0.77	0.070	0.20	26%
OVA	Ovarian cancer	402	0.91	0.098	0.29	32%
PBC	Primary biliary cirrhosis	125	2.70	0.268	0.78	29%
PBC2	"	105	2.55	0.303	0.89	35%
RBC	Breast cancer	1518	1.09	0.053	0.15	14%
SEER AT	"	731	1.69	0.067	0.20	12%
SEER CT	"	1084	1.83	0.049	0.14	8%
SEER DE	"	1540	1.86	0.048	0.14	8%
SEER HI	"	235	2.07	0.114	0.33	16%
SEER IA	"	1269	1.76	0.053	0.15	9%
SEER NM	"	400	1.71	0.084	0.24	14%
SEER SE	"	1184	1.93	0.052	0.15	8%
SEER SF	"	1270	1.92	0.051	0.15	8%
SEER UT	"	386	1.59	0.096	0.28	18%
STE	Cardiovascular disease	460	1.25	0.084	0.25	20%
WHI2	"	1628	1.36	0.043	0.13	10%
WHI3	"	515	1.21	0.079	0.23	19%
WHI4	"	331	0.77	0.083	0.24	31%

Table 5.1: Minimum δ detectable in validation studies, for various real datasets

5.2.3 Minimum δ for various datasets

As already described, when using calculation Sig-1 the difference that can be detected in a validation study (δ) is constrained by the size of the initial study: $\delta > \sigma_1 zz$, where σ_1 is the standard error of D in the initial study. To see how this translates in real life, Table 5.1 shows the minimum δ that could be detected in validation studies with $\alpha = 0.05$ and power 90%, for the various studies outlined in Chapter 3, using the models found using MFP with $p = 0.05$ described in Appendix B.

Of all these datasets, the smallest minimum δ that can be detected is 0.13; this size δ is only seen in the largest studies with more than 1000 events. The smallest studies with around 100 events have the largest minimum δ in this list, of around 0.7-0.9, surely too large to be of any practical use.

This raises a question: when is an estimate of D so poor – in terms of being based on too small a study – that it isn't worth basing a validation study on it? Firstly, there is little value in recruiting thousands of patients to a study where the non-inferiority margin is so large as to be useless. Secondly, with high $SE(D)$ and a low number of events, there is such a large amount of uncertainty around the estimate of D it may be that researchers are better off not using the sample sizes in this chapter, but instead considering the estimate of D as a target parameter (effectively assuming $SE(D) = 0$), and using the calculations that will be presented in the next chapter.

5.3 Simulation study: significance based calculation Sig-1

To check the validity of the sample size calculation based on significance testing (Sig-1) a simulation study is needed. As we are now concerned with sample size in terms of number of events, it is important that we can specify exactly the number of events in each simulated dataset. The procedure for this is as follows.

5.3.1 Simulating datasets with exact numbers of events

For a dataset with exactly e_1 events and exact censoring proportion $cens$, generate $\frac{e_1}{1-cens}$ records as follows:

1. Generate dataset with $2(\frac{e_1}{1-cens})$ records (twice as many as required) and approximate censoring proportion $cens$ using the method outlined in 3.3.2.

2. Randomly select e_1 records ending in failure, and $\frac{e_1}{1-cens} - e_1$ censored records, to form the final dataset.

5.3.2 Method

The method for testing calculation Sig-1 using a simulation study is as follows.

Note that in this simulation study (and all others in this chapter and Chapter 6), 200 bootstraps are used to obtain $SE(D)_{boot}$, in order to keep the running time of the studies reasonable. According to equation 4.1, the lowest accuracy resulting from 200 bootstraps, given the sizes of datasets used in this chapter and the next, should be around 15% of the value of $SE(D)$; but the majority of scenarios tested will have better accuracy than this

(between 5%–10%). We feel this is a reasonable compromise of estimate accuracy and time required.

1. Generate dataset with $e_1/(1-cens)$ records, exactly e_1 events and exact (random) censoring proportion *cens* using the method outlined in 5.3.1.

2. Compute the D value for this sample; this is D_1. Note that since $X \sim N(0,1)$, $var(\beta X) = \beta^2$, so we have $D = \beta\kappa$, where $\kappa = \sqrt{8/\pi}$ (as shown in Section 3.3).

3. Bootstrap this sample 200 times, calculating D each time. Use these 200 values of D to estimate the standard error of D; this is s_1.

4. Compute e_2 from δ (the maximum difference in D that we will tolerate), e_1, s_1, α and power using equation Sig-1. Note that e_2 will vary across the repetitions of these steps.

5. Simulate a new sample with $e_2/(1-cens)$ records, exactly e_2 events and exact censoring proportion *cens* using the same methods as step 1, under a proportional hazards model with linear predictor $(\beta - \delta/\kappa)X$. This sample is regarded as created under the null hypothesis of inferiority: $H_0 : D \leq \beta\kappa - \delta$ (specifically, under the assumption that $D = \beta\kappa - \delta$).

6. Compute the D value for this sample; this is D_{20}.

7. Bootstrap the sample 200 times, calculating D each time. Use these 200 values of D to estimate the standard error of D; this is s_{20}.

8. Simulate a new sample with $e_2/(1-cens)$ records, exactly e_2 events and exact censoring proportion *cens* under a proportional hazards model with linear predictor βX, using the methods in step 1. This sample is regarded as created under the alternative hypothesis of non-inferiority: $H_1 : D \geq \beta\kappa - \delta$ (specifically, under the assumption that $D = \beta\kappa$).

9. Compute the D value for this sample; this is D_{21}.

10. Bootstrap the sample 200 times, calculating D each time. Use these 200 values of D to estimate the standard error of D; this is s_{21}.

11. Compute test statistics z_0 and z_1 for testing H_0 and H_1 respectively:

$$z_0 = \frac{D_{20} - (D_1 - \delta)}{\sqrt{s_1^2 + s_{20}^2}}$$

$$z_1 = \frac{D_{21} - (D_1 - \delta)}{\sqrt{s_1^2 + s_{21}^2}}.$$

12. Repeat the whole procedure (steps 1-11) 2000 times and store the results.

13. The type I error rate is estimated by the proportion of observations for which $z_0 > \Phi^{-1}(1-\alpha)$; the power is estimated by the proportion of observations for which $z_1 > \Phi^{-1}(1-\alpha)$.

These steps will be repeated for various values of e_1 (750, 1500), power (80%, 90%), *cens* (0%, 50%), β (1, 1.5; hence $D = 1.6, 2.4$) and δ (0.4, 0.5). α=0.05 for all.

5.3.3 Results

Table 5.2 shows the results we would hope for if the sample size calculation Sig-1 was correct. The type I error is close to 5% and the power is close to 80% or 90%, not showing any particular bias above or below the desired value. The standard errors of the estimates of type I error and power are given; these were calculated using the usual binomial variance $np(1-p)$ where $n = 2000$ (the number of simulations) and p is the type I error or power.

How does censoring affect sample size?

Although the censoring proportion in the dataset is not directly included in the calculation of sample size, it does have an indirect effect, as can be seen in Table 5.2. For an initial study with constant D and constant number of events, the number of events required for the validation study decreases as the proportion of censoring in the initial study increases (but number of events stays constant). However, while the number of events decreases, the number of patients required will increase with censoring. This means that if the censoring proportion in the validation study is higher than in the initial study, it is likely that results from the validation study will be less precise than were planned, because not enough patients were recruited.

Simulation parameters					Observed				
e_1	β	Power	δ	cens	e_2	% Type 1 (*se*)		% Power (*se*)	
750	1.0	80%	0.4	0%	320	5.3	(0.51)	80.8	(0.88)
				50%	270	4.3	(0.46)	81.2	(0.87)
			0.5	0%	177	4.9	(0.49)	80.3	(0.89)
				50%	153	4.5	(0.47)	81.4	(0.87)
		90%	0.4	0%	535	5.5	(0.52)	91.4	(0.63)
				50%	435	4.7	(0.48)	91.6	(0.62)
			0.5	0%	272	5.1	(0.50)	90.4	(0.66)
				50%	230	4.2	(0.46)	91.1	(0.64)
	1.5	80%	0.4	0%	730	5.2	(0.51)	79.7	(0.90)
				50%	525	4.4	(0.47)	80.5	(0.89)
			0.5	0%	340	5.6	(0.53)	79.8	(0.90)
				50%	264	4.7	(0.48)	83.0	(0.84)
		90%	0.4	0%	1929	5.4	(0.52)	90.5	(0.66)
				50%	1043	5.0	(0.50)	91.9	(0.61)
			0.5	0%	583	4.4	(0.47)	89.1	(0.70)
				50%	432	4.3	(0.46)	90.6	(0.65)
1500	1.0	80%	0.4	0%	261	4.4	(0.47)	82.0	(0.86)
				50%	227	4.9	(0.49)	82.3	(0.85)
			0.5	0%	158	4.8	(0.49)	81.6	(0.87)
				50%	139	4.7	(0.48)	83.8	(0.82)
		90%	0.4	0%	388	4.6	(0.48)	91.0	(0.64)
				50%	334	4.1	(0.45)	91.8	(0.61)
			0.5	0%	227	5.1	(0.50)	90.6	(0.65)
				50%	198	4.5	(0.47)	92.6	(0.59)
	1.5	80%	0.4	0%	480	4.7	(0.48)	80.2	(0.89)
				50%	386	4.5	(0.47)	82.0	(0.86)
			0.5	0%	274	5.1	(0.50)	79.8	(0.90)
				50%	227	4.1	(0.45)	82.1	(0.86)
		90%	0.4	0%	766	4.9	(0.49)	90.3	(0.66)
				50%	590	4.2	(0.46)	91.3	(0.63)
			0.5	0%	410	4.7	(0.48)	90.1	(0.67)
				50%	330	4.3	(0.46)	92.1	(0.60)

Table 5.2: Simulation study results for significance-based calculation Sig-1

How does the value of D affect sample size?

The magnitude of D seen in the original study has an indirect effect on sample size, as can be seen in Table 5.2. Higher values of D require larger sample sizes. Thus, if the value of D in the validation study is different to that seen in the original study, then the precision of results from the validation study will not be as desired. If the value of D is lower in the validation study, then the study will have greater precision to estimate D. If the value of D is higher in the validation study, then the study will have less precision. However, in the latter scenario a higher value of D may mean that slightly lower precision in its estimate is acceptable, so all may not be lost. This scenario is discussed in more detail in Chapter 7.

5.3.4 Conclusion

The significance-based sample size calculation developed appears to work well. The calculation, Sig-1, shows no systematic errors in power or α as a result of censoring, but the estimates of D from the validation study may be less precise if the censoring proportion or magnitude of D is higher than it was in the initial study.

When considering this significance based calculation it is important for researchers to be aware of the constraint it places on the minimum detectable difference in the validation study. If the initial study is small, then the minimum detectable difference will be quite large, and may be too large to be of use in planning future studies. In this situation one of the other calculations proposed in this and the next chapter may be more appropriate.

5.4 Sample size calculation based on CI for D

In the previous section we considered a traditional significance based calculation. We may alternatively wish to base our sample size calculation on the desired precision of the estimate of D, in the form of the width of its confidence interval (CI).

To this end we extend the work in Section 5.2 to obtain a formula for the CI of the D estimated in the second study. D is normally distributed (as shown in Section 3.4), thus the $(1-\alpha)\times 100\%$ two-sided CI for D is

$$\widetilde{D}_2 \pm z_{1-\alpha/2}\sqrt{var(D)}, \tag{5.5}$$

where $\widetilde{D}_2$ is the estimate of D from the new study. As we have defined $\lambda = e_1\sigma_1^2 = e_2\sigma_2^2$, we can obtain λ from the first study and write $var(D)$ in the validation study as $\sigma_2^2 = \lambda/e_2$. Thus 5.5 becomes

$$\widetilde{D}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\lambda}{e_2}}. \tag{5.6}$$

We can finally substitute $\lambda = e_1\sigma_1^2$ back in and rearrange equation 5.6 to give the required number of events e_2 for our validation study. If we wish the $(1-\alpha)\times 100\%$ CI for the resulting estimate of D not to be larger than $\pm w$ we must rearrange

$$w = z_{1-\alpha/2}\sqrt{\frac{e_1\sigma_1^2}{e_2}}$$

to be in terms of e_2, which gives us the final calculation:

$$e_2 = e_1\sigma_1^2\left(\frac{z_{1-\alpha/2}}{w}\right)^2. \tag{CI-1}$$

Note that this equation is not dependent on the magnitude of D, and the only restriction on the value of w – the half-width of the CI – is that it must be greater than zero.

We can also write this equation in terms of λ:

$$e_2 = \lambda\left(\frac{z_{1-\alpha/2}}{w}\right)^2.$$

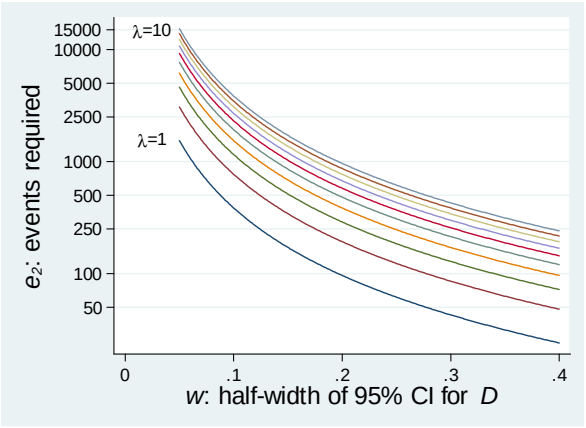

Figure 5.3: CI-1 example: events required vs w: various λ from 1 to 10. Note log scale for events.

5.4.1 Example

Using the same example as in Section 5.2.1, assume we have an original study with 299 events, $D_1 = 1.226$ and $\sigma_1 = 0.105$. If we wish our estimate of D in the second study, $\widetilde{D}_2$, to have a 95% confidence interval of total width 0.25 (so $w = 0.125$), we require

$$e_2 = 299 \times 0.105^2 \times \left(\frac{z_{0.975}}{0.125}\right)^2 = 811 \text{ events.}$$

If we want a much narrower CI of total width 0.1, the sample size increases to 5066.

Once again we will use the parameters of this example to illustrate some of the relationships between sample size and the various parameters in the calculation.

5.4.2 Effect of parameters on calculation CI-1

Again, the more precise we wish our estimate of D to be (in terms of w), the larger the study must be. Figure 5.3 shows the variation in e_2 versus w, for $\lambda = e_1\sigma_1^2$ of 1, 2, ... 10, and illustrates that for many situations a half-width of 0.1 requires more than 1000 events; sometimes many more than 1000. For the example in Section 5.4.1 with $\lambda = 3.3$, if we want a study with fewer than 1000 patients, we must accept a CI of half width > 0.10.

Unlike calculation Sig-1, the sample size of the first study (e_1) and the estimate of $SE(D)$ from the first study (σ_1) only enter calculation CI-1 through the parameter $\lambda = e_1\sigma_1^2$. As a result of this, if λ is assumed to be fixed for a particular data structure and model (as per the proportionality assumption described and tested in Chapter 4), then

the sample size of the validation study is independent of the size of the previous study. This is in contrast to calculation Sig-1 in which σ_1 appears in the calculation additionally to λ (see expression 5.4) and so the size of the first study does have an indirect effect on the calculation through its effect on σ_1, even when λ is considered fixed.

5.5 Simulation study: CI based calculation CI-1

In order to test our proposed sample size calculation CI-1, we must perform a simulation study to test whether the sample size output from the calculation gives the desired confidence interval width.

5.5.1 Method

1. Generate dataset with $e_1/(1-cens)$ records, exactly e_1 events and exact censoring proportion *cens* using the method outlined in 5.3.1.

2. Compute the D value for this sample; this is D_1. Estimate its standard error using a bootstrap with 200 replications; this is s_1.

3. Compute e_2 from e_1, s_1, α and the desired half-width of CI w using equation CI-1.

4. Simulate a new sample with $e_2/(1-cens)$ records, exactly e_2 events and exact censoring proportion *cens* records using the same methods as step 1.

5. Calculate D in this dataset and record it; this is $\widetilde{D}$.

6. Repeat steps (2) and (3) 2000 times. Note that while e_1 is fixed, e_2 varies over the 2000 repetitions since D_1 and s_1 will vary according to the dataset generated in step 1.

7. The proportion of repetitions for which $\widetilde{D} \in (\beta\kappa - w, \beta\kappa + w)$ gives the % CI which has half-width w in the simulated dataset. This should approximate $1-\alpha$, if the sample size calculation is correct.

These steps will be repeated for various values of e_1 (750, 1500), *cens* (0%, 40%, 80%), β (1, 1.5; thus D=1.6, 2.4), w (0.05, 0.1, 0.2). $\alpha = 0.05$ for all.

Simulation Parameters				Observed (95% CI)		
					% of $\widetilde{D}$ (*se*)	
e_1	β	w	*cens*	mean e_2	within $\beta\kappa \pm w$	
750	1.0	0.10	0%	2193	95.0	(0.49)
			50%	1731	95.0	(0.49)
		0.20	0%	550	94.3	(0.52)
			50%	435	95.7	(0.46)
		0.30	0%	247	94.8	(0.50)
			50%	194	96.0	(0.44)
	1.5	0.10	0%	3607	93.9	(0.54)
			50%	2742	95.1	(0.49)
		0.20	0%	904	95.0	(0.49)
			50%	682	95.0	(0.49)
		0.30	0%	401	94.1	(0.53)
			50%	307	95.5	(0.47)
1500	1.0	0.10	0%	2194	94.8	(0.50)
			50%	1731	94.8	(0.50)
		0.20	0%	552	95.3	(0.48)
			50%	434	94.5	(0.51)
		0.30	0%	245	94.8	(0.50)
			50%	192	95.0	(0.49)
	1.5	0.10	0%	3608	94.7	(0.50)
			50%	2730	95.0	(0.49)
		0.20	0%	894	95.5	(0.47)
			50%	682	95.0	(0.49)
		0.30	0%	399	94.3	(0.52)
			50%	305	95.1	(0.48)

Table 5.3: Simulation study results for CI based sample size calculation CI-1

5.5.2 Results

The results are given in Table 5.3. The observed percentage proportion of $\widetilde{D} \in (\beta\kappa - w, \beta\kappa + w)$ is very close to 95% which suggests that the calculation works well. Table 5.3 also shows that the required sample size in many cases is rather large, and that (as expected) there is little difference in e_2 as a result of doubling the sample size of the initial study (here, from 750 to 1500).

From these results we can gain an idea of how two parameters which are not directly included in the sample size calculation – D and censoring rate – actually affect sample size.

5.5.3 How does censoring affect sample size?

Similarly to Sig-1, censoring has an indirect effect on sample size in the CI-based calculation, as can be seen in Table 5.3. A higher censoring proportion in the initial study means

slightly fewer events are required in the validation study, however more patients will be needed. Again, if the censoring proportion in the initial study is lower than in the validation study, it is likely that confidence intervals will have lower than expected coverage, because not enough patients were recruited.

We did a further simulation study to try and quantify the effect of censoring misspecification on sample size and power.

Method

The methods used were broadly the same as outlined in section 5.5. Briefly, an 'initial study' with 50% censoring was simulated using these methods; e_2 was calculated, and then five validation studies with $e_2/0.50$ patients were simulated (separately), with approximately 20%, 30%, 40%, 60%, 70% and 80% censoring. Thus when generating the simulated data for the second study, the number of patients was based on 50% censoring but the distribution of censoring times was based on a different censoring proportion. This was repeated 2000 times for each combination of parameters, for both $e_1 = 1500$ and $e_1 = 750$.

Results

The results for $e_1 = 1500$ and $e_1 = 750$ were very similar, so only the former are presented, in Table 5.4. As well as the proportion of $\widetilde{D}$ within the desired width of CI, we also present the half-width of 95% CI actually observed in the resulting validation dataset; since even if coverage is too low, the observed 95% CI may not be much wider than was desired.

As expected, a higher proportion of censoring in the validation study led to a lower than expected proportion of $\widetilde{D}$ being within the prescribed limits $(\beta\kappa - w, \beta\kappa + w)$ and hence wider 95% CIs. With 70% censoring instead of 50%, the 95% CI width was around 0.24 instead of the planned 0.20, and 0.48 instead of planned 0.40. For 80% censoring instead of 50%, observed widths were 0.28 instead of 0.20 and 0.58 instead of 0.40.

Conversely lower censoring proportions led to $(\beta\kappa - w, \beta\kappa + w)$ containing more than 95% of output values, and so slightly narrower 95% CIs; for 20% or 30% censoring instead of 50%, widths were around 0.18 instead of 0.20, and 0.36–0.37 instead of 0.40.

There was little difference between the desired width and the widths observed when the validation dataset had 40% or 60% censoring instead of 50%.

Simulation Parameters						Observed (95% CI)			
	cens			*cens*	planned		% of $\widetilde{D}$ (*se*)		half-width
e_1	init. study	β	w	val. study	mean e_2	mean e_2	within $\beta\kappa \pm w$		of 95% CI
1500	50%	1.0	0.10	20%	1725	2760	98.0	(0.32)	0.083
				30%	1725	2415	96.7	(0.40)	0.090
				40%	1725	2070	96.3	(0.42)	0.093
				60%	1725	1380	92.6	(0.59)	0.110
				70%	1725	1035	88.3	(0.72)	0.123
				80%	1725	690	83.5	(0.83)	0.144
			0.20	20%	434	694	98.2	(0.30)	0.165
				30%	434	608	96.9	(0.39)	0.181
				40%	434	521	96.1	(0.44)	0.187
				60%	434	347	92.4	(0.59)	0.220
				70%	434	260	90.1	(0.67)	0.241
				80%	434	174	84.4	(0.81)	0.288
		1.5	0.10	20%	2724	4358	97.9	(0.32)	0.087
				30%	2724	3813	97.1	(0.38)	0.090
				40%	2724	3269	96.4	(0.42)	0.094
				60%	2724	2179	91.9	(0.61)	0.113
				70%	2724	1634	90.3	(0.66)	0.119
				80%	2724	1090	83.4	(0.83)	0.140
			0.20	20%	683	1093	97.7	(0.34)	0.174
				30%	683	956	96.9	(0.39)	0.184
				40%	683	820	96.3	(0.42)	0.187
				60%	683	546	93.2	(0.56)	0.213
				70%	683	410	89.8	(0.68)	0.241
				80%	683	273	82.9	(0.84)	0.290

Table 5.4: Simulation study results for CI based sample size calculation CI-1: misspecification of censoring rate, $e_1 = 1500$

5.5.4 How does D affect sample size?

Although D is not a parameter in the sample size calculation CI-1, it has an effect on sample size through λ. Table 5.3 shows that the higher D is, the larger the sample size (number of events) required. We ran another simulation study to try and quantify the effect of misspecifying D – for example in the situation where D in the validation study turns out to be higher or lower than the D in the initial study.

Method

An 'initial study' with $D = 1.6$ ($\beta = 1.0$) was simulated using the methods in Section 5.5, and then two validation studies were simulated; one with $D = 0.8$ ($\beta = 0.5$) and one with

$D = 2.4$ ($\beta = 1.5$). Thus when generating the simulated data for the second study, the number of events was based on $D = 1.6$, but the distribution of survival and censoring times were based on either $D = 0.8$ or $D = 2.4$. This was repeated 2000 times for each combination of parameters.

Results

The results are given in Table 5.5. As expected, observing a higher D than was expected in the validation study leads to a lower proportion than 95% of $\widetilde{D}$ being in the interval $(\beta\kappa - w, \beta\kappa + w)$. However, with the parameters used here – going from $D = 1.6$ to $D = 0.8$ – the resulting 95% CI is not much wider than was planned; generally a total width of 0.21–0.22 rather than 0.20, and 0.42–0.44 rather than 0.40. Going from $D = 1.6$ to $D = 2.4$ leads to wider intervals (0.25-0.28 instead of 0.20, 0.50-0.54 instead of 0.40). However, once again it should be noted that a higher than expected D in the validation study may mean that a wider CI is acceptable than was originally desired.

If the D in the validation study was lower than expected, the validation study will have narrower CIs than planned. Going from $D = 1.6$ to $D = 0.8$ results in intervals of 0.13 instead of 0.20 and 0.26–0.28 instead of 0.40.

5.5.5 Conclusion

The sample size calculation CI-1 based on CI width works well, giving the desired confidence interval width over a variety of parameter values. The sample size for obtaining a confidence interval of half width less than 0.1 is generally very large, increasing for higher values of D. A point of note for this calculation based on precision of estimates is that under the assumption that λ is fixed for a particular covariate distribution and model, the sample size required for the validation study is independent of the size of the initial study.

The calculation works well for censored datasets under the assumption that the censoring rate in the validation study will be the same as the rate seen in the original study. If the censoring rate is higher in the validation study, then the desired CI for D will be wider than expected; a lower censoring rate means CIs will be narrower than expected. Misspecification of D also has an effect; a higher D in the validation study than in the initial study means the validation study will have slightly wider than expected CIs, but

the difference is quite small. A lower D means that the CIs will be slightly narrower than expected.

Simulation Parameters					Observed (95% CI)			
	β			β	mean	% of $\tilde{D}$ (se)		half-width
e_1	init. study	w	*cens*	val. study	e_2	within $\beta\kappa \pm w$		of 95% CI
750	1.0	0.10	0%	0.5	2211	98.6	(0.26)	0.064
				1.5	2211	86.2	(0.77)	0.113
				2.0	2211	78.4	(0.92)	0.131
			50%	0.5	1739	98.1	(0.31)	0.067
				1.5	1739	90.0	(0.67)	0.101
				2.0	1739	79.4	(0.91)	0.127
		0.20	0%	0.5	549	98.6	(0.26)	0.128
				1.5	549	86.2	(0.77)	0.222
				2.0	549	78.7	(0.92)	0.266
			50%	0.5	434	97.8	(0.33)	0.139
				1.5	434	89.4	(0.69)	0.203
				2.0	434	80.2	(0.89)	0.258
1500	1.0	0.10	0%	0.5	2205	98.8	(0.25)	0.065
				1.5	2205	87.5	(0.74)	0.109
				2.0	2205	78.3	(0.92)	0.129
			50%	0.5	1732	98.3	(0.29)	0.069
				1.5	1732	89.7	(0.68)	0.101
				2.0	1732	79.3	(0.91)	0.130
		0.20	0%	0.5	550	99.1	(0.22)	0.132
				1.5	550	87.5	(0.74)	0.217
				2.0	550	77.1	(0.94)	0.276
			50%	0.5	432	97.9	(0.32)	0.138
				1.5	432	89.2	(0.69)	0.207
				2.0	432	80.4	(0.89)	0.259

Table 5.5: Results of simulation study of D for CI based sample size calculation CI-1: misspecification of D

5.6 Discussion

We have considered the issue of required study sample size in the situation where a previous study is used to estimate the required parameters such as D and $SE(D)$. For this situation we have developed and explored two sample size calculations; one based on significance testing and one on the desired precision of the resulting estimate of D (in terms of confidence interval width).

The calculation based on significance testing (Sig-1) appears to work well, giving the correct type I error and power, for both uncensored and censored data (under the assumption that the censoring rate in the validation study will be the same as in the initial

study). One consequence of using the significance based calculation is that if the initial study was small, the minimum detectable difference may be too large to be of practical use.

The calculation based on precision of estimates (CI-1) also works well, and there is no lower limit on the desired width of confidence interval; although the sample size may be prohibitively high if a narrow interval is desired. This calculation is independent of the size of the initial study. It shows no apparent bias with censored data (again, under the assumption that the previous and validation study have the same rate of censoring), although it should be remembered that the calculation gives the number of events required, rather than number of patients.

For these two calculations, if D or the censoring rate seen in the validation study are markedly different to their values in the initial study, this will affect the precision of the estimate of D from the validation study. In particular, an increase in D or an increase in censoring rates will see lower observed precision in the validation study than was planned, while a decrease in D or censoring rates will result in higher precision. The magnitude of the effects are difficult to quantify as these parameters are not included in the calculation but our simulation studies seem to suggest that no concern arises if the difference in censoring is 10% or less (in absolute terms).

The consequences of overestimation of D in the initial study should not be ignored; due the the large numbers of patients required in general for these studies, overestimating D could increase sample size by hundreds or thousands of patients. In Chapter 3, we found that the optimism present in D can be quite large when study size is small and the model of interest was selected using an automatic selection procedure. If this is the case, it is likely that the estimate of D from a small initial study is too high, which could lead to an inflated sample size for a validation study. This makes it important in this situation to determine the optimism present in D, in order that a realistic value of D can be obtained for use in the sample size calculation.

If researchers suspect that the value of D or censoring proportion may be markedly different in a future study to what they were in the initial study, it may be prudent to perform simulation studies to obtain a range of likely sample sizes.

Chapter 6

Sample size calculations using a point estimate of D

6.1 Introduction

In the previous chapter we considered sample size calculations for the scenario where the researcher has access to suitable individual patient data from some previous study and has already formulated a model for the situation at hand. Thus an estimate of D (and $SE(D)$) is available and the model known, and it is desired to validate this estimate for the model in a new dataset.

The second possible scenario, which we cover in this chapter, is where researchers have some estimate of D and wish to test it in a new dataset. This is akin to the situation we usually see in a clinical trial: a difference in effect size between two groups is postulated and a study designed to test this. The main difference between this scenario and the scenario in Chapter 5 in terms of derivation of sample size calculations is that an estimate of $SE(D)$ is not used here, we just have a target value of D.

This target value will be based on some amount of evidence; on one hand the number may be just plucked from the range of likely values of D, on the other hand it might be based on a few published articles or a D reported in a similar disease or group of patients. Again, this is similar to how a target parameter is arrived at for a clinical trial; investigators may have nothing more than a clinically significant difference to guide them, or there may be previous similar studies that give them an idea of the effect size they are likely to see.

This scenario covers a more diverse range of possible situations than the scenario presented in Chapter 5. It may be that researchers wish to externally validate a published model for which the authors reported a value of D, but no $SE(D)$, hence the calculations in Chapter 5 cannot be used. Alternatively – and probably more commonly – they may wish to develop a new model, and have little idea of what the value of D might be, let alone $SE(D)$. Another possible situation is where estimates of D and $SE(D)$ are available, but a significance based calculation is desired and the minimum difference in D to be detected (δ) using calculation Sig-1 is too large to be of any use.

As in the previous chapter, for this scenario we develop and test two sample size calculations, the first based on significance testing (which we term Sig-2) and confidence interval width (CI-2), and give examples of their use.

6.2 Sample size calculation based on significance test

Here D is specified in advance and the desire is to determine the number of events needed to detect $\widetilde{D_2}$ (the estimate of D in the new study) falling below $D - \delta$ with power $1 - \beta$ at a one-sided significance level α. This effectively assumes that e_1 is infinite and σ_1^2 is 0. We can calculate the 'asymptotic' sample size e_2 as e_1 approaches ∞ and σ_1^2 approaches 0 as follows.

Starting with equation Sig-1 and substituting $\sigma_1^2 = \lambda / e_1$ we have

$$\begin{aligned} e_2 &= e_1 \left[\left(\frac{\delta}{\sigma_1 zz} \right)^2 - 1 \right]^{-1} \\ &= \frac{\lambda}{(\delta / zz)^2 - \lambda / e_1}. \end{aligned}$$

We want to compute the asymptotic sample size e_2 as $e_1 \to \infty$, which is

$$e_2 = \frac{\lambda zz^2}{\delta^2}. \tag{6.1}$$

To use this calculation, we need an estimate of λ. Thus the next step is to substitute the formula for λ in terms of D which we derived in Chapter 4, that is

$$\lambda = c_0 + c_1 D^{1.9} + c_2 (D \cdot cens)^{1.3},$$

where $c_0 = 2.66$, $c_1 = 1.26$, and $c_2 = -1.65$. Our sample size calculation is now

$$e_2 > \left(\frac{zz}{\delta}\right)^2 \left(2.66 + 1.26D^{1.9} - 1.09(D \cdot cens)^{1.3}\right). \qquad \text{(Sig-2)}$$

In order to keep consistent notation with Chapter 5 we continue to call the sample size for the new study e_2; although technically it is not a second study here as there is no first study. A point of note is that unlike the significance based calculation Sig-1 in Chapter 5, the only limit on δ is that it must be positive (e_2 tends to infinity as δ approaches 0).

Censoring proportion

Calculation Sig-2 includes *cens*, the proportion of censoring in the dataset. We believe that most researchers will have a good idea of the likely censoring proportion in their proposed study by the time they come to do a sample size calculation. They should know approximate survival rates for their disease and have planned the length of study follow up, so they should have a reasonable estimate of the proportion of patients who will not have had an event by the time of analysis. We would recommend that a sensitivity analysis is done by researchers; i.e. the calculation be repeated for a range of likely censoring proportions, to see how this affects the number of events and patients required.

6.2.1 Example

Suppose we want to detect a reduction in D of $\delta = 0.3$ with 90% power at the one-sided 5% significance level. Thus $zz = z_{1-\alpha} + z_{1-\beta} = 1.645 + 1.282 = 2.927$ and the sample size required is

$$e_2 = \frac{\lambda \times 2.93^2}{0.3^2} = 95.12\lambda.$$

Thus if we believe $D = 1.2$ and estimated 50% censoring (approximately the same as in the breast cancer example in Chapter 5), then

$$\lambda = 2.66 + 1.26 \times 1.2^{1.9} - 1.09(1.2 \times 0.5)^{1.3} = 3.88$$

and $e_2 = 370$.

This example will be used to inform the next section which considers how various parameters in the calculation Sig-2 affect sample size.

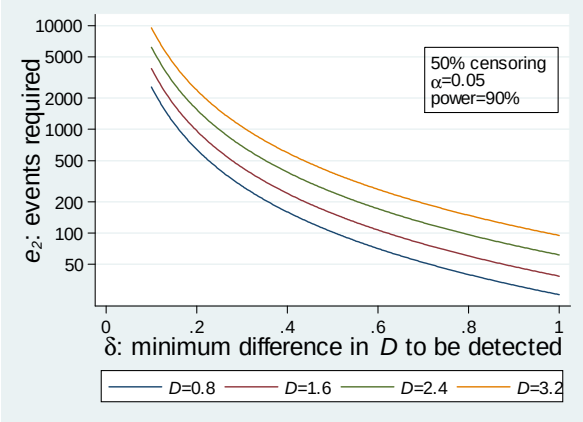

Figure 6.1: Sig-2: events required vs δ, for various D. Note log scale for events

6.2.2 Effect of parameters on calculation Sig-2

As with all sample size calculations, the smaller the difference to be detected (δ), the larger the sample size. Figure 6.1 shows the drop in number of events required in the example in Section 6.2.1, as δ increases from 0.1 to 0.5 for various values of D.

As calculation Sig-2 includes both D and the level of censoring as parameters, we can easily see the relationships between these two quantities and sample size.

Increasing D increases λ and hence increases sample size. Thus if the final value of D in a validation study is larger than was anticipated in the planning stages, this will likely mean the estimate of D in the study has less precision than was planned. Likewise, if the value of D is smaller than expected, the estimate of D in the validation study will be more precise. The increase in study size with D is more marked the smaller δ is, as shown in Figure 6.1.

Increasing censoring leads to a small decrease in number of events required, because equation 4.2 for λ includes a negative term for censoring, so λ decreases slightly as censoring increases. However although the number of events required may have decreased, the increased censoring leads overall to an increase in number of patients, which can be quite large. Figure 6.2 shows the relationship between level of censoring and (a) events and (b) patients, over δ. If at the time of analysis, the proportion of censored patients in the validation study is greater than was expected at the planning stages, the validation study is likely to have lower precision, because the original sample size calculation

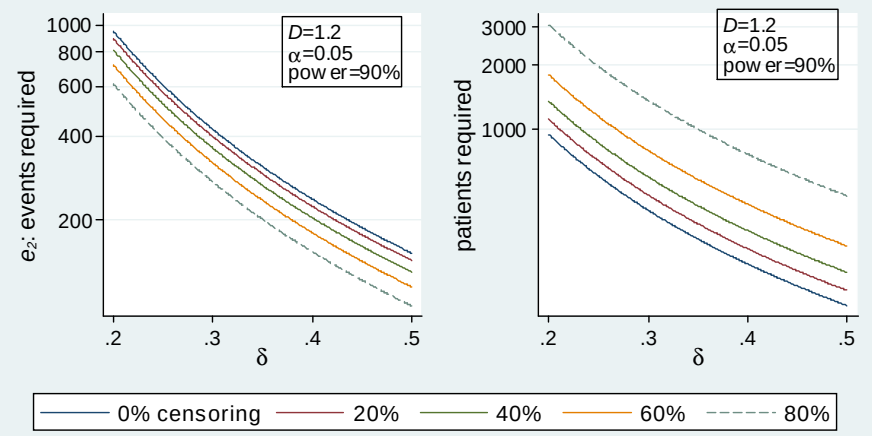

Figure 6.2: Sig-2: events (left) and patients (right) required vs δ. Note log scale for events

assumed a lower proportion of censoring (and so although slightly more events were needed, fewer patients were recruited).

6.3 Simulation study: significance based calculation Sig-2

To check the validity of the sample size calculation based on significance testing a simulation study is needed.

6.3.1 Method

1. Select δ, α, power and β (note $D_1 = \beta\kappa$ where $\kappa = \sqrt{8/\pi}$).

2. Calculate e_2 from the equation Sig-2 using the selected δ, α and power.

3. Generate dataset with $e_2/(1 - cens)$ records, exactly e_2 events and exact (random) censoring proportion *cens* using the method outlined in 5.3.1, under a proportional hazards model with linear predictor $(\beta - \delta/\kappa)X$, where $X \sim N(0, 1)$. This sample is regarded as created under the null hypothesis of inferiority: $H_0 : D \leq \beta\kappa - \delta$ (specifically, under the assumption that $D = \beta\kappa - \delta$).

4. Compute the D value for this sample; this is D_{20}. Estimate its standard error using a bootstrap with 200 replications; this is s_{20}.

5. Generate dataset with $e_2/(1 - cens)$ records, exactly e_2 events and exact censoring proportion *cens* using the method outlined in 5.3.1, under a proportional hazards

model with linear predictor βX. This sample is regarded as created under the alternative hypothesis of non-inferiority: $H_1 : D \geq \beta\kappa - \delta$ (specifically, under the assumption that $D = \beta\kappa$).

6. Compute the D value for this sample; this is D_{21}. Estimate its standard error using a bootstrap with 200 replications; this is s_{21}.

7. Compute test statistics z_0 and z_1 for testing H_0 and H_1 respectively:

$$z_0 = \frac{D_{20} - (D_1 - \delta)}{\sqrt{s_{20}^2}}$$

$$z_1 = \frac{D_{21} - (D_1 - \delta)}{\sqrt{s_{21}^2}}.$$

   Note that $D_1 = \beta\kappa$, and since D_1 is considered fixed, the standard error of D_1 is 0 (this quantity was termed s_1 in Chapter 5).

8. Repeat steps (3) to (7) 2000 times and store the results.

9. The type I error rate is estimated by the proportion of observations for which $z_0 > \Phi^{-1}(1 - \alpha)$; the power is estimated by the proportion of observations for which $z_1 > \Phi^{-1}(1 - \alpha)$.

These steps will be repeated for various values of power (80%, 90%), *cens* (0%, 40%, 80%), β (1, 2; hence D=1.6, 3.2) and δ (0.4, 0.5). α=0.05 for all.

6.3.2 Results

The results from this study are given in Table 6.1. Power and type I error are generally as expected.

For fixed combinations of β, power and δ, there is a slight decrease in observed power with increasing censoring, which is more noticeable when $\beta = 2.0$. This is expected as our model slightly overestimates λ when the proportion of censoring is low, while slightly underestimating it when censoring is higher (shown in Section 4.7), and the effect increases as D increases. This means that power is generally slightly higher for the scenarios with 0% censoring than for the scenarios with 80% censoring; however the observed differences are small.

The observed type I error appears to be slightly reduced below the expected 5% for the simulations with $\beta = 2.0$; again this may be related to the errors in estimation of λ. However, if this decrease is a real effect, it is only small, and it is better that the observed error is lower than expected, rather than higher.

Simulation parameters				Observed				
β	Power	δ	cens	e_2	% Type 1 (*se*)		% Power (*se*)	
1.0	80%	0.4	0	228	5.1	(0.50)	82.8	(0.84)
			40	205	4.3	(0.46)	78.8	(0.91)
			80	170	4.4	(0.47)	80.1	(0.89)
		0.5	0	146	5.3	(0.51)	81.3	(0.87)
			40	131	4.2	(0.46)	83.2	(0.84)
			80	109	4.6	(0.48)	79.1	(0.91)
	90%	0.4	0	316	5.1	(0.50)	91.0	(0.64)
			40	283	4.4	(0.47)	91.8	(0.61)
			80	235	5.3	(0.51)	92.2	(0.60)
		0.5	0	202	4.9	(0.49)	91.5	(0.62)
			40	181	4.5	(0.47)	93.0	(0.57)
			80	151	4.8	(0.49)	90.3	(0.66)
2.0	80%	0.4	0	543	4.7	(0.48)	77.1	(0.94)
			40	485	3.8	(0.44)	81.9	(0.86)
			80	400	4.4	(0.47)	82.5	(0.85)
		0.5	0	348	3.9	(0.44)	79.8	(0.90)
			40	311	4.0	(0.45)	82.3	(0.85)
			80	256	4.7	(0.48)	83.2	(0.84)
	90%	0.4	0	752	3.5	(0.42)	89.8	(0.68)
			40	672	3.6	(0.42)	89.8	(0.68)
			80	554	4.9	(0.49)	90.9	(0.64)
		0.5	0	482	3.5	(0.42)	88.6	(0.71)
			40	430	3.5	(0.42)	90.7	(0.65)
			80	355	4.1	(0.45)	91.2	(0.63)

Table 6.1: Simulation study results for significance based calculation Sig-2

6.3.3 Conclusion

This sample size calculation based on significance testing appears to work well; although slight errors in power (of one or two percent) were seen as a result of imperfect estimation of λ. If D and the level of censoring are correctly estimated then the resulting study will have the precision required, or very close to it. If D or the censoring proportion is underestimated, then the precision of estimates will be lower; conversely, overestimating either of these two parameters will result in higher precision. The type I error may be slightly lower than expected (by one percent or so) when D is greater than 3, but such values of D are rare in practice, as will be shown in Chapter 8 (Figure 8.8).

6.4 Sample size calculation based on CI for D

In order to develop a sample size calculation based on the width of the $(1-\alpha)\times 100\%$ two-sided confidence interval (CI) for D for the situation where a target value of D is known, we must substitute λ back in to equation CI-1, to give us

$$e_2 = \lambda \left(\frac{z_{(1-\alpha/2)}}{w} \right)^2 .$$

Using the same estimate of λ from D as in the previous section, our final sample size calculation is

$$e_2 = \left(\frac{z_{(1-\alpha/2)}}{w} \right)^2 \left(2.66 + 1.26D^{1.9} - 1.09(D \cdot cens)^{1.3} \right). \qquad \text{(CI-2)}$$

Once again, in order to use this calculation researchers must have some idea of the magnitude of D, and the proportion of censoring they are likely to observe.

6.4.1 Example

If we wish to estimate D with a 95% confidence interval of half-width 0.15, then the sample size calculation is

$$\begin{aligned} e_2 &= \lambda \left(\frac{z_{0.975}}{0.15} \right)^2 \\ &= 170.7 \times \lambda. \end{aligned}$$

If we believe that $D = 1.2$ and the censoring level will be 50%, as in the example in the previous section, then $\lambda = 3.88$ and $e_2 = 663$.

6.4.2 Effect of parameters on calculation CI-2

The narrower the confidence interval desired, the larger the sample size output by calculation CI-2. Figure 6.3 shows graphically the relationship between half-width w and the number of events required for a 95% CI for the example in Section 6.4.1, using values of D from 0.8 to 3.2 and 50% censoring. The width of confidence interval has probably the greatest impact on sample size, especially for higher values of D. This graph also allows

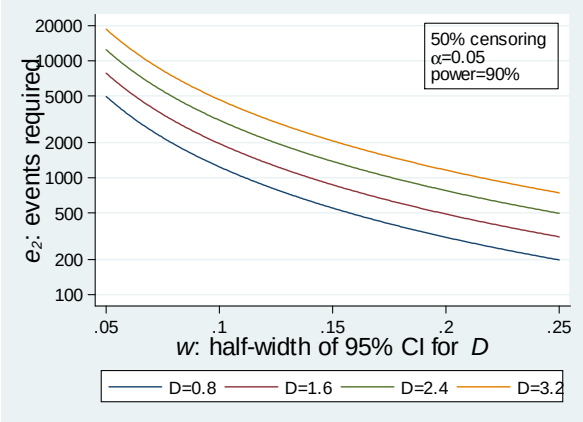

Figure 6.3: CI-2: events required (log scale) vs w. Note log scale for events

us to consider the trade off between events required and CI width; researchers may want to know what precision a certain study size will 'buy' them given D.

As for all the other sample size calculations in this and the previous chapter, increasing D increases sample size, and this relationship is stronger when w is small. This is also illustrated in Figure 6.3. This graph clearly shows the importance of calculating a range of sample sizes, especially when the value of D is not very certain, since it can have a large effect on the number of events required.

As for calculation Sig-2, increasing censoring leads to a small decrease in the number of events required, but an increase in the number of patients required. Figure 6.4 shows separately events and patients required vs w for a fixed value of $D = 1.2$ and five different censoring levels. If the proportion of censored patients in the validation study is larger than was expected at the planning stages, the validation study is likely to have wider confidence intervals than desired.

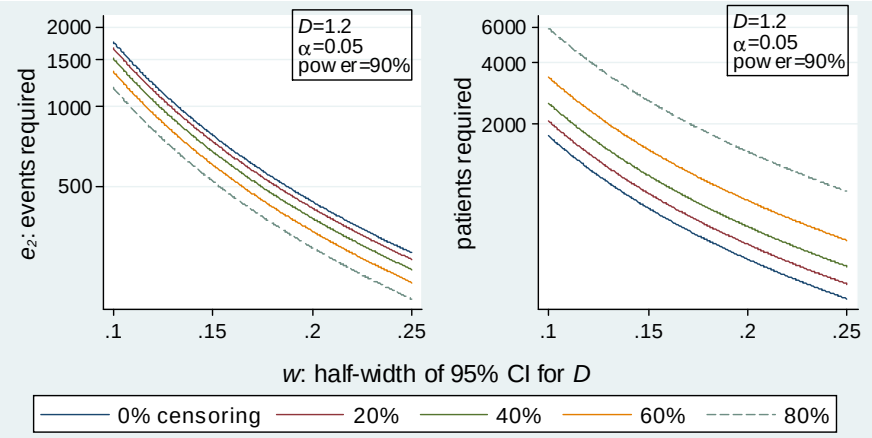

Figure 6.4: CI-2: events (left) and patients (right) required vs w. Note log scale for events

6.5 Simulation study for CI based calculation CI-2

A simulation study is performed to check the validity of sample size calculation CI-2.

6.5.1 Method

1. Compute e_2 from D, $cens$, α and the desired half-width of CI w using equation CI-2.
2. Generate dataset with $e_2/(1-cens)$ records, exactly e_2 events and exact censoring proportion $cens$ using the method outlined in 5.3.1, under a proportional hazards model with linear predictor βx.
3. Calculate D in this dataset and record it; this is $\widetilde{D}$.
4. Repeat steps (2) and (3) 2000 times.
5. The proportion of repetitions for which $\widetilde{D} \in (\beta\kappa - w, \beta\kappa + w)$ gives the % CI which has width $\pm w$ in the simulated dataset. This should approximate $1-\alpha$, if the sample size calculation and estimation of λ are correct.

These steps will be repeated for various values of $cens$ (0%, 40%, 80%), β (1, 2; thus D=1.6, 3.2), w (0.1, 0.2, 0.3), and α (0.05, 0.10).

6.5.2 Results

The results are given in Table 6.2 and show that the proportion of $\widetilde{D} \in (\beta\kappa - w, \beta\kappa + w)$ is close to that desired across all parameter combinations. As seen with the significance-

based calculation Sig-2, the small errors in our estimation of λ means that we would expect to see a slightly higher than desired proportion of $\widetilde{D}$ falling within the interval $(\beta\kappa - w, \beta\kappa + w)$ when $cens = 0\%$ and a slightly lower proportion when $cens = 80\%$. This pattern is observed in some combinations of parameters but the differences are not large.

Simulation Parameters					Observed (95% CI)			Observed (90% CI)		
						% of $\widetilde{D}$ (se)			% of $\widetilde{D}$ (se)	
β	D	w	$cens$	λ	mean $e2$	within $\beta\kappa \pm w$		mean $e2$	within $\beta\kappa \pm w$	
1.0	1.6	0.10	0%	5.72	2199	94.5	(0.51)	1549	90.1	(0.67)
			40%	4.80	1844	96.3	(0.42)	1299	91.7	(0.62)
			80%	3.45	1324	94.9	(0.49)	933	89.9	(0.68)
		0.20	0%	5.72	550	95.7	(0.46)	388	90.2	(0.67)
			40%	4.80	461	95.2	(0.48)	325	89.8	(0.68)
			80%	3.45	331	94.3	(0.52)	234	88.0	(0.73)
		0.30	0%	5.72	245	95.5	(0.47)	173	88.6	(0.71)
			40%	4.80	205	95.0	(0.49)	145	91.4	(0.63)
			80%	3.45	148	94.2	(0.52)	104	88.9	(0.70)
2.0	3.2	0.10	0%	14.09	5414	94.9	(0.49)	3813	89.2	(0.70)
			40%	11.81	4539	95.9	(0.45)	3197	92.2	(0.64)
			80%	8.48	3259	95.2	(0.48)	2296	92.2	(0.63)
		0.20	0%	14.09	1354	95.3	(0.47)	954	90.5	(0.66)
			40%	11.81	1135	95.7	(0.45)	800	91.2	(0.63)
			80%	8.48	815	95.7	(0.45)	574	91.1	(0.64)
		0.30	0%	14.09	602	94.2	(0.52)	424	89.5	(0.69)
			40%	11.81	505	95.9	(0.45)	356	91.2	(0.64)
			80%	8.48	363	95.3	(0.47)	256	90.3	(0.66)

Table 6.2: Simulation study results for CI based calculation CI-2

6.5.3 Conclusion

The sample size calculation based on CI width works well. If D and the level of censoring are correctly estimated then the resulting confidence interval will be the desired width. If D or the censoring proportion is underestimated, then the CI will be wider than desired. Conversely, overestimating either of these two parameters will result in a narrower CI than was planned.

Depending on the strength of prior information on D and the censoring proportion, it may be pertinent for researchers to perform a number of sample size calculations in order to obtain a range of study sizes which cover the most likely eventualities. This is straightforward with calculations Sig-2 and CI-2 since they include D and censoring proportions as explicit terms.

6.6 Discussion

In this chapter we have presented sample size calculations for the situation where an estimate or guess of D is available, but there is no estimate of the uncertainty around it. This is likely to be a common scenario; we envisage the most likely situation to be where researchers have a target value of D to investigate, but there may also be occasions where D was estimated from a previous study but no estimate of $SE(D)$ is available, or where estimates of D and $SE(D)$ are available but researchers do not believe the source provides reliable evidence.

As in Chapter 5, both the significance based (Sig-2) and the confidence interval based (CI-2) calculation appear to be correctly powered and both show no bias associated with censoring. Calculation Sig-2 did show some minor fluctuations in observed power due to the imperfect nature of our estimation of λ, specifically that observed power was sometimes slightly too low for the lowest censoring level considered (0%), and slightly too high for the highest censoring level (80%). However, the errors were a matter of 3% (absolute) at most so are not too concerning.

The sample size calculations in this chapter have the advantage that they contain parameters for D and censoring, so it is easy for researchers to calculate a a range of sample sizes for the proposed study in order to consider a variety of possible values of D and censoring proportions. We particularly recommend researchers do this, since estimates of D and censoring rates may not be based on strong evidence if a previous good-quality study is not available.

Having these calculations available may be very useful, but in the situation where no previous estimate of D is available, a researcher's first question is likely to be, how do we come up with an educated guess of the value of D to input into these calculations? In Chapters 8 and 9 we will collate and present values of D for various research areas, and anticipate that this 'library' will provide a starting point for research where no comparable previous studies have been performed.

In the next chapter we will illustrate the use of the sample size calculations presented in this and the previous chapter, using real life studies as starting points.

Chapter 7

Sample size examples

In this chapter we will more comprehensively illustrate the four sample size calculations developed in Chapters 5 and 6. First we will perform each calculation using observed parameters from ten real datasets. This will cover a wide variety of real scenarios and will show how the required numbers of events and patients change according to the various parameters of the dataset. Secondly, we will fix the number of events and calculate how much precision this sample size will 'buy' us under each calculation, again using real datasets for estimates of the parameters required. This will show more clearly how the four sample size calculations perform (in terms of precision obtained) when all other variables are held equal. Finally, we will show that precision can be considered as a proportion of D in calculations Sig-2 and CI-2, and illustrate how this can be exploited to avoid the loss of precision seen when D is higher than anticipated in a study.

Let us first recapitulate the four calculations and notation. First, let z_x be the x-quantile of the standard normal distribution.

Calculation Sig-1: Validating D from previous study: significance based

When estimates of D and the standard error of D are available from a previous study, the number of events required to detect a difference in D of δ with one-sided α and power $1-\beta$ is

$$e_2 = e_1 \left[\left(\frac{\delta}{\sigma_1 zz} \right)^2 - 1 \right]^{-1}, \qquad \text{(Sig-1)}$$

where e_1 is the number of events and σ_1 the standard error of D from the previous study, and $zz = z_{1-\alpha} + z_{1-\beta}$. For this calculation, there is a restriction on the difference in D that can be detected: $\delta > \sigma_1 zz$.

Calculation CI-1: Validating D from previous study: CI based

When a previous study is available, the number of events required in order that an estimate of D has a two-sided 95% confidence interval (CI) of half width w is

$$e_2 = e_1 \sigma_1^2 \left(\frac{z_{1-\alpha/2}}{w} \right)^2 . \tag{CI-1}$$

Calculation Sig-2: Point estimate of D: significance based

When researchers have a target value of D but no previous estimate of its standard error, the number of events required to detect a difference in D of δ with one-sided α and power $1 - \beta$ is

$$e_2 = \left(\frac{zz}{\delta} \right)^2 \left(2.66 + 1.26 D^{1.9} - 1.09 (D \cdot cens)^{1.3} \right) , \tag{Sig-2}$$

where D is the point estimate of D, $cens$ is the estimated proportion of censored records in the final study dataset, and $zz = z_{1-\alpha} + z_{1-\beta}$ as before.

Calculation CI-2: Point estimate of D: CI based

When researchers have a target value of D but no previous estimate of its standard error, the number of events required in order that an estimate of D has a two-sided 95% confidence interval (CI) of half width w is

$$e_2 = \left(\frac{z_{(1-\alpha/2)}}{w} \right)^2 \left(2.66 + 1.26 D^{1.9} - 1.09 (D \cdot cens)^{1.3} \right) . \tag{CI-2}$$

7.1 Effect of dataset parameters on sample size

For each of the ten real datasets, the required parameters will be given and the four calculations performed.

For the first calculation Sig-1, a δ close to the minimum δ detectable is chosen; for the third calculation Sig-2 $\delta = 0.2$ is used (hence the two calculations are not always directly comparable in this table). For the CI based calculations (CI-1 and CI-2) $w = 0.1$ is used, so these two calculations are comparable.

For Sig-1 and Sig-2, $\alpha = 0.05$ and power is 90%; for CI-1 and CI-2, the confidence intervals are always 95%.

Table 7.1 shows the number of events and patients required for the ten datasets under each of the four calculations, for these selected values. More detailed results showing how sample size varies with w and δ for the four calculations are given in Appendix D.

								Events (patients) required by calculation							
	Events				$SE(D)$		Chosen	Sig-1		Sig-2		CI-1		CI-2	
Dataset	(e_1)	Patients	*cens*	D	(σ_1)	Min δ	δ for Sig-1			$\delta = 0.2$		$w = 0.1$		$w = 0.1$	
APC	338	475	29%	0.85	0.095	0.28	0.35	578	(813)	711	(1000)	1172	(1648)	1276	(1794)
GLI	273	411	34%	1.15	0.120	0.35	0.4	918	(1383)	819	(1233)	1511	(2275)	1470	(2214)
LEG	97	200	52%	2.07	0.267	0.78	0.85	529	(1091)	1261	(2600)	2657	(5479)	2262	(4664)
LVA	128	137	7%	1.43	0.231	0.68	0.75	555	(595)	1086	(1163)	2624	(2809)	1948	(2085)
MYE	856	1057	19%	0.77	0.070	0.20	0.25	1750	(2161)	705	(871)	1612	(1991)	1265	(1563)
PBC	125	312	60%	2.70	0.268	0.78	0.85	716	(1788)	1690	(4219)	3449	(8609)	3032	(7568)
RBC	1518	2982	49%	1.09	0.053	0.15	0.2	2291	(4501)	731	(1436)	1639	(3220)	1311	(2576)
SEER DE	1540	13533	89%	1.86	0.048	0.14	0.2	1500	(13182)	771	(6776)	1364	(11987)	1382	(12145)
SEER NM	400	4422	91%	1.71	0.084	0.24	0.3	818	(9043)	690	(7628)	1085	(11995)	1238	(13687)
STE	460	3413	87%	1.25	0.084	0.25	0.3	940	(6975)	591	(4385)	1247	(9253)	1060	(7865)

Table 7.1: Example sample size calculations based on parameters of real datasets

Significance based calculations

Calculation Sig-1 – the significance based calculation for validating D from previous study – is unique among the four calculations in that it is the only one where the difference to be detected is constrained (beyond the restriction $\delta > 0$ which applies to all four calculations). If the previous study was relatively small the minimum difference can be quite large, and even the largest datasets we consider in this chapter, with approximately 1500 events, have a minimum δ of around 0.15. However the sample size drops quite rapidly as the δ increases, and a δ of about 0.05 above the minimum generally produces a more reasonable sample size in terms of numbers of events. It is worth remembering that this calculation is for a non-inferiority study and so if the minimum δ detectable is large, this may not be the best calculation to use, since a positive result from such a study will not add much information to what is already known.

For the two datasets where the δ to be detected in Table 7.1 is 0.2 for both Sig-1 and Sig-2 (RBC and SEER DE), we can compare the calculations directly and see that Sig-1 outputs 2291 and 1500 events for the RBC and SEER DE parameters, while Sig-2 outputs 731 and 771. In fact, to detect a particular δ with a fixed power and α, the number of events required according to calculation Sig-1 (call it $e_{\text{Sig-1}}$) is always larger than the number required according to calculation Sig-2 (call it $e_{\text{Sig-2}}$). To show this, consider calculation Sig-1 in terms of λ (Formula 5.4 from Chapter 5):

$$\begin{aligned} e_{\text{Sig-1}} &= \frac{\lambda}{(\delta/zz)^2 - \sigma_1^2} \\ &= \frac{\lambda(\frac{zz}{\delta})^2}{1 - (\frac{zz}{\delta})^2\sigma_1^2} \\ &= \frac{e_{\text{Sig-2}}}{1 - \left(\frac{\sigma_1 zz}{\delta}\right)^2} \end{aligned}$$

When using calculation Sig-1, we have the restriction $\delta > \sigma_1 zz$, hence

$$\begin{aligned} \frac{\sigma_1 zz}{\delta} &< 1 \\ 1 - \left(\frac{\sigma_1 zz}{\delta}\right)^2 &< 1 \\ e_{\text{Sig-1}} &> e_{\text{Sig-2}}. \end{aligned}$$

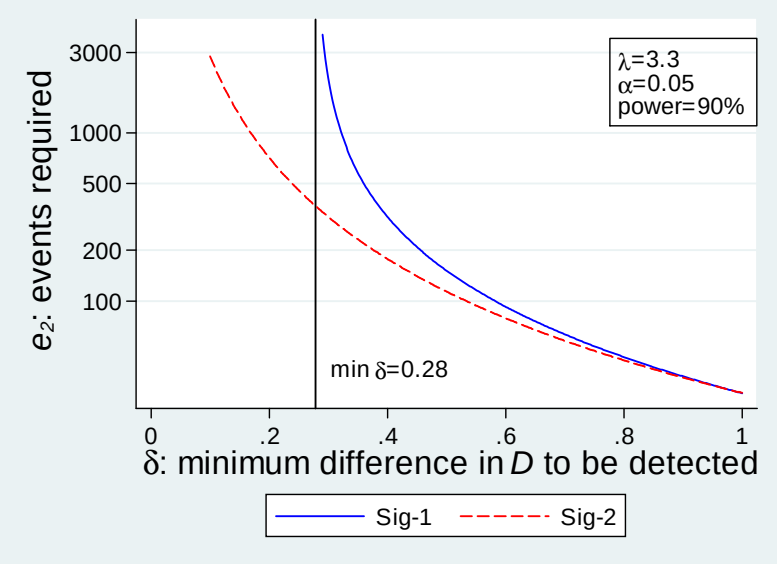

Figure 7.1: Events required vs δ: comparison of calculations Sig-1 and Sig-2. Note log scale of y axis

Under calculation Sig-1, as $\delta \rightarrow \sigma_1 zz$, the number of events $e_{\text{Sig-1}}$ approaches infinity; but using Sig-2, the lower limit on δ is 0, so $e_{\text{Sig-1}}$ and $e_{\text{Sig-2}}$ diverge as δ approaches $\sigma_1 zz$. This is illustrated in Figure 7.1, which uses the parameters of dataset APC as a basis for the calculations.

CI-based calculations

Calculation CI-1 – the CI based calculation for validating D from previous study – generally gives reasonable sample sizes for a half-width of 0.1. It is only when D is high (over 2.0, say) and the previous study was quite small (less than 300, say) that this sample size is pushed over 2000 events. Of course, if the censoring proportion is expected to be high this may still be an unfeasibly high sample size.

The sample sizes required by CI-1 and CI-2 are generally fairly similar. The only difference between CI-1 and CI-2 is that CI-1 uses observed parameters σ_1 and e_1 to estimate λ, while CI-2 estimates λ using the prediction model Equation 4.2 developed in Chapter 4. Thus in a single situation, any difference in output sample sizes from the two calculations is due to error in predicting λ with D and *cens*. Such errors can be positive or negative (as shown in Table 4.5), and thus the sample size from CI-2 may be either higher or lower than the sample size from CI-1.

7.2 Effect of calculation choice on precision

A different way of comparing the sample size calculations is to consider what precision (in terms of δ or w) each of the four sample size calculations can 'buy' with a fixed number of events. To do this we need to rearrange the four calculations to be in terms of δ or w; that is

$$\text{Sig-1: } \delta = \sigma_1 zz\sqrt{\tfrac{e_1}{e_2}+1} \qquad \text{CI-1: } w = \sigma_1 z_{1-\alpha/2}\sqrt{\tfrac{e_1}{e_2}}$$

$$\text{Sig-2: } \delta = zz\sqrt{\tfrac{\lambda}{e_2}} \qquad \text{CI-2: } w = z_{1-\alpha/2}\sqrt{\tfrac{\lambda}{e_2}}$$

where $\lambda = 2.66 + 1.26D^{1.9} - 1.09(D \cdot cens)^{1.3}$ as previously. Using these expressions, we can calculate what precision a particular number of events (e_2) will buy us, assuming for Sig-1 and CI-1 a first study with observed parameters e_1, D and σ_1, and for Sig-2 and CI-2 estimates of D and $cens$. The precision obtained with study sizes of 200, 500 and 1000 events (δ or CI half-width w) is given in Table 7.2 for each of the four calculations and parameters provided by the ten real datasets.

						Precision obtained from a second study of given size											
						200 events				500 events				1000 events			
	Events				$SE(D)$	Sig-1	Sig-2	CI-1	CI-2	Sig-1	Sig-2	CI-1	CI-2	Sig-1	Sig-2	CI-1	CI-2
Dataset	(e_1)	Patients	*cens*	D	(σ_1)	δ	δ	w	w	δ	δ	w	w	δ	δ	w	w
APC	338	475	29%	0.85	0.095	0.46	0.38	0.24	0.25	0.36	0.24	0.15	0.16	0.32	0.17	0.11	0.11
GLI	273	411	34%	1.15	0.120	0.54	0.40	0.27	0.27	0.44	0.26	0.17	0.17	0.40	0.18	0.12	0.12
LEG	97	200	52%	2.07	0.267	0.95	0.59	0.36	0.34	0.85	0.32	0.23	0.21	0.82	0.22	0.16	0.15
LVA	128	137	7%	1.43	0.231	0.87	0.47	0.36	0.31	0.76	0.29	0.23	0.20	0.72	0.21	0.16	0.14
MYE	856	1057	19%	0.77	0.070	0.47	0.38	0.28	0.25	0.34	0.24	0.18	0.16	0.28	0.17	0.13	0.11
PBC	125	312	60%	2.70	0.268	1.00	0.58	0.42	0.39	0.88	0.37	0.26	0.25	0.83	0.26	0.19	0.17
RBC	1518	2982	49%	1.09	0.053	0.45	0.38	0.29	0.26	0.31	0.24	0.18	0.16	0.25	0.17	0.13	0.11
SEER DE	1540	13533	89%	1.86	0.048	0.41	0.39	0.26	0.26	0.28	0.25	0.17	0.17	0.22	0.18	0.12	0.12
SEER NM	400	4422	91%	1.71	0.084	0.43	0.37	0.23	0.25	0.33	0.23	0.15	0.16	0.29	0.17	0.10	0.11
STE	460	3413	87%	1.25	0.084	0.45	0.34	0.25	0.23	0.34	0.22	0.16	0.15	0.30	0.15	0.11	0.10

Table 7.2: Precision in estimate of D obtained from studies with 200, 500 or 1000 events, according to calculations Sig-1–CI-2

As seen in Table 7.1, calculations CI-1 and CI-2 behave similarly, both resulting in approximately the same precision in all ten situations.

In all cases calculation Sig-1 gives less precision than Sig-2 with the same number of patients. Additionally, increasing the number of events in the second study has a lesser impact on the precision available from Sig-1 than from the other three calculations. This is because for calculations CI-1, Sig-2 and CI-2, the precision available is proportional to $\sqrt{1/e_2}$, which means that increasing e_2 by a factor of k reduces w or δ by a factor of $\sqrt{k}$. The precision available from calculation Sig-1, however, is proportional to

$$\sqrt{\frac{e_1}{e_2} + 1}.$$

This means that increasing the number of events in the second study from e_2 to $k \cdot e_2$ increases the precision available by a factor of

$$\sqrt{k}\sqrt{\frac{e_1 + e_2}{e_1 + ke_2}},$$

which is less than $\sqrt{k}$ and dependent on both e_1 and e_2 .

So, using calculation Sig-1, if $e_1 = 400$, doubling e_2 from 100 to 200 reduces δ by a factor of 1.29, whereas doubling from 200 to 400 reduces δ by a factor of 1.22. If $e_1 = 1000$, the corresponding factors for the same increases in e_2 are 1.35 and 1.31. When using calculations CI-1, Sig-2 and CI-2, doubling e_2 always reduces δ or w by a factor of $\sqrt{2} = 1.41$.

7.3 Precision as a percentage of D

Having D – but importantly, not $SE(D)$ – explicit in the calculations Sig-2 and CI-2 allows the possibility of not directly specifying δ or w as values but rather as functions of D itself. This allows the situation where we can specify a sample size which allows a δ of, say, 10% of D, regardless of the actual value of D.

This has two key benefits. Firstly, when D is thought to be high (say, $D \geq 2$) the sample sizes required by Sig-2 and CI-2 can get very large if the same fixed value of δ or w is desired as for a smaller D. Considering δ or w as a proportion of D instead may lead researchers to think more carefully about what level of precision is actually acceptable

when D is large, which could lead to more relaxed constraints and a smaller sample size. Secondly, it should prevent the loss of precision seen if the D observed in the new study is higher than anticipated in the planning stages.

We are able to do this with calculations Sig-2 and CI-2 because the prediction model for λ implicitly adjusts $SE(D)$ as D changes; it cannot be done with Sig-1 and CI-1 as $SE(D)$ is an explicit term in these two calculations.

The theory is similar for both Sig-2 and CI-2; in order to explain it more clearly we focus on Sig-2.

7.3.1 Sig-2

Let us replace δ in calculation Sig-2 with pD, where p is the percentage of D that we are happy with, expressed as a proportion (so for 10%, $p = 0.1$). This gives us

$$e_2 = \left(\frac{zz}{pD}\right)^2 \left(2.66 + 1.26D^{1.9} - 1.09(D \cdot cens)^{1.3}\right). \tag{7.1}$$

Figure 7.2 shows the sample size profiles for a desired precision of 10%, 15% and 20% of D (assuming 30% censoring in this case). We can see that when we specify δ as a percentage of D, the sample size actually decreases as D increases, the opposite to what happens when δ is a fixed value. This means that we can choose a sample size for a particular value of D, and if the D in the new study turns out to be higher than this, we still have the percentage precision we originally aimed for (assuming the proportion of censored observations is still the same).

However, if D in the new study is lower than planned, then we lose precision quite rapidly and will not be able to achieve the δ that we aimed for. A pragmatic solution to enable precision to be controlled regardless of whether D is higher or lower than originally planned, is to specify both a fixed value of δ and a percentage of D that are acceptable. At each possible value of D take whichever option of the two gives the larger value of precision (and hence smaller sample size). Thus we have a 'composite' sample size calculation.

For example, we may say that we want precision of $\delta = 0.15$ or 10% of D, whichever requires the smaller sample size at each possible value of D. We can plot the required sample size versus D for both $\delta = 0.15$ and $\delta = 0.1D$ and read off the resulting profile,

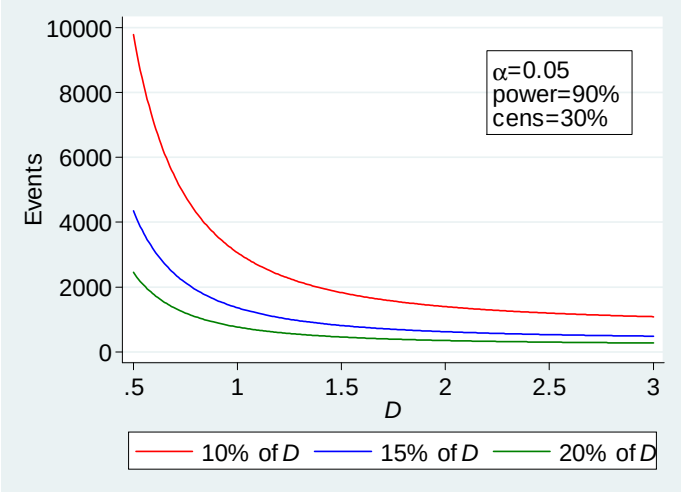

Figure 7.2: Sample sizes required from Sig-2 for δ considered as a percentage of D

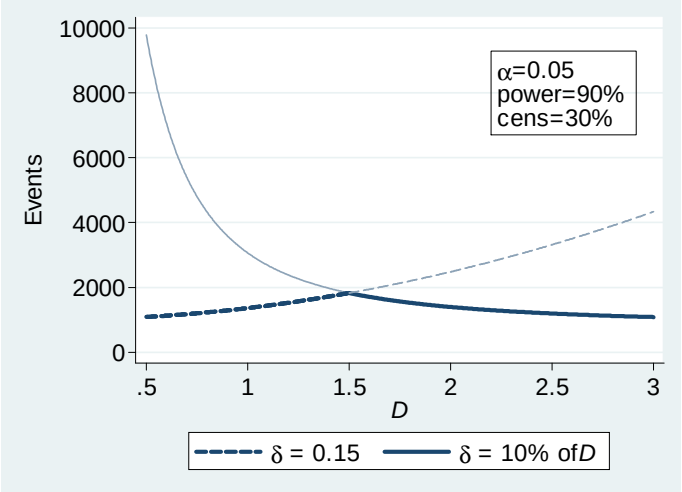

Figure 7.3: Events vs D for composite Sig-2 calculation: $\delta = 0.15$ or 10% of D

shown in Figure 7.3. The heavier lines give the required sample size at each value of D; the maximum occurs at $0.15/0.1 = 1.5$ and requires 1827 events.

Thus, in this particular situation a sample size of 1827 events (2610 patients) would achieve the desired precision regardless of the value of D output from the new study. From looking at Figure 7.3 we can see that as D either increases or decreases away from the peak at $D = 1.5$, the required sample size decreases (to 1089 events when $D = 0.5$, and 1084 patients when $D = 3$). This means that if the value of D from a study with 1827 events turns out to be greater than 1.5, the observed precision will be better than 10% of D, and if D turns out to be less than 1.5, the observed precision will be better than 0.15.

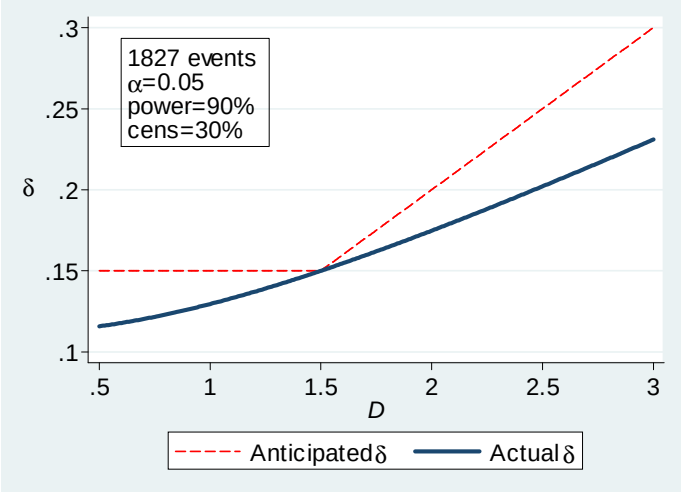

Figure 7.4: Anticipated and actual precision vs D from study with 1827 events

The actual precision obtained can be calculated quite simply by rearranging equation 7.1, and this is shown in Figure 7.4. The planned precision ($\delta = 0.15$ or $\delta = 0.1D$) is shown by a red dashed line, while the actual precision that would be obtained with 1827 events is shown as a solid navy line. The two lines coincide at $D = 1.5$.

7.3.2 CI-2

We can use the same procedure for calculation CI-2, this time replacing w – the half-width of the confidence interval – with kD. The resulting calculation is

$$e_2 = \left(\frac{z_{(1-\alpha/2)}}{kD}\right)^2 \left(2.66 + 1.26D^{1.9} - 1.09(D \cdot cens)^{1.3}\right).$$

Figure 7.5 shows how this required sample size varies with D, for confidence interval half-widths of 10%, 15%, and 20% of the value of D.

As for calculation Sig-2, a composite sample size calculation may be used to ensure a minimum precision is achieved across the possible range of D.

Table 7.3 gives the required sample sizes output from calculations Sig-2 and CI-2, for various composite precision limits and censoring levels.

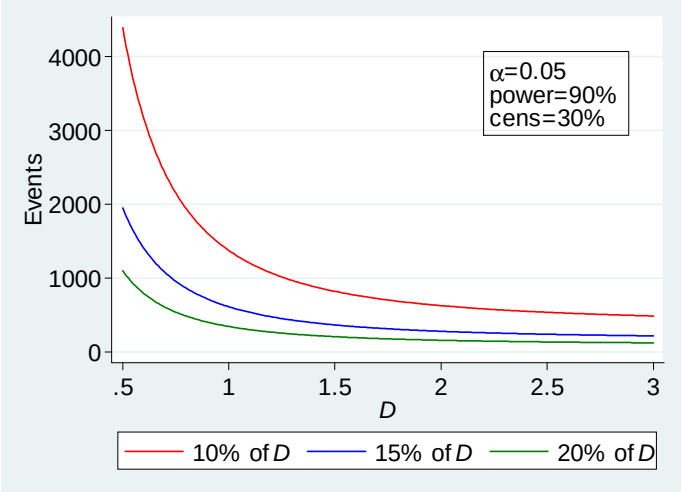

Figure 7.5: Sample sizes required from CI-2 for w considered as a percentage of D

Precision		Censoring	Sig-2		CI-2	
Fixed δ or w	δ or w as % of D	proportion	Events	(Patients)	Events	(Patients)
0.1	10%	20%	3183	(3979)	1428	(1785)
0.1	10%	40%	2928	(4880)	1314	(2190)
0.1	10%	60%	2630	(6575)	1180	(2950)
0.1	10%	80%	2300	(11500)	1032	(5160)
0.15	10%	20%	1918	(2398)	1120	(1400)
0.15	10%	40%	1726	(2877)	1074	(1790)
0.15	10%	60%	1501	(3753)	1020	(2550)
0.15	10%	80%	1253	(6265)	959	(4795)
0.2	10%	20%	1469	(1837)	861	(1077)
0.2	10%	40%	1312	(2187)	774	(1290)
0.2	10%	60%	1129	(2823)	674	(1685)
0.2	10%	80%	926	(4630)	562	(2810)
0.1	20%	20%	2497	(3122)	555	(694)
0.1	20%	40%	2393	(3989)	520	(867)
0.1	20%	60%	2272	(5680)	479	(1198)
0.1	20%	80%	2138	(10690)	434	(2170)
0.15	20%	20%	1237	(1547)	659	(824)
0.15	20%	40%	1159	(1932)	589	(982)
0.15	20%	60%	1068	(2670)	507	(1268)
0.15	20%	80%	967	(4835)	416	(2080)
0.2	20%	20%	796	(995)	357	(447)
0.2	20%	40%	732	(1220)	329	(549)
0.2	20%	60%	658	(1645)	295	(738)
0.2	20%	80%	575	(2875)	258	(1290)

Table 7.3: Sample sizes required by calculations Sig-2 and CI-2 for various composite precision limits and censoring levels

7.4 Discussion

7.4.1 Which calculations should be used?

Performing the four sample size calculations for a variety of different input parameters gives a good idea of how they perform in real life. In general, to obtain a fairly precise estimate of D, more than 500 events are required, and to detect a reasonably small difference in D more than 1000 events; often 2000 or more. If there is no previous study data with which to estimate $\lambda = e \cdot var(D)$, then Sig-1 and CI-1 cannot be used; however if an estimate of λ is available, how do researchers choose between using Sig-1 or Sig-2, or CI-1 or CI-2?

CI based calculations

Firstly, the decision between CI-1 and CI-2. If it is desired to base sample size on the confidence interval around the estimate of D in the new study, CI-1 and CI-2 perform approximately equivalently. Any difference between CI-1 and CI-2 is caused by errors in the prediction of λ from equation 4.2; however, since in CI-2 there is no first study, we cannot know the magnitude of this error until the proposed study is finished. Additionally, in Chapter 4 we found that we can't say for certain that the value of λ in a first study is going to be more accurate than λ predicted from D and *cens* using equation 4.2, when it comes to estimating λ in the second study (although our research in Chapter 4 into the sampling distribution of λ and accuracy of equation 4.2 is quite limited).

CI-2 is more flexible than CI-1 as it includes D and *cens* explicitly in the calculation, making it easy for researchers to calculate a range of sample sizes for different values of these two parameters. For this reason we would recommend using CI-2 rather than CI-1, even if an estimate of $SE(D)$ is available from a previous study. We would however add the caveat, that if a previous study exists and λ from that previous study is very different to the λ predicted from equation 4.2, researchers should consider the sample sizes output by both CI-1 and CI-2, as they may be quite different and we cannot be sure which estimate of λ is more accurate.

Significance based calculations

If a significance based calculation is preferred, we have seen in this chapter that there can be a large difference between the sample sizes output by Sig-1 and Sig-2 for the same input parameters. The restriction on δ implicit in Sig-1 makes this calculation less flexible than Sig-2; but even aside from this, Sig-1 will always output a larger sample size than Sig-2, everything else being equal. So, if researchers do have an estimate of $SE(D)$ and hence λ from a previous study, should they ignore it and use calculation Sig-2, just to have a smaller sample size? We cannot see a strong argument against this.

Firstly, as already mentioned in Chapter 4, we cannot be sure that the value of λ from a first study is going to be more accurate than a λ predicted using equation 4.2; so using all available information from the previous study is not necessarily going to provide a 'truer' sample size. Secondly, in clinical trials of time-to-event data, the standard error of the effect size (hazard ratio) is not generally involved in the sample size calculation, even though an estimate of it might be available; so we don't believe there is a philosophical argument in favour of using every scrap of available information, simply because it is there to be used. Thus we can't see a good reason to insist that Sig-1 should always be used if an estimate of $SE(D)$ is available. There are also two practical arguments in favour of Sig-2. As for CI-2, using Sig-2 means that a range of sample sizes can be easily calculated for different values of D and *cens*. Finally, the reduction in D to be detected is constrained in Sig-1, often with such a high lower bound that it is effectively useless. For this reason our current recommendation would be for researchers to use Sig-2 over Sig-1.

7.4.2 Precision as proportion of D

We have shown that considering the precision of estimates of D as a proportion of the value of D may be a useful way to avoid the loss of precision experienced when D is higher than anticipated in a study. Using a composite sample size made up of both a fixed value of δ or w and a percentage of D means additionally that precision is not reduced when D is lower than planned. This somewhat mitigates the worry about whether an obtained estimate of D is accurate enough, as the precision from the planned study will be at least as good as originally desired regardless of the value of D observed. Although this isn't a traditional way to consider sample size we feel it is quite a practical solution,

especially when the target value (of D, in this case) is not very certain. Using this method means that researchers will know the minimum precision they will have before doing the study, rather than having to do a post hoc calculation if their study comes back with a different D than anticipated.

Chapter 8

Systematic review of D values

8.1 Introduction

In Chapter 6 we looked at the situation where researchers wish to validate a point estimate of D in a new dataset. In this scenario the point estimate of D must be provided in order to calculate sample size; but how are researchers to assign a value to D for their disease area, dataset and model combination if there is no prior dataset to guide them? Although D depends on many things (what data is collected, the number of patients and events, the model used, etc.) and no two situations will be exactly comparable, any realistic estimate is better than none.

In this chapter we describe a review to determine D values in a number of disease areas. In addition to reviewing published values of D, we also consider a transformation from Harrell's c to D. c is widely reported in the literature, and being able to use this quantity as the basis of an estimation for D would provide a large pool of potential D estimates from a wide range of diseases.

The values found, along with values of D calculated in previous chapters, will form the basis of a 'D Library' which could be used by researchers when planning a prognostic study. In this chapter we describe our methods and present initial results from the review. The library itself is presented in more detail in Chapter 9.

8.2 Methods

In previous chapters we have already calculated D and λ for a number of datasets in different diseases; these will form the first part of the library. The bulk of the library will be formed from twin literature searches of D values and c values in published articles. The search strategy and results for these reviews are presented in Sections 8.3 and 8.5. In Section 8.4 an empirical conversion for obtaining D from c is developed.

8.3 Literature search for papers reporting D or R^2_D

8.3.1 Aim

The aim of this literature search was to find published papers which either quote a value of D for a particular dataset and model, or which provide enough other information such that D can be calculated. This search was originally performed in February 2011 and updated in July 2011.

8.3.2 Search strategy

This literature search consisted of a citation search in Web of Science to find all papers citing Royston and Sauerbrei's (2004) original paper on D. All papers were read and any which present D, R^2 (any type) or Harrell's c-index for real data were recorded.

Publications were excluded if they were purely methodological, not in human medicine, not available as full text online or not reporting analyses of time-to-event data.

8.3.3 Results

This section comprises a brief overview of the number of papers found and which quantities were reported. A more thorough discussion of the results is given in Chapter 9.

71 papers were found to have cited the original Royston and Sauerbrei (2004) paper. 23 papers were excluded as purely methodological. Two papers were excluded for not having full text available online. This left 44 papers, which between them describe 142 models. Of these 44 papers, 34 actually reported values of D or R^2_D for a total of 108 models. Table 8.1 shows which relevant quantities were included in these 34 papers in addition to either D or R^2_D. 12 papers out of the 34 (32%) did not report the number of

Quantities reported						Number of papers
D	$SE(D)$	95 % CI for D	R^2_D	Harrell's c	e	
1	0	0	0	1	1	7
1	0	1	0	1	0	6
1	0	1	0	1	1	6
1	0	1	0	0	1	3
1	1	0	0	1	1	2
1	0	0	0	0	0	2
0	0	0	1	1	0	2
1	0	0	0	0	1	2
0	0	0	1	0	0	1
0	0	0	1	1	1	1
1	0	1	1	1	1	1
1	0	1	0	0	0	1

Table 8.1: Quantities reported in the 34 papers reporting D or R^2_D. 0 indicates absence and 1 presence of that quantity

events in the dataset used to develop the model, a proportion similar to the 30% found by Mallett et al. (2010). 10 papers did not actually calculate D or R^2_D for any models despite referencing the Royston and Sauerbrei paper. 8 of these 10 reported Harrell's c, the other two reported neither D nor c.

Of the 34 papers that did report D or R^2_D, 17 reported models predicting death or disease progression events in patients with a particular disease. These 17 comprised 11 in cancer, 2 each in cardiac disease and respiratory disease, 1 in HIV and 1 in neurology. The other 17 papers reported risk models predicting first disease events in healthy patients. Of these 17, 10 looked at cardiac endpoints, three at bone fracture, two at occurrence of diabetes, one at renal disease, and one looked at a variety of endpoints (cardiac, liver, renal disease and cataracts).

Figure 8.1 gives a histogram of the 101 values of D available from the 34 papers. The distribution is roughly Normal, however there is a slight positive skew. The mean is 1.62 and the median is 1.47. The highest value is $D = 3.44$.

Table 8.1 shows that 25 of these papers contained both D or R^2_D and Harrell's c. These papers included a total of 89 models and so could contribute 89 data points to the process of modelling the relationship between c and D.

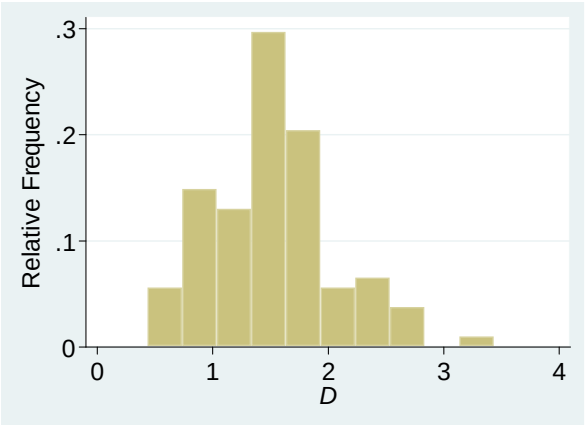

Figure 8.1: Histogram of 101 D values from first literature search

8.4 Converting Harrell's c-index to D

The 34 articles found which report D form a good basis for a library of D values, but being few in number they are limited in practical use; many researchers will not be able to find a suitable estimate of D amongst these 34 papers. Thus we wish to expand the library to cover more disease areas, and also make it as easy as possible for researchers to find a realistic estimate of D themselves, for the situation where there isn't a suitable estimate available in our library. One way to achieve both of these aims is to develop a transformation from a more widely used measure of prognostic value to D.

A suitable target measure of prognostic value is Harrell et al.'s (1984) c-index (see Chapter 2 for more details about this measure). The c-index is widely reported in prognostic studies and should provide a rich source of estimates of D for our sample size calculation, if a reasonable transformation can be found.

We begin our search for a transformation with a method proposed by White (2011) to convert D to c.

8.4.1 Proposed conversion (White)

Underlying the theory of D is the assumption that the linear predictor (prognostic index) of the model is normally distributed; that is $\beta' x \sim N(\mu, \sigma^2)$. This is the key step in a proposed theoretical conversion between D and the c-index, developed by White (2011).

The c-index in time-to-event data is the proportion of pairs where predictions and outcomes are concordant; that is, the patient with lower risk (according to the model) has

a lower hazard and hence survives longer than the patient with higher risk. Equivalently, if we assume no censoring, it is the expected probability that for any randomly chosen pair of observations in the dataset, the patient with lower PI survives longer (PI_2, with survival time T_2) than the patient with higher PI (PI_1, with survival time T_1); that is,

$$c = E\left[P(T_2 > T_1)\right].$$

over all random pairs of patients in the dataset where $PI_1 > PI_2$.

Now let us assume that survival times are exponentially distributed with a constant hazard h, and that we model survival using the Cox proportional hazards model. Let θ be the hazard ratio between the patient with the higher PI (PI_1) and the patient with the lower PI (PI_2). This implies the hazard ratio θ is greater than 1, as risk of an event occurring increases with PI. Thus for a randomly chosen pair, the patient with the higher PI (PI_1) has hazard $h\theta$ and thus survival function $S(t) = \exp(-h\theta t)$; the patient with the lower PI (PI_2) has hazard h and function $S(t) = \exp(-ht)$.

We now use a general result that if random variables $X \sim Exp(\frac{1}{\mu_x})$ and $Y \sim Exp(\frac{1}{\mu_y})$, then

$$\begin{aligned} P(X > Y) &= \frac{\mu_x}{\mu_x + \mu_y} \\ &= \frac{\frac{1}{\mu_y}}{\frac{1}{\mu_y} + \frac{1}{\mu_x}} \end{aligned}$$

(Casella and Berger, 2001). Using this result, for our randomly chosen pair we have $T_1 \sim Exp(h\theta)$ and $T_2 \sim Exp(h)$ and so

$$\begin{aligned} P(T_2 > T_1) &= \frac{h\theta}{h\theta + h} \\ &= \frac{\theta}{\theta + 1}. \end{aligned}$$

We have defined the hazard ratio $\theta = \frac{\exp(PI_1)}{\exp(PI_2)} = \exp(PI_1 - PI_2)$ so

$$P(T_2 > T_1) = \frac{\exp(PI_1 - PI_2)}{\exp(PI_1 - PI_2) + 1}$$

Thus returning to our definition of c, which considers all the possible pairs of patients in the dataset where $PI_1 > PI_2$,

$$\begin{aligned} c &= E\left[P(T_2 > T_1)\right] \\ &= E\left[\frac{\exp(PI_1 - PI_2)}{\exp(PI_1 - PI_2) + 1}\right] \\ &= E\left[logit^{-1}(PI_1 - PI_2)\right] \end{aligned}$$

where $logit^{-1}(x)$ is the inverse logit function $\frac{\exp(x)}{1+\exp(x)}$. We assume that the prognostic index is normally distributed, $PI_1 \sim N(\mu, \sigma^2)$ and $PI_2 \sim N(\mu, \sigma^2)$; thus for two randomly selected patients their difference is also normally distributed: $PI_1 - PI_2 \sim N(0, 2\sigma^2)$.

Thus $E\left[PI_1 - PI_2\right] = E\left[Z\sqrt{2}\sigma\right]$ and so

$$c = E\left[logit^{-1}(Z\sqrt{2}\sigma)\right], \tag{8.1}$$

where $Z \sim N(0, 1)$.

Since $D = \kappa\sigma^*$, where σ^* is an estimate of the standard deviation of the prognostic index values under the assumption of Normality and $\kappa = \sqrt{8/\pi}$ (Royston and Sauerbrei, 2004), we can express σ as $\sigma = \frac{D}{\sqrt{8/\pi}} = \frac{D\sqrt{\pi}}{\sqrt{8}}$. Substituting this into (8.1) we can obtain an equation for c in terms of D:

$$c = E\left[logit^{-1}\left(\frac{ZD\sqrt{\pi}}{2}\right)\right]. \tag{8.2}$$

This expectation can be calculated using Gauss-Hermite quadrature or numerical integration. Thus for a given D we can calculate c.

Table 8.2 gives the transformation of various values of D to c according to equation 8.2, as well as the backwards transform of D to c. Some useful points in this relationship are that when $D = 0.6$, $c = 0.6$ also, and that a D of 1 transforms to a c of 0.66 ($\frac{2}{3}$).

Performance - simulated data

First we assessed the performance of equation 8.2 in simulated survival data with random censoring. This data was simulated using the same methods as outlined in Chapter 5, Section 5.3.2. Datasets of 1000 patients were simulated 2000 times for each β value from 0.2–2.0 in steps of 0.2, and each censoring proportion of 0%, 20%, 40%, 60%, and 80%. We

$D \longrightarrow c$		$D \longrightarrow c$	
0	0.500	2.6	0.801
0.2	0.535	2.8	0.812
0.4	0.570	3.0	0.822
0.6	0.602	3.2	0.831
0.8	0.632	3.4	0.839
1.0	0.659	3.6	0.847
1.2	0.684	3.8	0.854
1.4	0.707	4.0	0.860
1.6	0.727	4.2	0.866
1.8	0.745	4.4	0.872
2.0	0.761	4.6	0.877
2.2	0.776	4.8	0.882
2.4	0.789	5.0	0.886

$c \longrightarrow D$		$c \longrightarrow D$	
0.50	0	0.76	1.99
0.52	0.11	0.78	2.26
0.54	0.23	0.80	2.58
0.56	0.35	0.82	2.97
0.58	0.46	0.84	3.43
0.60	0.60	0.86	4.00
0.62	0.72	0.88	4.73
0.64	0.86	0.90	5.71
0.66	1.00	0.92	7.07
0.68	1.16	0.94	9.07
0.70	1.34	0.96	12.15
0.72	1.53	0.98	17.56
0.74	1.74	1.00	∞

Table 8.2: Conversion tables for c, D using White's transformation

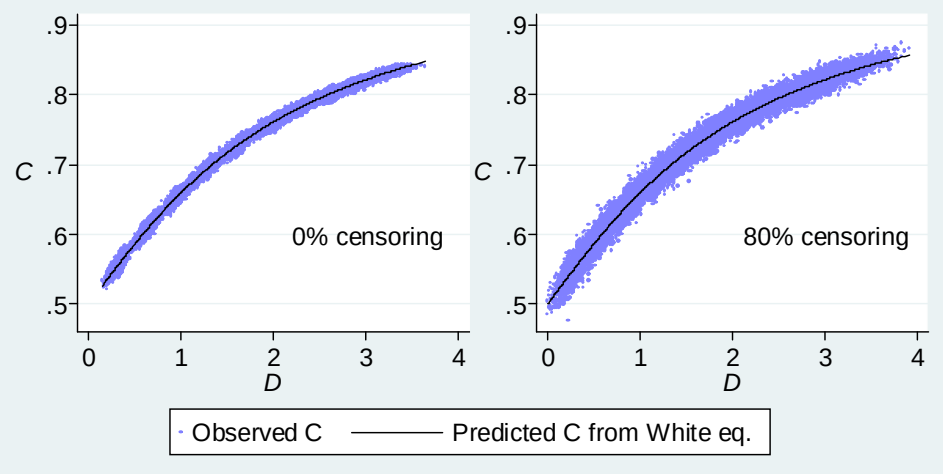

Figure 8.2: Predicted c from equation 8.2 vs D overlaid with observed c from simulated datasets, 0% and random 80% censoring

expected that the transformation may not work well with censored data, since its development included the assumption that data was not censored. However, we found that the magnitude of c did not change with level of censoring and the prediction worked very well for all β and censoring proportions, as shown in Figure 8.2, which is a plot of observed and predicted c separately for 0% and 80% censoring.

We repeated this study with purely administrative censoring, which is where a potential censoring time is fixed and known for all patients in advance (for example, when a date is fixed for data freeze for final study analysis; all patients who have not failed by this time will be censored on that date). Gonen and Heller (2005) found that c increased slightly – up to 2 or 3 percent – with increasing administrative censoring. To simulate

datasets of patients with purely administrative censoring we followed the procedure described in Chapter 4, Section 4.7.4:

1. Simulate N survival times (T_s) by following the procedure in 3.3.2.

2. Calculate a censoring time T_c for the nth record as follows:

$$T_c = \frac{n}{N} r + f$$

   where r was the length of the study recruitment period and f the study follow up period. This assumes that entry of patients to the trial was uniformly staggered over the recruitment period, and that all patients were censored at the end of the follow up period if they had not failed by this time.

3. Records where $T_c < T_s$ were considered censored at time T_c; records where $T_s < T_c$ were considered failures at time T_s. No other censoring was performed.

Once again r and f were set to 2 and 4 years and datasets of size $N = 2000$ were generated. Steps (1) to (3) were repeated 1000 times for each β value of 0.2–2.0 in steps of 0.2, and censoring levels of 40% and 80%. The desired censoring proportion was achieved by changing the baseline hazard h.

With administrative censoring we found that there was an effect on c with censoring: c increased with increasing censoring and the effect was stronger the higher β was. Thus c was under-predicted when $D \geq 1$ ($\beta \geq 0.63$) and the censoring proportion was high. Figure 8.3 shows the observed and predicted c vs D for 0% censoring and 80% administrative censoring.

Performance - real data

Finding that the transformation generally worked well with some simulated data we moved on to assessed how well it worked with real data. Figure 8.4 plots the conversion, and overlays the D and c from the 26 datasets used in chapter 9 (calculated using a MFP model with $\alpha = 0.05$), and also those from the 89 articles from the literature search which reported both D (or R^2_D) and c. Figure 8.4 seems to show that the conversion underestimates c when $D \geq 1.5$ approximately.

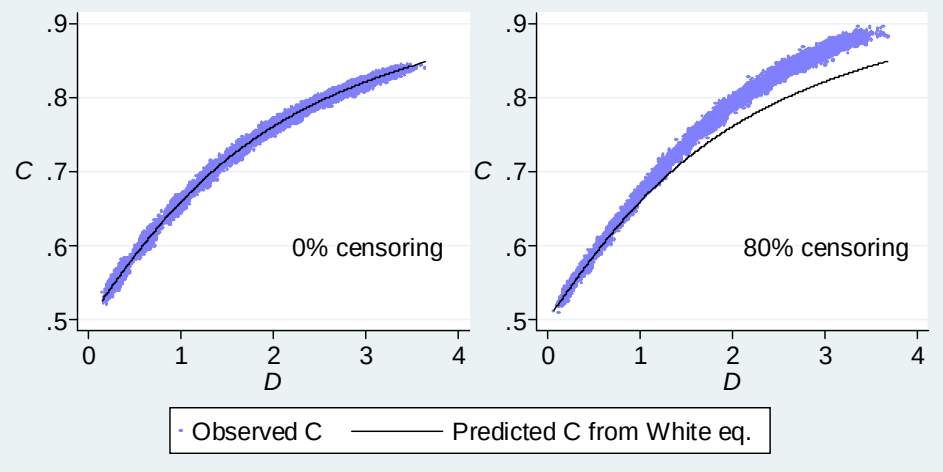

Figure 8.3: Predicted c from equation 8.2 vs D overlaid with observed c from simulated datasets, 0% and administrative 80% censoring

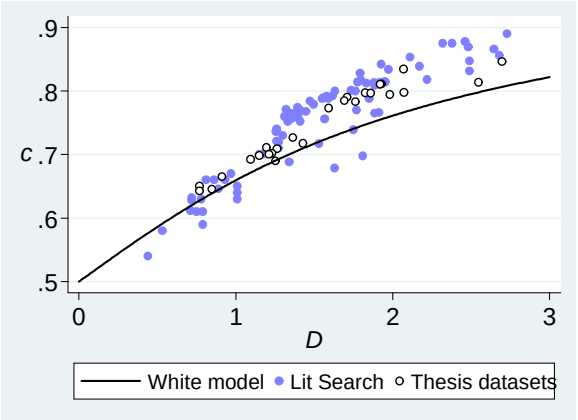

Figure 8.4: Plot of c vs D, using White's conversion, overlaid with D and c from articles found in literature search (blue dots); and from datasets used throughout the thesis (circle).

In summary, we found that equation 8.2 did not predict c well for real data and indeed the results from the real data generally showed even higher values of c than those in the simulated data results with 80% administrative censoring (Figure 8.5). This discrepancy is likely to be due in part to censoring – as we noted that administrative censoring affects the value of c – however the fact that the under-prediction of c is worse in the real datasets than in the simulated data implies there are other forces at play. We speculate that the additional discrepancy may be due to non-normality of the prognostic index but we have not investigated this.

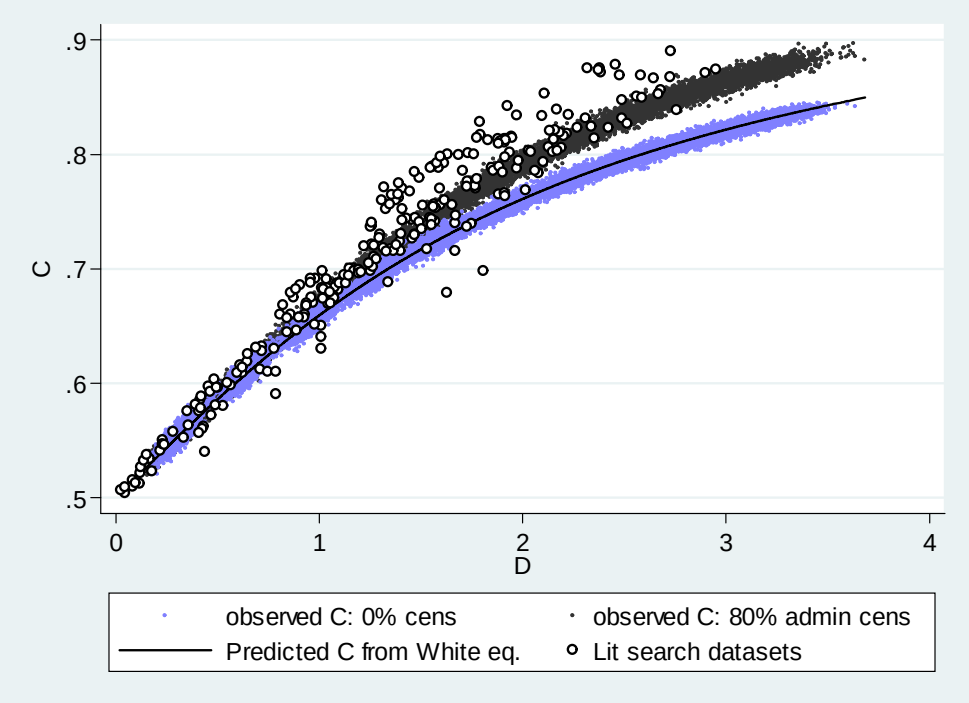

Figure 8.5: Predicted c from equation 8.2 vs D (line) overlaid with observed c from simulated datasets with 0% (blue dots) and 75% administrative (black dots) censoring; plus D from 89 literature search & thesis datasets (circles).

8.4.2 Empirical relationship between D and c

As this theoretical relationship does not seem to work sufficiently well in practice, an empirical relationship was sought to help us predict D from c.

Source of data

To strengthen the relationship additional datapoints were added to the graph in Figure 8.4 above. These were obtained from the Surveillance, Epidemiology, and End Results Program datasets for (1) leukemia, myeloma and lymphoma, (2) colorectal cancers, (3) urinary cancers and (4) female genital cancers (SEER, 2000). These datasets were further broken down into particular disease areas, and some further into the 9 geographical registry areas of SEER, to give us a total of 179 data points. SEER datasets all utilise the same common variable set, so the datasets were inspected and potentially useful variables selected manually for each disease area. The same `mfp` command line was used for each disease area, but models were selected separately for each registry within that disease area. An α for selection of 0.05 was used for most datasets; although after inspection of results, the process was repeated using $\alpha = 0.50$ for some of the smaller datasets, purely

to get a model with higher D and 'fill in' as much of the likely D range as possible. Further details on the `mfp` command lines used for each disease area are given in Appendix B. In each case, the specifics of the final model and whether or not it was a good prognostic model for the disease were unimportant for this exercise; all that was required was a value of D and a value of c for one model.

Optimism was not calculated for any of the estimates of D or c from the thesis datasets, since most of the papers reporting D and c did not appear to have attempted to either adjust for optimism or externally validate the models developed. Additionally, adjusting for optimism may lead to the problem of negative values of D which do not really have an equivalent on the c scale, since $c < 0.5$ implies that the model does have predictive value; just in the opposite direction to that expected.

After the values were obtained and plotted it became clear that all the values of D were greater than 0.5, which meant there was a gap in the graph for $0 < D < 0.5$. In order to ensure that the empirical model developed from this was accurate across the full range of D, one SEER disease area (with all 9 registries) was selected from each of the 4 main cancer areas above and used to develop 'poor' models (again, described in Appendix B). The list of potentially included variables for these poor models was short; only including a few demographic variables and no disease-related variables, to ensure that the resulting D was very low.

The final spread of 294 data points from the three sources (179 from SEER, 89 from literature search, 26 from thesis datasets) is shown in Figure 8.6. The highest values of D obtained were just under $D = 3$. Higher values would have been desirable but are rarely seen in practice.

Model development

Using these 294 points, D was regressed on c using fractional polynomials. The regression line was forced through the point $D = 0,\ c = 0.5$, which corresponds to a model which has no predictive value. A clustering variable was used to identify models evaluated on the same dataset, to allow cluster-robust variance estimation. The resulting model found that an FP2 model with powers (1,3) provided the best fit. The final model was

$$D = 5.48(c - 0.5) + 10.59(c - 0.5)^3. \tag{8.3}$$

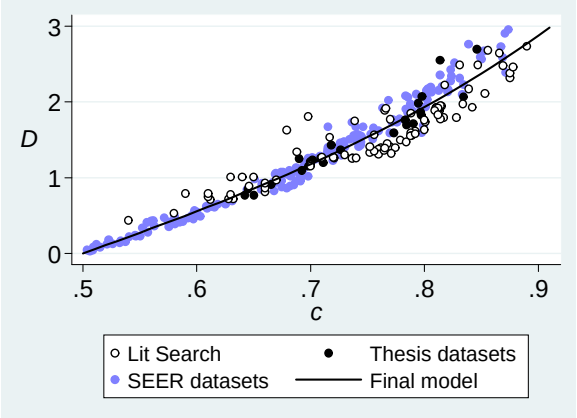

Figure 8.6: Plot of 294 datapoints used to create model for predicting D from Harrell's c: from literature search (circles), thesis datasets (black) and SEER datasets (blue); overlaid with final model line

Figure 8.6 shows this model overlaid over the data used to develop it.

The model seems a good fit for this data but for higher values of c it does not perform as well. When $c = 1$, $D = \infty$; so the upper tail of the line in the figure should perhaps be rising more steeply as it passes $c = 0.9$. However, it is unlikely that values of D greater than 3 would be seen in a real dataset, so we should not be too concerned about this.

Another issue relates to the effect of administrative censoring on c reported by Gonen and Heller (2005). It could be argued that there should be a term included in the model to account for such censoring, however even with high censoring the largest difference observed in c by Gonen and Heller was not more than 3%. As we can see from the spread of data in Figure 8.6, our conversion is not an exact method and should not be considered as such, thus any effect of censoring is small enough to be ignorable here.

The fact that we did not adjust any of the thesis dataset D and c estimates for optimism here – and most of the D and c estimates from the literature search were also not adjusted – should also be considered. It is possible that if optimism-adjusted D and c values were used to develop the transformation, the curve in Figure 8.6 would look different, however we cannot speculate on this as we have not looked in any detail at the optimism present in c. It is possible that the proportion of optimism in c estimated from a model would be the same as in D from the same model, and thus the transformation would still hold. The lack of adjustment for optimism in the development of the transformation does not concern us too greatly, but researchers should perhaps be wary of using it for transforming values

of c from very small studies, or from studies where serious overfitting is expected for whatever reason.

Formula 8.3 can be used to calculate D from any value of c. Table 8.3 transforms given values of c to D and also on to R^2_D (up to $c = 0.90$, to avoid extrapolating beyond the data used to develop the conversion). This table also presents the reverse transformation, D to c.

$c \longrightarrow$	$D \longrightarrow$	R^2_D	$c \longrightarrow$	$D \longrightarrow$	R^2_D
0.50	0.00	0.00	0.71	1.25	0.27
0.51	0.05	0.00	0.72	1.32	0.29
0.52	0.11	0.00	0.73	1.39	0.32
0.53	0.16	0.01	0.74	1.46	0.34
0.54	0.22	0.01	0.75	1.54	0.36
0.55	0.28	0.02	0.76	1.61	0.38
0.56	0.33	0.03	0.77	1.69	0.40
0.57	0.39	0.03	0.78	1.77	0.43
0.58	0.44	0.04	0.79	1.85	0.45
0.59	0.50	0.06	0.80	1.93	0.47
0.60	0.56	0.07	0.81	2.01	0.49
0.61	0.62	0.08	0.82	2.10	0.51
0.62	0.68	0.10	0.83	2.19	0.53
0.63	0.74	0.11	0.84	2.28	0.55
0.64	0.80	0.13	0.85	2.37	0.57
0.65	0.86	0.15	0.86	2.47	0.59
0.66	0.92	0.17	0.87	2.56	0.61
0.67	0.98	0.19	0.88	2.66	0.63
0.68	1.05	0.21	0.89	2.77	0.65
0.69	1.11	0.23	0.90	2.87	0.66
0.70	1.18	0.25			

$D \longrightarrow$	c	$D \longrightarrow$	c
0.0	0.50	1.6	0.759
0.1	0.518	1.7	0.772
0.2	0.536	1.8	0.784
0.3	0.556	1.9	0.796
0.4	0.572	2.0	0.808
0.5	0.590	2.1	0.820
0.6	0.607	2.2	0.831
0.7	0.624	2.3	0.842
0.8	0.641	2.4	0.853
0.9	0.657	2.5	0.863
1.0	0.673	2.6	0.874
1.1	0.688	2.7	0.884
1.2	0.703	2.8	0.893
1.3	0.717	2.9	0.903
1.4	0.732	3.0	0.912
1.5	0.745		

Table 8.3: Conversion tables for c to D to R^2_D, and also from D to c, using empirical transformation equation 8.3

8.5 Literature search for papers reporting *c*

8.5.1 Aim

In the previous section we developed a transformation tool which can be used to obtain an approximate value of *D* from a given value of *c*. We now aim to collate some values of *c* which we can convert to *D* in order to widen the disease areas covered by our *D* library.

8.5.2 Search strategy

For this search, we performed a citation search in Web of Science to find all papers citing Harrell et al.'s (1996) paper which is often cited when papers use the *c*-index (other papers may also be referenced, for example Harrell et al., 1984). Only publications from 2010 were included. Papers were excluded if they are purely methodological, not in human medicine, not available as full text online or not reporting analyses of time-to-event data. This search was performed in April 2011.

8.5.3 Results

Again, this section includes just a brief overview of the number of papers found. A more thorough discussion of the results is performed in Chapter 9.

The search resulted in 207 papers from 2010, and after excluding methodological papers and those not in human medicine 175 remained. Scanning abstracts to exclude any papers which clearly did not use survival analysis, we were left with 114. Where use of survival analysis could not be determined from the abstract, the full methods section was read; this resulted in a further 23 papers being excluded, leaving 91. Finally, the results section of the remaining 91 papers were read, which left a body of 77 papers which actually reported Harrell's *c* for a model; in total 331 models were reported by these 77 papers.

Of these 77 papers, 60 described prediction of disease events in patients who had a particular disease (true prognostic models), while 17 predicted onset of disease in healthy patients (risk models). Of the 60 predicting disease events, 47 were in cancer, 9 were in cardiac diseases, two in liver diseases, and additionally two in other diseases: Chagas disease and Raynaud's phenomenon. Of the 17 papers predicting disease in healthy patients, 15 considered cardiac endpoints, one cancer and one dental caries.

c	$SE(c)$	95 % CI for c	e	Number of papers
1	0	0	0	24
1	0	1	0	8
1	0	0	1	29
1	1	1	0	1
1	1	0	1	3
1	0	1	1	12

Table 8.4: Quantities reported in the 77 papers reporting c. 0 indicates absence and 1 presence of that quantity

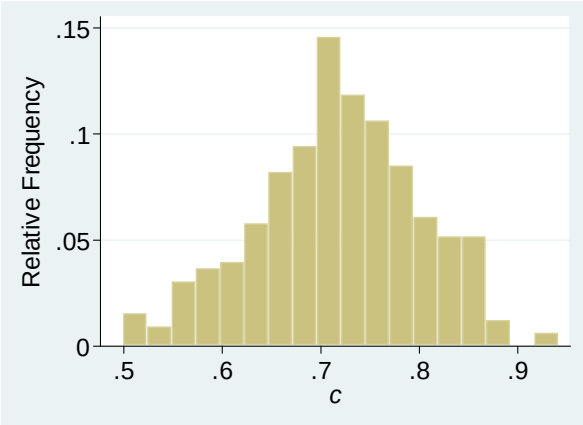

Figure 8.7: Histogram of 331 values of c obtained from second literature search.

Figure 8.7 gives a histogram of the 331 values of c given by the papers. They are approximately normally distributed, ranging from 0.50 to 0.94 and with mean and median both 0.72.

8.6 Literature searches: summary of combined results

Once the values of c were converted to D we had 480 values of D in total, across all papers and models within papers. This comprised 140 from the first literature search (108 of which were D values; an additional 32 were values of c converted to D), 314 from the second literature search (314 values of c converted to D) and 26 thesis datasets (26 D values). Note that the 190 models from SEER datasets used to develop the $c \rightarrow D$ transformation are not included in the library. Figure 8.8 shows a histogram of these 480 values. The values range from 0 – 3.44, their mean is 1.40 (median 1.38), and standard deviation 0.584. Since D has the interpretation of the log hazard between the two groups

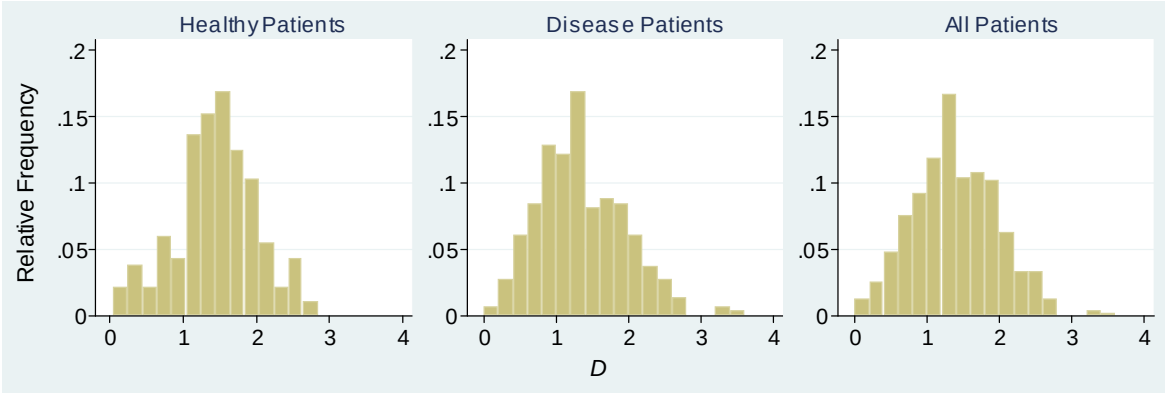

Figure 8.8: Histograms of values of D for (left) 184 models in healthy patients, (middle) 296 models in patients with disease, (right) all 480 models

formed by dichotomising the model prognostic index at its median, we can translate a D of 1.4 to a hazard ratio of $e^{1.4} = 4$.

8.6.1 Risk models in healthy subjects

These models are developed to predict onset or incidence of some disease or condition in apparently healthy subjects. Although risk models are not technically prognostic models, they are still aiming to predict an event and it is still of primary interest to researchers to measure and compare their predictive value. As we anticipate that our sample size calculations could be used with risk models just as easily as with prognostic models, we included them in our literature search. Figure 8.8 shows a histogram of the 184 values of D from such models. These values have a minimum of 0.05 and maximum of 2.73. Their mean and median are both 1.47 and standard deviation is 0.550.

8.6.2 Prognostic models in patients with disease

This section covers papers which develop and / or validate models in patients with a disease, and attempt to predict some disease-related event, or death. Figure 8.8 shows a histogram of the 296 values of D from such models. These values have a minimum of 0 and maximum of 3.44. They have mean 1.30 and median 1.35 – so somewhat lower than the models in healthy patients – and their standard deviation is 0.601.

8.7 Discussion

In this chapter we have outlined methodology used to collect values of D, through literature searches and a model developed to convert Harrell's c to Royston and Sauerbrei's D. The model developed is based on fractional polynomials and was used to convert the c values found, which gave us a final total of 480 values of D. We have presented histograms for these values altogether and split by whether the models were based in healthy patients or patients with disease; and found that the former show slightly higher values of D.

In the next chapter we will summarise and comment on this collection of D values, and describe some values of particular interest. A full narrative description on the papers found in the literature searches can be found in Appendix E, where D values are presented by disease and endpoint to form a basic library.

Chapter 9

D Library

9.1 Introduction

In the previous chapter we outlined the methods used to perform two literature searches for values of Harrell's c and Royston & Sauerbrei's D, and developed a model to predict D from c. The full 'library' of D values resulting from the two literature searches is described in Appendix E. In this chapter we present a broad overview of the findings: describing the disease areas covered by the search and highlighting some results of interest. We present first the results from risk models predicting first disease events in healthy patients, and then the results from prognostic models predicting death or disease progression events in patients who already have a particular disease. Here, as in Appendix E, we do not differentiate between D values which were originally reported in the paper and values which were derived from Harrell's c using the transformation developed in Chapter 8. For brevity, where relevant we use the notation D_F to indicate a D value for a female-only group and D_M for male-only.

For full references for all papers, see the D Library bibliography at the end of the thesis.

9.2 Risk models in healthy subjects

These models are developed to predict onset or incidence of some disease or condition in apparently healthy subjects. The majority of such papers found in our search concerned

the prediction of cardiovascular (CV) events of various types, but a few other diseases were also included.

9.2.1 Incidence of cardiovascular disease

Most of these papers considered the endpoint of cardiovascular disease (CVD), generally defined as coronary heart disease (CHD) plus cerebrovascular disease, that is: myocardial infarction, coronary heart disease, stroke, and transient ischaemic attacks. Most such papers presented separate results for women and men and interestingly all found that their models predicted better in women than in men. A majority of papers used established risk prediction scores or models based on such scores, and as a result D values across these papers were reasonably similar in magnitude. For example, Hippisley-Cox et al. (2007) derived the QRISK score for predicting CVD, finding $D_F = 1.55$ and $D_M = 1.45$ ($R^2_D = 36\%$ and 33%) in the validation cohort. Later the score was externally validated, with $D_F = 1.56$ (37%) and $D_M = 1.39$ (32%) (Collins and Altman, 2009; Hippisley-Cox et al., 2008a). The successor to this score, QRISK2, was also developed by Hippisley-Cox et al. (2008b), reporting $D_F = 1.80$ (44%) and $D_M = 1.62$ (39%) in the validation cohort; the score was independently validated by Collins and Altman (2010) who found $D_F = 1.66$ (40%) and $D_M = 1.45$ (33%). Yet another QRISK score, this time based on lifetime risk of CVD, was developed by the same group, finding $D_F = 1.93$ (47%) and $D_M = 1.79$ (43%) (Hippisley-Cox et al., 2010).

Several papers sought to predict CHD events alone, D values were similar to those for CVD and again the pattern was seen of higher values of D in women than men.

Figure 9.1 shows graphically the values of D for models predicting CVD or CHD events. The dots correspond to the mean of the D values found for that endpoint; the lines are the 95% CI for the D values based only on the number of studies included, not the number of patients or events within studies as this information was not consistently reported.

Other cardiovascular endpoints covered in the search included cardiac death, heart failure, coronary heart disease death, and stroke. The majority of these papers found D_F in the range 1.3-2.0 ($R^2_D = 29\%$–49%) and D_M in the range 1.2–1.8 (26%–44%), similar to the values seen for models predicting CVD and CHD events.

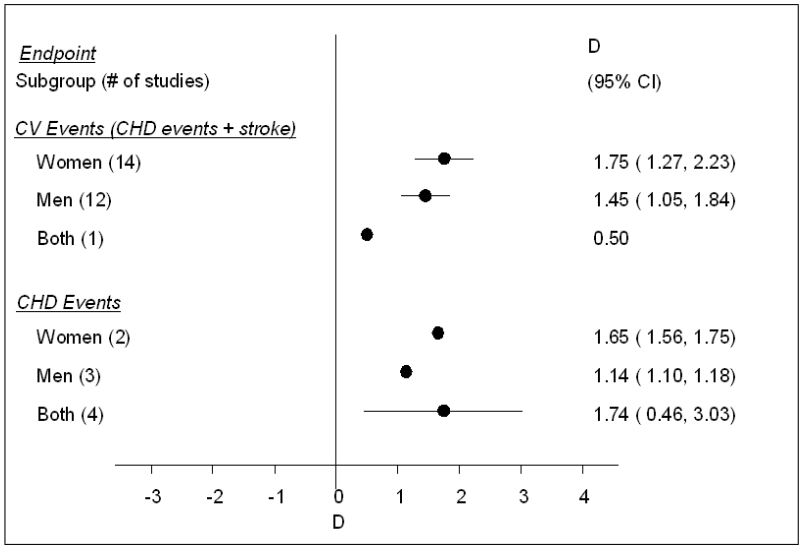

Figure 9.1: Forest-type plot of D for CV and CHD events in healthy subjects.

One paper looking at the performance of the Framingham Risk Score (D'Agostino et al., 2008) in hypertensive patients found very low D values for a variety of cardiac endpoints: in particular $D = 0.39$ for CHD events ($R^2_D = 4\%$) (Nelson et al., 2010). The authors concluded that this risk score discriminated poorly in hypertensive patients, and this illustrates an important point that we will elaborate later: the less diverse the population, the lower D is likely to be.

Figure 9.2 shows graphically the values of D for models predicting some of these other cardiac endpoints in healthy patients.

9.2.2 Incidence of other diseases

Other papers found included risk models for diabetes, bone fracture, chronic kidney disease, predicting specific side effects of statins, colorectal cancer, and dental caries. The models predicting diabetes and bone fracture tended to show quite high values of D, of around 1.7–2.5 ($R^2_D = 41\%$–60%) and 1.6-2.7 (38%–64%) respectively.

The single model predicting chronic kidney disease, which was based on a large general practice dataset, also found high values, $D_F = 2.32$ (56%) and $D_M = 2.38$ (57%) (Hippisley-Cox and Coupland, 2010b).

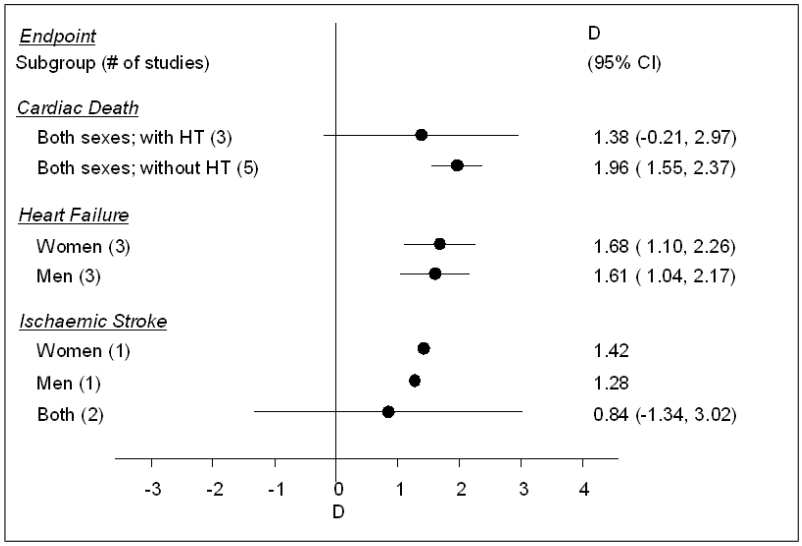

Figure 9.2: Forest-type plot of D for other cardiac events in healthy subjects. HT=hypertension

9.3 Prognostic models

This section covers papers which developed and / or validated models in patients with a disease, attempting to predict some disease-related event, or death; these are 'true' prognostic models.

9.3.1 Cardiovascular disease

13 papers were found which considered survival models in patients after cardiovascular (CV) disease events including heart failure, stroke, and MI; or after CV interventions such as coronary artery bypass graft and percutaneous coronary intervention. Most D values in this category were in the range 1.4–2.1 ($R^2_D = 32\%$–51%) with a cluster around 1.7–1.8 (41%–44%).

Figure 9.3 shows graphically the magnitude of D in a selection of the studies predicting CV events in patients with pre-existing conditions. Again, the dots correspond to the mean of the D values found for that endpoint and disease, and the lines are 95% CI.

9.3.2 Cancer

Most of the papers in our literature search were in cancer; 99 in total. We divide discussion by cancer site, and sometimes further by early and advanced disease or other characteristics where appropriate.

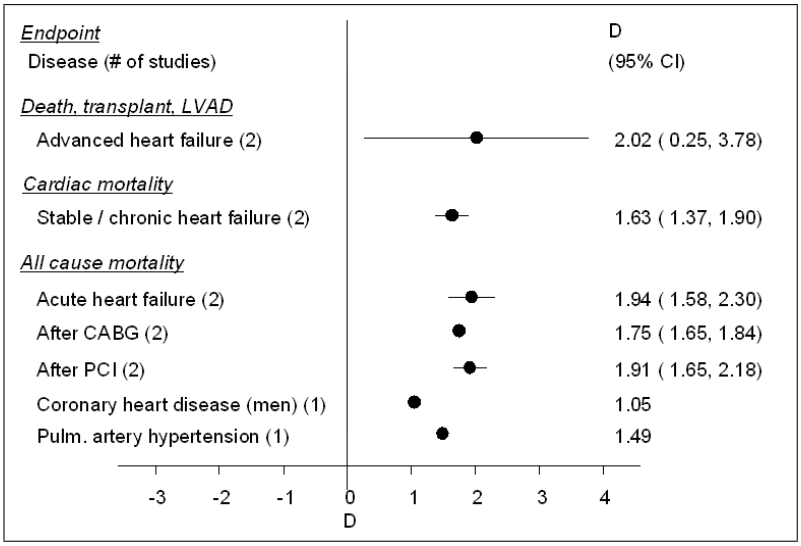

Figure 9.3: Forest-type plot of D for CV events in patients with existing CV condition. LVAD=left ventricle assist device, CABG=coronary arterial bypass graft, PCI=percutaneous coronary intervention

Breast cancer

Several papers considered prediction models for breast cancer, and 11 of the datasets used throughout this thesis were breast cancer datasets. The different patient groups (for example defined by stage of disease or hormone-receptor status) and outcome events used (including overall survival, cancer-specific survival, recurrence-free survival) makes it difficult to compare D values across models. One place where it is more appropriate to compare models, datasets and D values is the 9 SEER datasets used throughout the thesis. These datasets contain exactly the same information on breast cancer patients in 9 different geographical areas of the USA and as such the patients within each dataset should (arguably) be quite similar. The best models chosen within these datasets with an MFP procedure with $p = 0.05$ had D values ranging from 1.59–2.07, corresponding to R^2_D of 38%–51%. This implies that the 9 datasets each have subtly different relationships between variables and outcome.

Prostate cancer

The eight papers reporting on prostate cancer patients fell into various subcategories. The most common category consisted of papers seeking to predict biochemical recurrence in patients with organ-confined disease, who had been treated with radical prostatectomy

(RP). The values of D found in these papers varied fairly widely, from 1.2–2.1 ($R^2_D = 26\%$–51%) but due to differences in study follow up and the predictors used in the models, again it is difficult to compare them directly.

Another example of more heterogenous patient groups giving higher D values is provided in Cao et al. (2010). This paper tested a model in a group of prostate cancer patients with a Gleason score of 6–9, which gave $D_M = 1.38$ (31%); while the same model in the subset of patients with a Gleason score of 7 gave $D_M = 1.20$ (26%).

One other example of note is a paper on prostate cancer patients after radiation monotherapy treatment (Williams et al., 2006). As well as developing a new model to predict biochemical recurrence, the authors also sought to evaluate an existing model on two independent datasets, with 864 and 271 events respectively. The model showed $D_M = 1.50$ ($R^2_D = 35\%$) in the smaller dataset and $D_M = 1.02$ (20%) in the larger, showing that a validated model can produce quite different estimates of D when applied to different datasets. We will discuss some possible reasons for this and the impact it may have on use of our sample size calculations at the end of this chapter.

Renal cancer

Ignoring the different pathology types, the distribution of D values amongst papers in renal cancer appeared to be roughly bimodal. Values from models predicting overall survival were quite low, around $D = 1$ ($R^2_D = 19\%$), whereas models predicting disease specific survival were much higher, with 5 of 6 such values greater than 2 ($R^2_D = 49\%$).

Other cancers found in the literature search included bladder cancer (where there was no clear difference in D between models in patients with low and higher grade tumours) and liver cancer (two papers each tested three models and none exceeded $D = 1$, $R^2_D = 19\%$). Models in lung cancer were also rather low with $D < 1$ for all models across three papers, each considering different pathologies.

Figure 9.4 shows graphically the magnitude of D in a selection of the above cancer areas, with the endpoint of overall survival. Figure 9.5 shows D values for the endpoint cancer-specific survival, and Figure 9.6 shows D values for the endpoint progression-free survival.

Several papers were also found within the broad descriptions of gynaecological, head and neck, stomach and 'advanced' cancers, however the precise tumour types covered

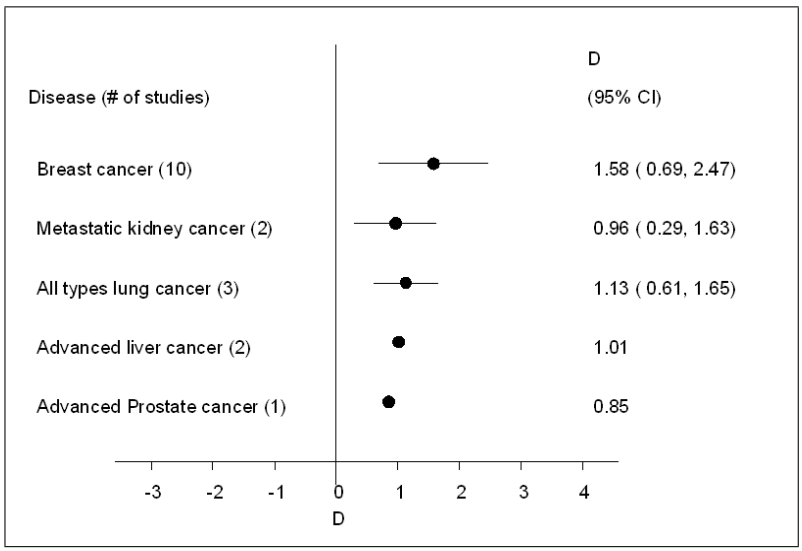

Figure 9.4: Forest-type plot of D for cancer papers with endpoint overall survival

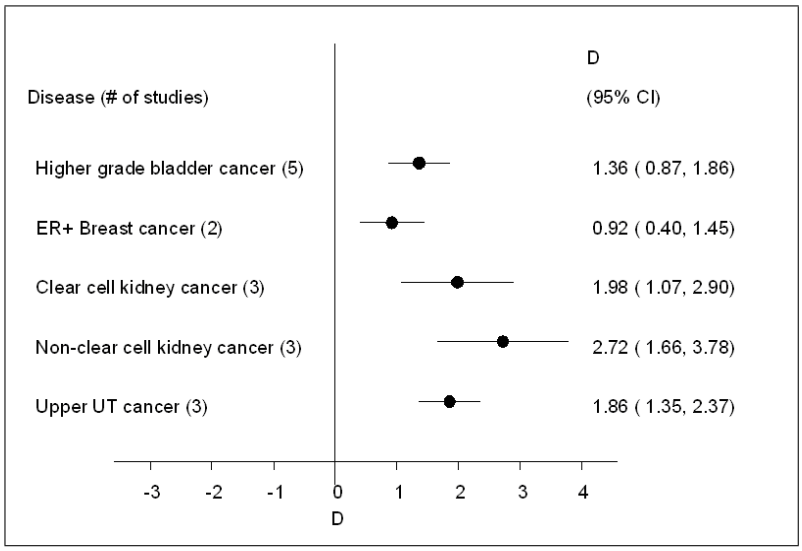

Figure 9.5: Forest-type plot of D for cancer papers with endpoint cancer specific survival. ER+ = estrogen-receptor positive, UT = urinary tract

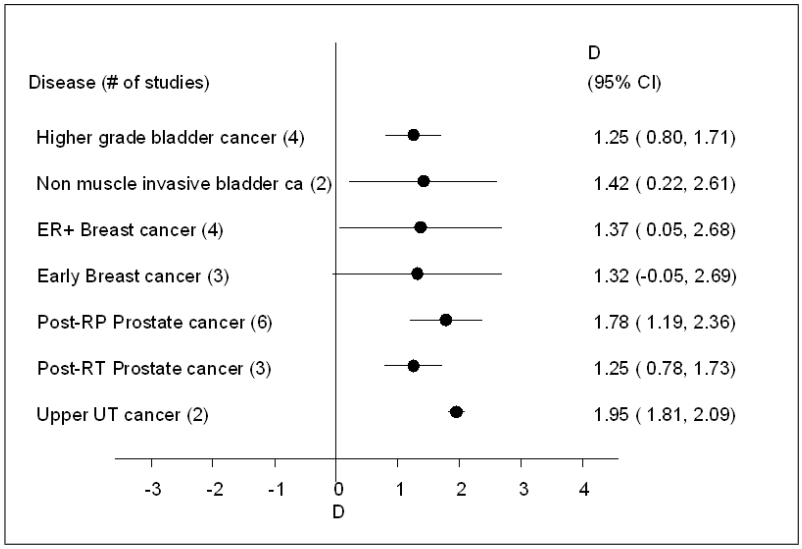

Figure 9.6: Forest-type plot of D for cancer papers with endpoint progression-free survival. ER+ = estrogen-receptor positive, RP = radical prostatectomy, RT = radiotherapy, UT = urinary tract

within these categories were often disparate. There were several papers reporting on outcomes in different classes of leukemia and lymphoma, and a handful of single papers in other cancers were also found. One omission which we felt was notable due to its frequency in the population, was colorectal cancer, for which only one paper returned in our search. Researchers requiring a value of D in any of these disease sites would likely need to do further literature searches in order to obtain a reliable value for their work.

9.3.3 Other diseases

Amongst the papers reporting prognostic models in areas other than cardiovascular disease and cancer, only HIV was seen more than once in our searches. Two of the datasets used throughout the thesis included patients with primary biliary cirrhosis. Our best models for these two datasets showed rather high D values of 2.7 ($R^2_D = 64\%$) and 2.55 (61%), however they were quite small in size.

9.4 Discussion

In this chapter we have described the results of literature searches performed to obtain values of D for prognostic models across a variety of disease areas, as well as for risk models predicting disease in healthy subjects. The primary aim of this literature search was to provide values of D for use with the sample sizes Sig-2 and CI-2 developed in Chapter 6, as these calculations require point estimates of D. Although many diseases had no representation in our searches, we envisage that it will be relatively easy for researchers to follow the same process we followed to obtain approximate D values for use in planning studies in their disease area. A literature search can be conducted to find suitable papers in the disease area of interest which report c values, and the c to D transformation given in Chapter 8 used to calculate D.

9.4.1 Differences in D within the same disease area

In some disease areas covered in this chapter, the same models were used with different datasets. On the whole, the larger the datasets used, the better the consistency in observed values of D from different papers. For example, the cardiovascular risk models QRISK and QRISK2 gave very similar values of D across independent datasets, since very large general practice databases (tens of thousands of events) were used both for their development and validation. For smaller datasets the values of D observed from the same model could be quite different across studies, as in the prostate cancer example described above (which included hundreds of events, so still reasonable sized by research standards). Such differences in D could be a result of the model not having been properly validated or calibrated, however it could also be due to datasets being different in important ways, some of which we discuss below. Observational studies in particular may show differences in the apparent predictive ability of the same model, due to their uncontrolled nature (Royston et al., 2004).

This heterogeneity in observed D values from the same model, let alone different models in the same disease area, highlights the difficulty of finding a single 'best' value of D for use in the sample size calculations. This leads us to recommend that a range of possible D values (and censoring proportions) are input into the calculations, to obtain a range of possible sample sizes. Alternatively, if there is a limit on the number of pa-

tients that can be included in the planned study, the calculations can be used to work out the power that will be obtained from various possible combinations of D and censoring proportion. We outlined in Chapters 5 and 6 the effect of the final D estimate in the new study being different to the value used in the sample size calculation; a higher D than planned means lower precision, a lower D than planned means higher precision.

It is unlikely that a study will be found in the literature which provides an excellent match to the planned study in all aspects. Researchers should read papers carefully and bear in mind the ways in which the planned study is likely to differ from the previous study found, and consider how they might affect D. Some aspects which may differ between studies which appear similar at first glance include whether the study's primary aim was model development or validation, the size of the study, the proportion of censored records, the model(s) used, and the case mix of the patient population.

Study type: validation or development It is important to consider whether a published study is primarily a model development or model validation study, as a study aiming to develop a new model is likely to report optimistic (over-estimated) measures of prognostic value if these estimates have not been externally validated. Even when a study reports both development and validation of a new model, the quality of this validation should be carefully considered, as some commonly used methods do not use completely independent data and so may still produce optimistic estimates. Examples of such methods are internal validation techniques such as data splitting, where the whole dataset is split into two parts: one for model development and one for validation; and temporal validation, where a later cohort of patients from the same institution is used to validate the model.

On the other hand, a study with the primary aim of validating an existing prognostic model on external data should produce estimates of prognostic value that can be considered free of optimism, assuming the validation cohort is truly independent of the development data. The study methods and data sources should be inspected to determine whether this is the case.

If it is suspected that the chosen estimate of D contains optimism, lower values of D should also be input into the calculation to see what difference this makes to sample size.

Study size Even before they have performed their sample size calculation, researchers will have some idea of the scale of study they wish to conduct. A study from the literature may be much smaller or much larger than this. If an estimate of D from the literature is based on a small dataset and has not been validated, it is likely to contain optimism; that is, the value of D is likely to have been overestimated. This shouldn't be ignored, as in Chapter 3 we saw that it is possible for optimism in estimates of D to make up 50% or more of the estimated value of D, when datasets have 100 or fewer events.

Censoring proportion Increased administrative censoring increases the value of c (and hence D output from the conversion) but only slightly – a few percent at most – and so we feel this can safely be ignored, since we are dealing in approximations here. Censoring proportion is a term in the sample size calculation so can be adjusted for and a range of possible proportions used if desired.

Different models In the validation situation (where the model to be used in the new study is already known) ideally a previous study utilising this model will be available for estimation of D. If not, an estimate of D from a similar model should be used, bearing in mind the subject specific knowledge of the researchers. If, for example, an important factor has been omitted from the 'similar' model which will be included in the new model, then researchers may decide that a slightly higher estimate of D is appropriate.

If the purpose of the study is to develop a new model, then obviously it is impossible to know what D will be for this model (as that is why the study is planned!). Here we suggest researchers choose an estimate of D after considering 'good' models in their disease area, bearing in mind the data they plan to collect on patients.

Another point to consider is how many factors are included in the model compared to events in the dataset. If the ratio of events to variables is lower than 20, and certainly lower than 10, again overfitting is quite likely and so the value of D reported may be optimistic. It is worth remembering, however, that such EPV calculations should ideally include the count of all candidate variables originally considered for inclusion in the model. Published papers are unlikely to give this information, so the EPV calculated from variables in the published model should be considered as an upper bound.

Case mix The studies found in the literature may be based on different case mixes to the new study. This may affect D in several ways. Firstly, the heterogeneity of the dataset impacts D. It is likely that a study with a wider patient demographic will output a higher D value than the same model used in a narrower group, simply because the wider group has more factors which can be used to differentiate the patients. This occurs due to the broader case mix resulting in a higher variance for the model PI, which in turn results in a higher D. This phenomenon is common to all measures of prognostic value, not just D.

It is difficult to know how to deal with differences in case mix between studies found in the literature and the proposed new study. Of course the closest match should be chosen; again it may be useful to calculate a range of sample sizes based on a range of D.

9.4.2 Other uses of D and the D library

The values of D collated here have other uses beyond our sample size calculations, and D in general can tell us more than simply the prognostic value of one model in one dataset.

Values of D from the same (validated) model fitted to various datasets can tell us something about how similar the datasets are. There may be obvious differences between datasets, for example in terms of the distribution of patient characteristics or observed outcomes, but if these aspects are similar in two datasets while the D values differ, there must be different relationships between variables and outcomes at work. This could be of use if researchers wish to ensure two datasets are similar; for example if a model is to be developed in one dataset and validated on another, fitting an existing model to both datasets may help determine if there are important differences between the datasets which could affect the planned validation process. In this situation it would be important to ensure that the model used for comparison is already validated, as otherwise optimism could affect its performance.

Looking at D in other datasets may also help reassure researchers that estimates in a new dataset are reliable. Let us take the example of assessing the additional prognostic value of a new factor, when added to an existing model. If the existing model has shown $D = 1.3$ in previous similar patient groups, then when fitting the model in a new dataset, we would hope to see a value of around $D = 1.3$. Additionally, if we see $D \simeq 1.3$ and adding the new factor to the model gives $D = 1.6$, then this reassures us that we have a reliable estimate of the extra value of the new factor.

Chapter 10

Discussion

10.1 Summary of thesis

Prognostic studies are often performed by researchers and so appear frequently in medical literature. The aim of such studies is generally to develop a multivariable model to predict the outcome of interest, and they often use time-to-event data analysed with the Cox proportional hazards model. Many prognostic studies are performed retrospectively and often without reference to sample size (Mallett et al., 2010), suggesting that a reasonably reliable prognostic study may often be more a matter of luck than of good planning. In this thesis we aimed to develop sample size calculations which could be used by researchers planning prognostic studies or developing multivariable models, either to ensure they have sufficient precision or to estimate the precision they do have from the available data.

10.1.1 Review of available sample size guidance

Our first step was to review the sample size calculations already available (Chapter 2). We found that a calculation is available for the situation where it is desired to detect the prognostic value of a proposed new factor, in addition to an existing multivariable model; however this is not suitable for the majority of prognostic research, where we are interested in the (combined) effect of all factors on the outcome. The only other sample size guidance available for the multivariable context is the Events Per Variable (EPV) recommendation which states that at least 5 or 10 events per candidate model variable should be available for estimated regression coefficients to have proper accuracy and

precision and the correct coverage. We did not feel that this recommendation was optimal for the prognostic study scenario, since it was developed with the regression coefficients of the model in mind, rather than its prognostic ability, and does not appear to have been much tested beyond the two papers in which the idea originated (Concato et al., 1995; Peduzzi et al., 1995).

Finding no sample size calculations or recommendations which were well suited for use with prognostic multivariable models, we considered development of a new calculation, rather than adaptation of an existing one. The first step was to consider how we would measure the prognostic value of a model. In Chapter 2 we considered reviews performed by Schemper and Stare (1996) and Choodari-Oskooei (2008), and chose Royston & Sauerbrei's D as our measure of prognostic value, due to its one-to-one relationship with a measure of explained variation (R^2_D), appealing interpretation and robustness to outliers in the prognostic index.

10.1.2 Investigation of some properties of D

In Chapter 3 we investigated D to further uncover some of its properties. First we considered how it varied with sample size; to see whether there were any important relationships which would affect any potential sample size calculations. We also wanted to see whether there were any obvious cut offs of sample size dividing 'too small' from 'big enough' across all datasets. Our main result here was that D was variable while the sample size was small, but converged to a particular value as the sample size increased. We found that roughly at least 30 EPV or 200-300 events were required for the estimate of D to be within 25% of the 'true' D, but that this varied across datasets, and so a more formal sample size calculation was required.

We also looked at the issue of optimism, finding that an estimate of D from a model found using an automated variable selection procedure was virtually guaranteed to be inflated to some degree due to overfitting. We used Harrell et al.'s (1996) bootstrap method for estimating optimism which found that with 20 EPV, or with 200 events, the observed estimate of D contained between 10%–50% optimism for the vast majority of the datasets considered. We did not validate this method, however, and Royston and Sauerbrei (2004) found that it may overestimate optimism in D.

We also found in Chapter 3 that the estimate of $SE(D)$ described by Royston and Sauerbrei (2004) consistently underestimated the true standard error, especially when D was high, and so proposed using a bootstrap estimate of $SE(D)$ instead. The bootstrap estimate is quite quick to obtain for a single dataset and we found that for simulated data it much more closely approximated the gold standard empirical standard error.

In Chapter 4 we introduced an important parameter for formal sample sale size calculations: λ, the product of the number of events and the variance of D in a dataset. We found that λ can be estimated in a dataset with reasonable precision as long as there are more than about 170 events in the dataset. We also developed a fractional polynomial prediction model for λ using D and the censoring proportion from the dataset; this performed fairly well when tested in real datasets, not showing any obvious bias. Currently we cannot say which estimate of λ will be a better predictor of the value in a planned new study: the point estimate from a previous study, or λ from our equation, using the values of D and censoring proportion from the previous study.

Finally, we showed that λ is reasonably independent of dataset size, assuming a particular covariate structure. This assumption was key to development of sample size calculations for use in two different situations.

10.1.3 Development of sample size calculations

The first situation (in Chapter 5) is where it is desired to validate a value of D obtained from a model in a previous study, and importantly, an estimate of $SE(D)$ is available as well as an estimate of D. Two calculations were presented. The first (Sig-1) is a significance based calculation, where the proposed new study is a non-inferiority study and the reduction in D that we are willing to tolerate (δ) is specified along with a one-sided type I error (α) and power. The second (CI-1) is based on the precision of the estimate of D that will be obtained from the new study, in terms of the width of its confidence interval. Both calculations assume that the censoring proportion in the new study is the same as in the previous study; although with differences of 10% or less, the difference in resulting coverage is minimal. Broadly, if the censoring proportion in the new study is higher than the proportion in the previous study, the study will have lower precision than planned, whereas a lower censoring proportion would mean the new study has higher precision than planned. Additionally, if the new study showed a D that was markedly different to

the previous study this would also impact the precision of the study; a higher D in the new study implying lower precision, a lower D implying higher precision. This slightly counter-intuitive effect occurs because the variance of D increases with D.

The second situation, described in Chapter 6, was where only a target value of D is available, and no estimate of $SE(D)$. In this situation we again have a significance based (Sig-2) and a precision based (CI-2) calculation. For both these calculations we used the prediction model we developed in Chapter 4 to estimate λ. As already described, this model requires an estimate of the censoring proportion of the new study. This should be approximately known by researchers, who should have a good idea of survival rates in their disease area, and will have planned the length of study follow up. The model also requires an estimate of D; obtaining this may be more problematic. At one end of the scale, a good estimate of D may be available from a previous study (for which individual patient data is not available to the researchers, and for which no estimate of $SE(D)$ is available), while at the other end D may have to be effectively guessed. However, as censoring proportion and the estimate of D are explicit in the sample size calculations for this scenario, it is straightforward to calculate a range of sample sizes for various possible values of these parameters, and we strongly recommend this is done to help decision making.

We have highlighted the issue of studies inadvertently having higher or lower precision than planned due to misspecification of important parameters in our sample size calculations, however it is worth noting that this is a potential problem in any scenario, including randomised clinical trials. Any divergence from the expected censoring rate or hazard ratio for treatment in a clinical trial would mean that the originally calculated sample size was either too large or too small; however post-hoc calculations of power are not routinely performed. Thus the sample size calculations we have developed do not differ in principle from the (widely accepted) calculations for randomised trials in this respect.

As a final point on the calculations presented in Chapters 5 and 6, we note that there are four further calculations that could have been presented in the same vein. The calculations developed here were based on two binary factors: firstly whether the quantity λ was estimated from a previous study (Sig-1 and CI-1), or from our estimating equation (Sig-2 and CI-2); and secondly whether the calculation was based on significance testing

(Sig-1 and Sig-2) or CI width (CI-1 and CI-2). There is a further factor of interest: whether the test statistic is based on a difference in D, or on the value of D itself. In this work we developed the calculations so that Sig-1 is based on a difference in D, while Sig-2, CI-1 and CI-2 are based on the value of D itself, but this could be changed to give an additional four calculations. Finally, Sig-1 and Sig-2 are presented as calculations for non-inferiority studies here, but by changing the α to two-sided rather than one-sided the study can be powered for superiority instead.

10.1.4 Application of sample size calculations

When testing all four of these calculations with parameters from 10 real datasets, with a difference in D of $\delta = 0.1$ (calculation CI-1) or a 95% confidence interval of total width 0.2 for D (Sig-2 and CI-2), we found that none gave a sample size of fewer than 500 events and most were much higher. Calculation Sig-1 has an implicit restriction on the minimum difference in D that can be detected, and sometimes this minimum can be rather high. Sample sizes output from Sig-1 approach ∞ as δ approaches its minimum, and so generally this calculation also outputs a rather high required number of events.

The large sample sizes output by all four calculations lead us to believe that underpowering of prognostic studies is common, as perusal of the literature reveals many publications with sample size of less than 100 events. Relaxing the δ to be detected, or the precision in the estimate of D, will of course reduce the sample size required. We found that the CI based calculations CI-1 and CI-2 performed similarly, while Sig-1 always output a higher sample size than Sig-2 with the same parameters.

We also considered the precision that a given sample size will 'buy' under each of the four calculations. We found that for most of the scenarios we considered, with 200 events the CI based calculations (CI-1 and CI-2) provided a CI with half-width w of around 0.25; while 500 events gave w of around 0.16 and 1000 events a w of around 0.11. For small first studies or higher estimated D, the w available was slightly higher. Increasing the number of events by a factor of k leads to an increase in precision from calculations CI-1 and CI-2 by a factor of $\sqrt{k}$. This relationship between events and precision also holds for the significance based calculation Sig-2, where 500 events gave a detectable difference in D of between 0.22 and 0.37, and with 1000 events, between 0.15 and 0.26. Calculation Sig-1, however, behaved differently because of its implicit restriction on δ. The precision

available from a fixed number of events using Sig-1 was always higher than that from Sig-2, and the relationship between number of events and precision more complex.

Finally, we considered the desired precision in calculations CI-1 and CI-2 as a proportion of the value of D. We found that this could be a useful way to avoid the loss of precision experienced when D is higher than anticipated in a study, and also to sidestep the large sample sizes typically required by the calculations when D is high. Further, using a composite sample size consisting of both a fixed value of δ or w and a percentage of D means additionally that the expected precision will be obtained whatever the value of D in the new study. This somewhat mitigates the worry about whether an obtained estimate of D is accurate enough, and we feel it is quite a practical solution for the situation where researchers planning a study are not very sure about the target value of the parameter to be estimated.

10.1.5 Improving utility of sample size calculations

To aid researchers in estimating D and thus improve the usability of the calculations Sig-2 and CI-2 for the second scenario, in the last part of this thesis a detailed literature search of papers reporting D from models developed in real studies was conducted. This resulted in 34 papers which presented estimates of D for either predicting various diseases in healthy patients (risk models), or for predicting disease events in patients with a particular disease. We wished to expand this collection further, and to do this developed an empirical-based method to convert values of Harrell's c-index to D. A literature search of papers reporting c resulted in estimates of D from a further 77 papers. The final D Library consisting of all estimates of D found during the literature searches is presented in Appendix E and it was summarised in Chapter 9.

We found that on average, risk models had slightly higher D values than true prognostic models, which may be due to the more heterogenous healthy patient populations used in the former. We also found that the same model could output quite different values of D in different datasets, which may have an impact on prognostic study planning. It is unlikely that researchers seeking an appropriate D value for use with our calculations will find a perfect match in the literature, and differences in study size, the models used, and case mix of the population deserve careful scrutiny.

10.2 Recommendations

10.2.1 Sample size recommendations

If researchers wish to develop a new or altered multivariable prognostic model, or validate an existing one, they must consider various questions to determine which sample size calculation to use. A suggested decision tree is given in Figure 10.1.

At this point, we primarily recommend calculations Sig-2 and CI-2 over Sig-1 and CI-1, regardless of whether an estimate of $SE(D)$ is available from a previous study. This is in part due to the restrictions on δ implicit in calculation Sig-1, which in many cases may mean the minimum δ to be detected is to large to be of use. Another issue is the uncertainty around estimates of $SE(D)$; or more accurately, estimates of λ (the product of $var(D)$ and number of events in the dataset, which is present in all four calculations). Our research on the sampling distribution of λ appeared to show that a single value of λ from a previous similar study is not necessarily going to be a better predictor of λ in a new study than a λ predicted using our prediction model (equation 4.2). Further research may shed more light on the prediction error from both methods of estimating λ, but currently we cannot recommend one method over the other. The flexibility of Sig-2 and CI-2 in terms of the ease with which researchers can use them to generate a range of sample sizes for various values of D and censoring proportion swings the balance in favour of these two calculations. This flexibility is especially important as there may not always be strong evidence to support one single value of D over all others. Calculations Sig-2 and CI-2 may be made even more flexible by considering the desired precision as a percentage of D, rather than as a fixed value, as shown in Chapter 8; or by using our idea of a 'composite' sample size.

Working through the decision tree in Figure 10.1, the first consideration is whether an estimate of D is available from a previous similar study, as this should be the best possible estimate. If it is, then researchers can proceed to use either calculation Sig-2 and CI-2 (together with an estimate of study censoring proportion). If no estimate of D is available from a previous study, it must be sought from another source. There may be a suitable value of D in our library (or in future publications reporting D), but if not, the next best option is to search for an appropriate value of c in the literature. If such a value

is found, it can be approximately converted using our conversion and again calculation Sig-2 or CI-2 used.

One question that arises here is the errors in our estimate of D caused by (i) a source study or publication not being similar in all respects to the planned study and (ii) the conversion from c to D. For example, if both options were available: is it better to use an estimate of D from a study found in the literature which is dissimilar in one or two important respects, or to use an estimate of c from a very similar study, converted to D? Currently we do not have an answer to this and it would be rather difficult to investigate. In Figure 10.1 an estimate of D is given priority.

If no suitable estimate of c can be found, then our recommendation – following the work in Chapters 8 and 9 – is to use $D = 1.4$ (which is equivalent to $R^2_D = 32\%$) as this was the approximate mean (and median) value of D from the prognostic studies found in the literature search. Calculation Sig-2 or CI-2 can then be used.

In any situation where researchers are not very certain about their target or 'best' estimate of D, they may wish to consider our idea of a composite sample size and choose both a fixed limit for precision (δ or w), and an acceptable percentage of D. The maximum sample size output across all possible values of D will result in a study with the desired precision, regardless of the value of D observed.

Alongside the issue of estimating a suitable value for D, is the issue of what is a suitable value to choose for either δ or w. While the interpretation of D may be reasonably straightforward for researchers who already have a grasp of survival analysis, the interpretation of a difference in D may be more difficult. One possible way to illustrate the meaning of such a difference may be to express it in terms of R^2_D; however it is important to note that the difference in explained variation will vary depending on the value of D. More work on this problem would be useful if the calculations are to be used and understood; researchers must have a grasp of what difference in D they are excluding or accepting with their study, and how it relates to the performance of their model.

Where researchers wish to detect the additional value of a single new factor on top of previously known factors, they have an additional option. They can use the calculations developed by Hsieh and Lavori (2000) to determine sample size for a planned study or to calculate the power available from the data they have; although technically these calculations are not based on prognostic ability of the factor but the size of its log hazard

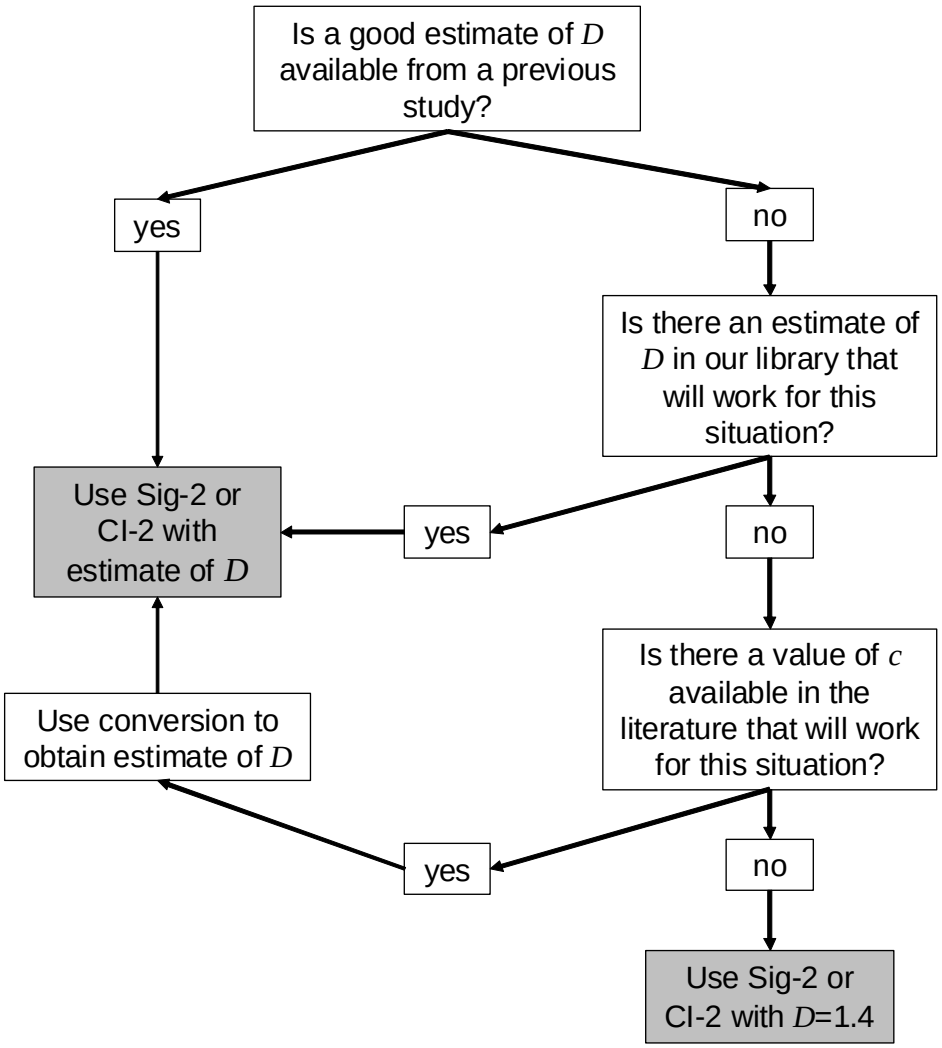

Figure 10.1: Flowchart to aid decision making about which sample size to use

ratio β in the Cox model. Note that in order to use Hsieh and Lavori's (2000) calculation, researchers must estimate the R^2 for predicting the new variable of interest with the existing factors, which will usually require a suitable existing dataset. If it is desired to test the additional prognostic value of a single new factor in terms of D, then the flowchart in Figure 10.1 and sample sizes calculations Sig-2 or CI-2 can still be used.

10.2.2 Obtaining estimates of D in the model validation context

In the model-validation context, if individual patient data from a suitable previous study is available to estimate D from, additional steps may be taken to estimate an optimism-

adjusted estimate of D in the original study, if it is suspected that the model in the study was overfitted. In Chapter 3 we used Harrell et al.'s (1996) enhanced bootstrap method to estimate optimism, however we did not validate this method in our data. Harrell (2001) describes some other methods for estimation of optimism, while Royston and Sauerbrei (2004) developed an adjusted D quantity called D_{adj}. D_{adj} accounts for the portion of optimism which is a result of parameter uncertainty, but not that which is due to data-driven model selection techniques.

Where individual patient data is not available, if it is suspected that a model providing an estimate of D may be overfitted (for example, due to automatic variable selection procedures, non-validation of the model or a small dataset or EPV), the possible optimism in the estimate should be borne in mind. An unrealistically high D will result in higher sample size and thus a less efficient study. Due to lack of individual patient data, researchers will only be able to guess at the magnitude of the problem.

10.2.3 Post-study recommendations

Calculation of $SE(D)$

Regardless of the method used to calculate sample size, at the end of the study we recommend that the bootstrap estimate of $SE(D)$ is obtained as described in Chapter 3. 500 bootstraps should be sufficient to estimate $SE(D)$ to within 10% of its value as long as the study included 150 or more events. If the dataset is smaller than this accuracy may decrease to $\sim$15%; 1000 bootstraps may be used if more accuracy is desired.

Bootstrapping $SE(D)$ is particularly important if $D \geq 1.0$, but as it is a quick procedure we recommend it is done regardless of the value of D. Note that the `str2d` command in Stata does provide a bootstrap standard error option.

Optimism

In the model development context, we recommend that an optimism-adjusted estimate of D is calculated, especially if the variables in the final model were selected using an automatic selection procedure (in which case, the selection process must be included in the optimism estimation).

10.3 Remarks on D

During this work we have had the opportunity to reflect on D and its use and present some of our thoughts here.

When reviewing values of D from a literature search, it is important to consider that the absolute value of D can be changed simply by changing the case mix of the patients in the dataset. The more heterogenous the dataset, the more variables there are available to differentiate between patients in a model and thus the higher D becomes. This is common to all measures of prognostic value, not just D.

This effect means that it is important to give careful thought to what is a suitable population for developing a model on, especially if a prospective study is planned. Often a prognostic model is developed in order to help determine a treatment or course of action for a particular subgroup of patients. For example, the aim of the model may be to divide patients into risk groups in order to identify the highest risk patients who are most likely to benefit from an aggressive treatment. In this case, the model would likely be developed on a population of patients who currently receive the same or similar course of treatment, the aim being to learn whether we should in fact be treating these patients differently.

If the best model developed on such a group of patients still has a low D, then assuming no important available factors have been omitted, this tells us that currently we can't differentiate much further between patients within this group. If additionally these patients have a variety of different disease outcomes, or perhaps show the same outcome but over a wide range of time, then this is likely a sign that there is potentially more to learn about these patients. Thus, in a sense, D tells us how much we know about the disease in this homogenous group: the larger D is, the more we know.

D can also tell us more generally how much we know about a disease, or the causes of a disease, if it is calculated from a model developed on a wide variety of patients with the disease or at risk of the disease. For example, we found that risk models seeking to predict occurrence of diabetes generally had higher values of D than models predicting cardiovascular disease; in particular this was true for the models developed on very large general practice datasets. This implies that we know more about the causes of diabetes than we do about the causes of cardiovascular disease.

10.4 Further research

Various parts of this thesis would benefit from further research; either due to limitations in the work, lack of time to expand areas as far we would like, or interesting results which are not related to the main theme of the thesis but deserve further exploration.

Relationship between D and EPV

Our conclusions related to EPV were not as strong or generalisable as they could be, being based only on a selection of six real datasets. For a more systematic consideration of the EPV issue multivariable data with different covariate structures could be simulated, for example using the `drawnorm` function in Stata or more sophisticated methodology based on the covariate structure of existing datasets. This should enable further insight into aspects of D's relationship with EPV.

Optimism in D

In Chapter 3 we showed that unvalidated estimates of D can contain optimism when models are not pre-selected (and even when they are, due to parameter uncertainty). However, in the work done with real datasets in the remainder of the thesis, the optimism in D was not calculated, since Harrell et al.'s (1996) bootstrap method would have been too time-consuming in this situation, especially as we were also estimating $SE(D)$ with bootstrapping. Further research into the best method to adjust for optimism in D would be prudent as part of general efforts to improve the quality of prognostic studies and temper the hopes raised when a new prognostic model is reported which appears to be highly predictive. Any method must be efficient and reasonably easy to implement; it is not yet clear whether Harrell et al.'s (1996) bootstrap method fulfils this criteria when used with D. As part of this work, we feel further work on Royston and Sauerbrei's D_{adj} would be worthwhile.

Sampling distribution of λ

We defined and described λ in Chapter 4 and looked at its mean value in different datasets, but did not deeply consider its sampling distribution due to time limitations. More work would be beneficial here as the sampling distribution of λ was key in our

decision to recommend sample size calculations Sig-2 and CI-2 over Sig-1 and CI-1, even when an estimate of $SE(D)$ is available.

Prediction of λ

We developed an equation for predicting λ from D and the censoring proportion which seemed to work reasonably well in simulated data with random and administrative censoring, but was less accurate in real data. Further research to quantify the errors in the prediction of λ and potentially refine the prediction model would be valuable in order to improve calculations Sig-2 and CI-2. This work may also help determine more conclusively whether the Sig-2 and CI-2 calculations are indeed preferable when an estimate of $SE(D)$ is available.

Testing sample size calculations

Like all the simulated datasets produced for this thesis, simulated model PI data with random censoring were used to test the four sample size calculations developed in Chapters 5 and 6 for correct power and type I error (or coverage in the case of the CI based calculations). Repeating this testing with simulated multivariable data and data with different censoring patterns should strengthen our conclusions and may highlight situations in which the calculations do not result in the required power, type I error or coverage.

Effect of censoring on Harrell's c

Work done in Chapter 8 on a transformation from Harrell's c to D suggests that only administrative censoring affects the value of Harrell's c index, while random censoring does not seem have an effect. This distinction does not appear to have been made before, so it may be worthwhile looking further at this issue with more simulation study.

Other approaches to sample size

In this thesis we concentrate mainly on just one approach to sample size; a 'traditional' calculation based on significance or confidence intervals, based on one single measure of model performance. There are other ways we could have explored the problem; for example, considering the unknown 'true' model for a particular dataset and how to find a model which is 'almost as good as' this, or considering calibration of a model in terms

of accuracy of survival predictions. We could also have used different measures of model performance such as Harrell's C, or one of the other measures with suitable properties which were mentioned in Chapter 2.

10.5 Final conclusions

In this thesis we have worked to develop sample size calculations for development or validation of prognostic multivariable models, based on the D statistic, and also tried to ensure that they are practical for use in clinical research. Although the sample sizes output by the calculations tend to be large, we have given suggestions on how study size can be managed, for example by considering precision as a proportion of the measure of interest, rather than as a fixed value. We have also explored the D statistic further, investigating some of its properties and reviewing its value in different disease areas.

We hope that these calculations, the guidance provided for their use, and our work on D will help improve the quality of prognostic research. As well as the calculations being used to provide sample sizes for prospective studies, they can also be used for retrospective research, either to give the required sample size before suitable data is sought, or to calculate the resulting precision when a dataset has already been chosen. At the very least we hope that the existence of these calculations – the first formal sample size calculations developed for prognostic research – will encourage researchers to consider the issue of sample size as a matter of course when developing or validating prognostic multivariable models.

Appendix A

Datasets used in the thesis

This appendix describes the datasets used in this thesis.

A.1 Datasets used in Chapters 3, 4, 7, and 8

26 datasets are variously used throughout this thesis. In Chapter 3 a selection is used to illustrate the behaviour of D with changing sample size, and in Chapter 4 to investigate λ. In Chapter 7, ten of the datasets are used as starting points to give examples of the sample size calculations developed in Chapters 5 and 6. Finally, they are all used in Chapter 8 to contribute datapoints to the development of the model used to transform values of Royston and Sauerbrei's D to Harrell's c.

The 26 datasets are summarised in Table A.1 below and given an acronym for brevity. Most of the datasets chosen originated from clinical trials or cohort studies and have previously been used to illustrate modelling techniques or prognostic analyses. Some further information including the endpoint used and a reference for more detail is given below for each dataset. Note that in Table A.1, the variables column refers to the number of potential variables included in the `mfp` input string for that dataset; usually this was all variables available.

Dataset	Disease	Patients	Events	Variables
APC	Prostate cancer	475	338	14
FBC	Breast cancer	686	299	8
FOL	Follicular lymphoma	767	573	18
GLI	Malignant glioma	411	274	13
HOS	Cardiovascular disease	500	215	14
KCA	Kidney cancer	347	322	9
LEG	Leg ulcer	200	97	9
LVA	Lung cancer	137	128	5
MYE	Myeloma	1057	856	11
OVA	Ovarian cancer	474	402	8
PBC	Primary biliary cirrhosis	312	125	20
PBC2	"	216	105	5
RBC	Breast cancer	2982	1518	9
RBC5	"	2982	1518	14
RBC10	"	2982	1518	19
RBC15	"	2982	1518	24
SEER	Breast cancer			
SEER AT	"	3666	235	14
SEER CT	"	4009	386	14
SEER DE	"	4422	400	14
SEER HI	"	6923	731	14
SEER IA	"	11339	1084	14
SEER NM	"	12028	1269	14
SEER SE	"	13533	1540	14
SEER SF	"	13671	1184	14
SEER UT	"	14213	1270	14
STE	Cardiovascular disease	3873	460	24
WHI2	"	12017	1628	7
WHI3	"	2712	515	7
WHI4	"	1583	331	7

Table A.1: Datasets used throughout thesis

APC

From a trial in patients with advanced prostate cancer and is described and analysed more fully in Byar and Green (1980).

FBC

From a cohort study performed by the German Breast Cancer Study Group in women with primary node-positive breast cancer Sauerbrei and Royston (1999). The endpoint is recurrence-free survival.

FOL

Pooled data from various British National Lymphoma Investigation trials.

GLI

From a randomised trial of three chemotherapy regimes which recruited 447 patients, however in this analysis only the 411 patients with complete data are considered. The dataset is described fully in Sauerbrei and Schumacher (1992). An endpoint of overall survival (OS) is used.

HOS

From the Worcester Heart Attack Study (WHAS), a cohort study looking at factors associated with long-term survival after acute myocardial infarction. The original study included over 11,000 patients; this random subset of 500 is described in and used throughout Hosmer et al.'s (2008) book as an example dataset, with the endpoint of OS.

KCA

From the MRC RE01 randomised trial, this dataset is as used in Royston and Sauerbrei's (2008) book, with some missing covariate values imputed and OS as endpoint.

LEG

From a randomised clinical trial of a dressing in patients with venous leg ulcer Smith et al. (1992a). The endpoint for this data is time to complete healing of the ulcer.

LVA

Veterans Administration lung cancer trial presented in Appendix 1 of Kalbfleisch and Prentice (1980), with an endpoint of OS. In this trial, males with advanced inoperable lung cancer were randomized to a standard therapy and a test chemotherapy.

MYE

The dataset originally contained 1087 patients, however only the 1057 with complete covariate data were retained for this investigation (MacLennan et al., 1988). The endpoint is OS.

OVA

From a randomised trial of patients with advanced ovarian cancer, conducted in Italy by Valsecchi et al. (1996); the endpoint used is OS.

PBC

From a randomised trial of a drug in patients with primary biliary cirrhosis (PBC). This data is presented in Fleming and Harrington (1991), appendix D1. The endpoint is OS.

PBC2

From a randomised trial of 248 patients with PBC (Christensen et al., 1985); only the 216 patients with complete data are included in this dataset. OS is used as the endpoint.

RBC

This breast cancer dataset is as used in Royston and Sauerbrei's (2008) book, with missing values imputed and the endpoint recurrence-free survival. As one of the larger datasets, this one was chosen to investigate the effect of additional noise variables in Chapter 3. To this end, 5, 10 and 15 uniform(0.1,1.1) distributed random variables were simulated and added to the RBC dataset to make three new datasets used in the same way as the original dataset (RBC5, RBC10, RBC15).

SEER breast cancer

This dataset was originally from the Surveillance, Epidemiology and End Results (SEER) program (SEER, 2000); specifically the breast cancer portion of SEER 9, and patients diagnosed 1988-1997 inclusive. SEER 9 consists of cancer registries from 9 geographical areas of the USA (listed in Table A.1) and the investigation was carried out in each of the 9 registries separately. The endpoint is OS. The final datasets used in the thesis were cut-down

versions of the original SEER dataset, with only factors known to be prognostic kept in the dataset. Some of the prognostic factors have been imputed where missing.

STE

From the SMART (Second Manifestations of ARTerial disease) cohort study designed to identify predictors of future cardiovascular events in patients with clinical symptoms of cardiovascular disease. This study is described extensively in Steyerberg (2008).

WHI

The Whitehall I Study was a large cohort study conducted amongst male UK civil servants; with 18,000 recruited between 1967-1977. We split the dataset into three parts by job grade, and used an endpoint of cardiovascular death, as described in Marmot et al. (1978).

SEER datasets used in Chapter 8

In Chapter 8, various SEER datasets were used to obtain additional datapoints for building the empirical D to c transformation model (SEER, 2000). A large amount of data manipulation and recoding of variables was performed on the raw downloaded SEER datasets, as in their original form each disease area dataset is rather heterogeneous, comprising a number of cancer types in the same organ or part of the body. Additionally, although all SEER data uses a common dataset, the list of available variables is extensive; the small subset of variables which may actually be useful predictors was determined by inspection for each disease site individually. All Stata programming done on the datasets is available from the author if required.

Appendix B

Models fitted to real datasets

This Appendix describes the model fitting procedures used with the real datasets in Chapters 3, 4, 8 and 9. The same lines of code were used regardless of the changing p values inserted into 'p'. The datasets are described in Appendix A.

B.1 Datasets used in Chapters 3, 4, 7, and 8

Some of these datasets are used in several chapters. The same `mfp` command line is used every time the dataset is mentioned.

APC

```
. stset survtime, fail(cens)
. xi: mfp stcox age wt sbp dbp sz ap hg sg pf hx bm i.stage ekg rx, select('p')
> alpha('p')
```

FBC

```
. stset rectime, fail(censrec)
. xi: mfp stcox i.hormon x1 i.x2 x3 x4a x4b x5 x6 x7, select('p') alpha('p')
```

FOL

```
. stset stime, fail(status)
. xi: mfp stcox grade2 grade3 grade4 iblast iinerrt ifgrade hist2 hist3 stage2
> stage3 stage4 age sex iesr ilymph ialbumin extra bulk, select('p') alpha('p')
```

GLI

```
. stset survtime, fail(cens)
. xi: mfp stcox sex tsymp gradd1 gradd2 age karno surgd1 surgd2 convul cort epi
> amnesia ops aph trt, select(‘p’) alpha(‘p’)
```

HOS

```
. stset lenfol, fail(fstat)
. xi: mfp stcox age gender hr sysbp diasbp bmi cvd afb sho chf av3 miord mitype
> i.year, select(‘p’) alpha(‘p’)
```

KCA

```
. stset survtime, fail(cens)
. xi: mfp stcox age sex whod1 whod2 t_dt t_mt rem mets haem wcc trt, select(‘p’)
> alpha(‘p’)
```

LEG

```
. stset ttevent, fail(censored)
. xi: mfp stcox i.treatmnt weight diastbp ankpres ulcarea age mthson height
> deepppg, select(‘p’) alpha(‘p’)
```

LVA

```
. stset t, fail(cens)
. xi: mfp stcox kps diagtime age prior treat squamous small adeno, select(‘p’)
> alpha(‘p’)
```

MYE

```
. drop if cens==1 & time==0
. stset time, fail(cens)
. xi: mfp stcox age hb creat indexd1 indexd2 sb2 calcium igm albumin treat3
> treat4, select(‘p’) alpha(‘p’)
```

OVA

```
. stset survtime, fail(surv)
. xi: mfp stcox c1 c2 rt1 rt2 g1 g2 serous k1, select(‘p’) alpha(‘p’)
```

PBC

```
. stset survtime, fail(cens)
. xi: mfp stcox i.trt age sex asc hep spider edemad1 edemad2 bil chol alb cu ap
> sgot trig plt pro staged1 staged2 staged3, select('p') alpha('p')
```

PBC2

```
. stset time, fail(cens)
. xi: mfp stcox i.treat age bilir albumin cirrh central, select('p') alpha('p')
```

RBC

```
. stset reltime, fail(cens)
. xi: mfp stcox age meno sized1 sized2 i.grade nodes pgr er hormon chemo,
> select('p') alpha('p')
```

SEER breast cancer

```
. stset futime, fail(event)
. xi: mfp stcox age npos nex er pr eodsize raceblck spanish histduct bcs rt
> married i.lateral g2 g3 g4, select('p') alpha('p')
```

STE

```
. stset tevent, fail( event)
. xi:mfp stcox sex age diabetes cerebral cardiac aaa periph stenosis systbp
> diastbp lengtho weighto cholo hdlo ldlo trigo homoco gluto creato imto
> i.albumin i.smoking i.alcohol packyrs, select('p') alpha('p')
```

WHI datasets

```
. stset pyar, fail(chd)
. xi: mfp stcox cigs sysbp diasbp age ht wt chol, select('p') alpha('p')
```

B.2 SEER datasets used in Chapter 8

For the SEER datasets used in Chapter 8, a p value of 0.05 was generally used in the `mfp` command line; however sometimes a higher value was used if it resulted in a markedly higher D value. The `mfp` command lines used for each disease site are given below. Note that each disease site dataset was divided into the nine registry datasets (HI, UT, NM,

AT, CT, SE, IA, SF, DE) unless otherwise specified. A dataset was not divided into the nine registry datasets if it was small (less than about 100 events), or if upon dividing it up some of the resulting registry datasets were found to contain less than 100 events; in the latter situation the smallest registries were combined in groups of two or three.

All these SEER datasets were identified as survival data using the Stata code

```
. stset surv, fail(event)
```

Colorectal

Colon

```
. xi: mfp stcox marital spanish agedx rxrad rxsurgprim nodepos nodesexa raceblack
> tumoursize tstage1 tstage2 tstage3 tstage4 nstage1 nstage2 mstage grade2 grade3
> grade4 numprim i.histtype_main i.subtype, select(0.05) alpha(0.05)
```

Rectum

```
. xi: mfp stcox marital spanish agedx rxrad  nodepos nodesexa raceblack tumoursize
> tstage1 tstage2 tstage3 tstage4 nstage1 nstage2 mstage grade2 grade3 grade4
> numprim i.histtype_main i.subtype , select(0.05) alpha(0.05)
```

Female genital

Vulva

Registries HI and UT combined, registries NM and AT combined; all other registries kept separate.

```
. xi: mfp stcox marital spanish agedx rxrad raceblack nodesexa tumoursize tstage2
> tstage3 tstage4 nstage1 nstage2 mstage  numprim i.histtype_main, select(0.05)
> alpha(0.05)
```

Vagina

Not divided into registries; kept as one single dataset.

```
. xi: mfp stcox marital spanish agedx rxrad raceblack nodesexa tumoursize tstage2
> tstage3 tstage4 nstage1 nstage2 mstage  numprim i.histtype_main, select(0.05)
> alpha(0.05)
```

Cervix uteri

```
. xi: mfp stcox marital spanish agedx rxrad rxsurgprim raceblack  tumoursize
> tstage2 tstage3 tstage4 nstage1 mstage grade2 grade3 grade4 numprim
> i.histtype_main i.subtype, select(0.05) alpha(0.05)
```

Corpus uteri

```
. xi: mfp stcox marital spanish agedx rxrad  raceblack tumoursize tstage2 tstage3
> tstage4 nstage1 mstage grade2 grade3 grade4 numprim i.histtype_main, select(0.5)
> alpha(0.05)
```

Ovary

```
. xi: mfp stcox marital spanish agedx rxrad  raceblack tumoursize tstage2 tstage3
> nstage1 mstage grade2 grade3 grade4 numprim i.histtype_main, select(0.5)
> alpha(0.05)
```

Blood cancers

Hodgkin's lymphoma

Not divided into registries; kept as one single dataset.

```
. xi: mfp stcox  marital spanish sex agedx rxsurgprim rxrad raceblack cssite2
> numprim  ext2 ext3 ext4 i.lymphrec, select(0.05) alpha(0.05)
```

Acute lymphoid leukaemia

```
. xi: mfp stcox marital spanish sex agedx rxrad raceblack numprim i.lymphrec,
> select(0.05) alpha(0.05)
```

Chronic lymphoid leukaemia

```
. xi: mfp stcox marital spanish sex agedx rxrad raceblack numprim, select(0.05)
> alpha(0.05)
```

Lymphosarcoma and reticulosarcoma

```
. xi: mfp stcox  marital spanish sex agedx rxrad raceblack numprim ext2 ext3 ext4
> i.subtype i.lymphrec, select(0.05) alpha(0.05)
```

Monocytic leukaemia

Not divided into registries; kept as one single dataset.

```
. xi: mfp stcox  marital spanish sex agedx rxrad raceblack numprim, select(0.05)
> alpha(0.05)
```

Multiple myeloma

```
. xi: mfp stcox  marital spanish sex agedx rxrad raceblack numprim, select(0.05)
> alpha(0.05)
```

Acute myeloid leukaemia

```
. xi: mfp stcox marital spanish sex agedx rxrad raceblack numprim i.histtype_main,
> select(0.05) alpha(0.05)
```

Chronic myeloid leukaemia

```
. xi: mfp stcox marital spanish sex agedx rxrad raceblack numprim i.histtype_main,
> select(0.05) alpha(0.05)
```

Nodular lymphoma

```
. xi: mfp stcox  marital spanish sex agedx rxsurgprim rxrad raceblack numprim
> cssite2 ext2 ext3 ext4, select(0.05) alpha(0.05)
```

Urinary

Bladder

```
. xi: mfp stcox subtype marital spanish sex agedx nodesexa cssize ajcct1 ajcct2
> ajcct3 ajcct4 ajccn1 ajccn2 i.ajccm rxsurgprim rxrad raceblack numprim,
> select(0.05) alpha(0.05)
```

Renal cell carcinoma

```
. xi: mfp stcox marital spanish sex agedx laterality nodesexa cssize ajcct2
> ajcct3  ajcct4 ajccn1 ajccn2 ajccm rxsurgprim rxrad raceblack numprim,
> select(0.05) alpha(0.05)
```

Renal pelvis and ureter

Registries IA and NM combined; registries HI, AT and UT combined. All others kept separate.

```
. xi: mfp stcox marital spanish sex agedx laterality nodesexa cssize i.ajcct
> i.ajccn i.ajccm rxsurgprim rxrad raceblack numprim, select(0.05) alpha(0.05)
```

Urethra

Not divided into registries; kept as one single dataset.

```
. xi: mfp stcox marital spanish sex agedx nodesexa tumoursize tstage2 tstage3
> tstage4 nstage1 nstage2 mstage  rxrad raceblack numprim, select(0.05)
> alpha(0.05)
```

'Poor' models

To add points to the lower ranges of the D vs c graph (Figure 8.6), intentionally poor models were built for some of the SEER datasets by using a severely limited pool of candidate predictors. The `mfp` program lines used for this are outlined below.

Renal cell carcinoma

```
. xi: mfp stcox  spanish nodesexa raceblack, select(1) alpha(0.05)
```

Rectum

```
. xi: mfp stcox marital nodesexa, select(1) alpha(0.05)
```

Cervix uteri

```
. xi: mfp stcox marital raceblack numprim, select(1) alpha(0.05)
```

Multiple myeloma

```
. xi: mfp stcox marital spanish sex, select(1) alpha(0.05)
```

Appendix C

Stata code for bootstrap procedure

This Stata code written by Patrick Royston (unpublished) implements the method used for estimating the optimism in *D*. This method was described by Harrell et al. (1996) and is based on Efron's (1983) refined bootstrap method.

```
*! version 1.0.01 PR 20feb2009
program define bootdval, rclass
version 10
// "Validate" (in the Harrell sense) the D-measure for a prognostic survival model
// selected using mfp.
syntax varlist(min=1 numeric) [if] [in], REPs(int) SAVing(string) ///
 [replace seed(int 0) SELect(real 1) ALPha(real 0.05) *]

if `reps' < 1 {
        di as error "reps() must be a positive integer"
        exit 2001
}

// Check whether output file already exists
if "`replace'" == "" confirm new file `"`saving'"'

// Create the output file
tempname handle
postfile `handle' d1 r2d1 r2pm1 d2 r2d2 r2pm2 d_orig d_origse nevent ///
        using `"`saving'"', replace

// Mark the estimation sample
marksample touse

// Run mfp on the original data and report/store results of D analysis
tempvar xb
di as txt _n "Running mfp on the original sample..."
quietly mfp stcox `varlist' if `touse', select(`select') alpha(`alpha') `options'
mfp
predict `xb' if `touse', xb
str2d stcox `xb'
```

```
return scalar D = r(D)
return scalar r2 = r(r2)
return scalar r2pm = r(r2pm)

loc d_orig `r(D)'
loc d_origse `r(sD)'
loc nevent `r(events)'

drop `xb'

// Set the random number seed
if `seed' > 0 set seed `seed'

// Keep track of failed model-fitting attempts
local failed 0

tempname mfpmodel selectedmodel

// Do model selection on a boot sample, predict on original sample, get D stats
local i 1
quietly while `i' <= `reps' {
        preserve
        // bsample drops observations that are filtered out by if, in or missing
        // covariate values
        bsample if `touse'
        capture mfp stcox `varlist', select(`select') alpha(`alpha') `options'
        if c(rc) > 0  {
                // mfp failed, for some reason - omit this bootstrap sample
                local ++failed
                restore
        }
        else {
                if "`e(fp_fvl)'" == "" {
                        // no variables selected by mfp
                        local d1 0
                        local r2d1 0
                        local r2pm1 0
                        local d2 0
                        local r2d2 0
                        local r2pm2 0
                        restore
                }
                else {
                        // Estimate D and R2 stats on bootstrap and original samples
                        predict `xb', xb
                        genmfpvars
                        local model `r(vl)'
                        _estimates hold `mfpmodel'
                        // Refit the selected model (to be used for prediction on
                        //original data)
                        stcox `model'
                        _estimates hold `selectedmodel'
                        str2d stcox `xb'
                        local d1 = r(D)
                        local r2d1 = r(r2)
                        local r2pm1 = r(r2pm)
                        restore
                        _estimates unhold `mfpmodel'
                        genmfpvars
                        _estimates unhold `selectedmodel'
                        predict `xb' if `touse', xb
                        str2d stcox `xb'
                        local d2 = r(D)
```

```
                        local r2d2 = r(r2)
                        local r2pm2 = r(r2pm)
                        drop `xb'
                }
                local ++i
                post `handle' (`d1') (`r2d1') (`r2pm1') (`d2') (`r2d2') (`r2pm2') ///
                        (`d_orig') (`d_origse') (`nevent')
        }
        if mod(`i', 10) == 0 noi di as txt `i', _c
}
di
postclose `handle'
preserve
use `"`saving'"', replace
lab var d1 "D on bootstrap samples"
lab var r2d1 "R2(D) on bootstrap samples"
lab var r2pm1 "R2(PM) on bootstrap samples"
lab var d2 "D predicted on original sample"
lab var r2d2 "R2(D) predicted on original sample"
lab var r2pm2 "R2(PM) predicted on original sample"
qui sum d1
local md1 = r(mean)
qui sum d2
local md2 = r(mean)
local opt = `md1' - `md2'
di as txt _n "D(original sample) = " %8.4f as res return(D) ///
 as txt " optimism = " %6.4f as res `opt' ///
 as txt " corrected D = " %8.4f as res return(D) - `opt'
save `"`saving'"', replace
restore
if `failed' > 0 di as txt _n "[mfp failed in `failed' bootstrap replicate(s)]"
return scalar failed = `failed'
return scalar Dcorr = return(D) - `opt'
return scalar opt = `opt'
end

program define genmfpvars, rclass
version 10
// Generates variables from the most recent fit of MFP
local nxvar `e(fp_nx)'
local vl ""
forvalues i = 1 / `nxvar' {
        local pwrs `e(fp_k`i')'
        if "`pwrs'" != "." {
                local x `e(fp_x`i')'
                fracgen `x' `pwrs', replace
                local vl `vl' `r(names)'
        }
}
return local vl `vl'
end
```

Appendix D

Sample size: further examples

This Appendix contains tables of sample size required for the four sample sizes developed in Chapters 5 and 6, for ten of the real datasets described in Appendix B and various values of w and δ. See Chapter 7 for full details.

Each table includes four sub-tables for the four calculations: Sig-1 (significance based) and CI-1 (confidence interval (CI) based) which were for the situation where a previous estimate of D is to be validated and an estimate of $SE(D)$ is available; and Sig-2 (significance based) and CI-2 (CI based) for use where an estimate of $SE(D)$ is not available. Within each sub table the number of events (e_2) and the number of patients (pts) is given for various values of either δ (the difference in D to be detected for calculations Sig-1 and Sig-2) or w (the half-width of the desired CI, for calculations CI-1 and CI-2).

For Sig-1 and Sig-2, $\alpha = 0.05$ and power=90%.

For CI-1 and CI-2, the confidence intervals are always 95%.

APC

Sig-1			CI-1			Sig-2			CI-2		
δ	e_2	pts	w	e_2	pts	δ	e_2	pts	w	e_2	pts
0.3	2056	2890	0.05	4688	6589	0.05	11420	16049	0.05	5123	7200
0.35	578	813	0.1	1172	1648	0.1	2855	4013	0.1	1281	1801
0.4	316	445	0.15	521	733	0.15	1269	1754	0.15	570	802
0.45	209	294	0.2	293	412	0.2	714	1004	0.2	321	452
0.5	152	214	0.3	131	185	0.3	318	447	0.3	143	201
0.6	93	131	0.4	74	104	0.4	179	252	0.4	81	114

Table D.1: Sample size calculations based on APC study

GLI

Sig-1			CI-1			Sig-2			CI-2		
δ	e_2	pts	w	e_2	pts	δ	e_2	pts	w	e_2	pts
0.38	1598	2406	0.05	6041	9095	0.05	13119	19751	0.05	5885	8860
0.4	918	1383	0.1	1511	2275	0.1	3280	4939	0.1	1472	2217
0.43	547	824	0.15	672	1012	0.15	1458	2196	0.15	654	985
0.45	426	642	0.2	378	570	0.2	820	1235	0.2	368	555
0.5	266	401	0.3	168	253	0.3	365	550	0.3	164	247
0.6	143	216	0.4	95	144	0.4	205	309	0.4	92	139

Table D.2: Sample size calculations based on GLI study

LEG

Sig-1			CI-1			Sig-2			CI-2		
δ	e_2	pts	w	e_2	pts	δ	e_2	pts	w	e_2	pts
0.79	4357	8984	0.05	10626	21910	0.05	19181	39549	0.05	8604	17741
0.8	2008	4141	0.1	2657	5479	0.1	4796	9889	0.1	2151	4436
0.85	529	1091	0.15	1181	2436	0.15	2132	4396	0.15	956	1972
0.9	297	613	0.2	665	1372	0.2	1199	2473	0.2	538	1110
0.95	203	419	0.3	296	611	0.3	533	1099	0.3	239	493
1	153	316	0.4	167	345	0.4	300	619	0.4	135	279

Table D.3: Sample size calculations based on LEG study

LVA

Sig-1			CI-1			Sig-2			CI-2		
δ	e_2	pts	w	e_2	pts	δ	e_2	pts	w	e_2	pts
0.7	1772	1897	0.05	10496	11234	0.05	17871	19128	0.05	8016	8580
0.75	555	595	0.1	2624	2809	0.1	4468	4783	0.1	2004	2145
0.8	320	343	0.15	1167	1250	0.15	1986	2126	0.15	891	954
0.85	221	237	0.2	656	703	0.2	1117	1196	0.2	501	537
0.9	166	178	0.3	292	313	0.3	497	532	0.3	223	239
1	108	116	0.4	164	176	0.4	280	300	0.4	126	135

Table D.4: Sample size calculations based on LVA study

MYE

Sig-1			CI-1			Sig-2			CI-2		
δ	e_2	pts	w	e_2	pts	δ	e_2	pts	w	e_2	pts
0.22	5581	6892	0.05	6446	7960	0.05	11391	14066	0.05	5110	6310
0.25	1750	2161	0.1	1612	1991	0.1	2848	3517	0.1	1278	1579
0.3	748	924	0.15	717	886	0.15	1266	1564	0.15	568	702
0.35	447	552	0.2	403	498	0.2	712	880	0.2	320	396
0.4	305	377	0.3	180	223	0.3	317	391	0.3	142	176
0.5	173	214	0.4	101	125	0.4	178	220	0.4	80	99

Table D.5: Sample size calculations based on MYE study

PBC

Sig-1			CI-1			Sig-2			CI-2		
δ	e_2	pts	w	e_2	pts	δ	e_2	pts	w	e_2	pts
0.8	3087	7706	0.05	13796	34435	0.05	24627	61469	0.05	11047	27574
0.85	716	1788	0.1	3449	8609	0.1	6157	15368	0.1	2762	6894
0.9	395	986	0.15	1533	3827	0.15	2737	6832	0.15	1228	3066
0.95	268	669	0.2	863	2155	0.2	1540	3844	0.2	691	1725
1	200	500	0.3	384	959	0.3	685	1710	0.3	307	767
1.05	158	398	0.4	216	540	0.4	356	961	0.4	173	432

Table D.6: Sample size calculations based on PBC study

RBC

Sig-1			CI-1			Sig-2			CI-2		
δ	e_2	pts	w	e_2	pts	δ	e_2	pts	w	e_2	pts
0.18	4377	8599	0.05	6553	12873	0.05	11446	22485	0.05	5134	10086
0.2	2291	4501	0.1	1639	3220	0.1	2862	5623	0.1	1284	2523
0.23	1267	2489	0.15	729	1433	0.15	1272	2499	0.15	571	1122
0.25	950	1867	0.2	410	806	0.2	716	1407	0.2	321	631
0.3	554	1089	0.3	183	360	0.3	318	625	0.3	143	281
0.35	371	729	0.4	103	203	0.4	179	352	0.4	81	160

Table D.7: Sample size calculations based on RBC study

SEER DE

Sig-1			CI-1			Sig-2			CI-2		
δ	e_2	pts	w	e_2	pts	δ	e_2	pts	w	e_2	pts
0.18	2399	21082	0.05	5453	47920	0.05	10095	88712	0.05	4529	39800
0.2	1500	13182	0.1	1364	11987	0.1	2524	22181	0.1	1133	9957
0.23	917	8059	0.15	606	5326	0.15	1122	9860	0.15	504	4429
0.25	711	6249	0.2	341	2997	0.2	631	5546	0.2	284	2496
0.3	433	3806	0.3	152	1336	0.3	281	2470	0.3	126	1108
0.35	296	2602	0.4	86	756	0.4	158	1389	0.4	71	624

Table D.8: Sample size calculations based on SEER DE study

SEER NM

Sig-1			CI-1			Sig-2			CI-2		
δ	e_2	pts	w	e_2	pts	δ	e_2	pts	w	e_2	pts
0.25	11657	47946	0.05	4337	47946	0.05	9005	99551	0.05	4040	44663
0.28	1345	11995	0.1	1085	11995	0.1	2252	24896	0.1	1010	11166
0.3	818	5329	0.15	452	5329	0.15	1001	11067	0.15	449	4964
0.35	390	3007	0.2	272	3007	0.2	563	6224	0.2	253	2797
0.4	243	1338	0.3	121	1338	0.3	251	2775	0.3	113	1250
0.5	128	752	0.4	68	752	0.4	141	1559	0.4	64	708

Table D.9: Sample size calculations based on SEER NM study

STE

Sig-1			CI-1			Sig-2			CI-2		
δ	e_2	pts	w	e_2	pts	δ	e_2	pts	w	e_2	pts
0.28	1547	11479	0.05	4988	37009	0.05	8283	61457	0.05	3716	27572
0.3	940	6975	0.1	1247	9253	0.1	2071	15366	0.1	929	6893
0.35	448	3324	0.15	555	4118	0.15	921	6834	0.15	413	3065
0.4	280	2078	0.2	312	2315	0.2	518	3844	0.2	233	1729
0.5	147	1091	0.3	139	1032	0.3	231	1714	0.3	104	772
0.6	93	691	0.4	78	579	0.4	130	965	0.4	59	438

Table D.10: Sample size calculations based on STE study

Appendix E

D library

In the Chapter 8 we outlined the methods used to perform two literature searches for values of Harrell's *c* and Royston & Sauerbrei's *D*, and to develop a model to predict *D* from *c*. This appendix presents the results of these searches, refined and presented narratively as a library of *D* values for a selection of disease areas.

Only the highest value of *D* from a paper is included, however if more than one endpoint is considered, or multiple independent datasets, then more than one value of *D* may be reported. Multiple values may also be given where the dataset is divided by an exclusive subgroup such as gender, along with a result for the whole dataset.

In this appendix we present first the results from risk models predicting first disease events in healthy patients, and then the results from 'true' prognostic models predicting death or disease progression events in patients who already have a particular disease. Within these two main categories each different disease or diagnosis area is discussed separately, and within these subcategories different endpoints are separated if necessary.

Only *D* values are referred to and we do not differentiate between *D* values which were originally reported in the paper and values which were derived from Harrell's *c* using the transformation developed in Chapter 8. Many papers predicting events in healthy patients gave separate values of *D* for females and males; additionally for some diseases such as breast cancer and prostate cancer, studies were performed in single sex groups. For brevity, we use the notation D_F to indicate a *D* value for a female-only group and D_M for male-only.

For full references for all papers, see the *D* Library bibliography at the end of the thesis.

E.1 Risk models in healthy subjects

These models are developed to predict onset or incidence of some disease or condition in apparently healthy subjects.

E.1.1 Incidence of cardiovascular events

The majority of papers predicting onset of disease concerned the prediction of cardiovascular (CV) events. These papers have been grouped by the outcome being predicted and so the subsection headings refer to the endpoint of the study. Some papers sought to predict different outcomes using the same model, so some papers have results in more than one of the subsections below. Two papers predicting CV outcomes in hypertensive patients are presented separately.

Cardiovascular disease

Many papers considered the endpoint of cardiovascular disease (CVD), generally defined as coronary heart disease (CHD) plus cerebrovascular disease; that is: myocardial infarction, coronary heart disease, stroke, and transient ischaemic attacks. Most presented separate results for women and men; most used established risk prediction equations or based their models on such models, and as a result most reported D values were reasonably similar in magnitude.

Hippisley-Cox et al. (2007) derived and validated the QRISK score (on two cohorts obtained by splitting the same general practice dataset), and found $D_F = 1.55$ and $D_M = 1.45$. Two papers (Collins and Altman, 2009; Hippisley-Cox et al., 2008a) externally validated QRISK, both using the same validation dataset, unsurprisingly they obtained the same results: $D_F = 1.56$ and $D_M = 1.39$.

Hippisley-Cox et al. (2008b) derived a successor to QRISK, QRISK2, splitting a large dataset into development and validation cohorts. In the validation cohort QRISK2 gave $D_F = 1.80$ and $D_M = 1.62$. Another paper externally validated QRISK2, finding $D_F = 1.66$ and $D_M = 1.45$ (Collins and Altman, 2010).

Yet another QRISK score, this time based on lifetime risk of CVD, was developed by the same group, finding $D_F = 1.93$ and $D_M = 1.79$ (Hippisley-Cox et al., 2010).

Other papers reported $D_F = 2.01$ (Wood and Greenland, 2009), $D_F = 1.95$ (Cook et al., 2006), $D_F = 1.58$ (de la Iglesia et al., 2011), $D_F = 2.34$ (Bozorgmanesh et al., 2010), $D_F = 1.62$ (Rutten et al., 2010), $D_F = 1.90$ (Paynter et al., 2010); and $D_M = 1.35$ (de la Iglesia et al., 2011), $D_M = 1.81$ (Bozorgmanesh et al., 2010), $D_M = 1.14$ (Rutten et al., 2010) and $D_M = 1.25$ (Araujo et al., 2010).

A few papers reported quite different values for D. Rubinshtein et al. (2010) considered the value of imaging in predicting cardiovascular events (adjusted for Framingham Risk Score) and found $D = 0.50$. Paolo et al. (2010) used a slightly modified endpoint, excluding stroke and transient ischaemic attack from the definition of cardiovascular event, and found $D = 0.85$.

Cardiac / cardiovascular death

Two papers used the endpoint CVD mortality which is death from cardiovascular disease; that is, from coronary heart disease (CHD) or cerebrovascular disease. The best model in Weiss et al. (2010) reported $D = 2.17$. Sehestedt et al. (2010) looked at whether adding markers of subclinical organ damage to the SCORE model improved risk prediction, finding a best model with $D = 2.10$. Hurley et al. (2010) considered different ethnic groups and found $D = 2.04$ for non-hispanic whites, $D = 1.81$ for non-hispanic blacks, and $D = 1.67$ for Mexican-Americans.

Heart failure

Two papers looked at this endpoint and both used the Health ABC model. Butler et al. (2008) reported $D_F = 1.80$ and $D_M = 1.54$, while Kalogeropoulos et al. (2010) reported $D_F = 1.49$, $D_M = 1.28$, and $D = 1.39$ for both sexes.

The other paper in this category (Rutten et al., 2010) used a model of traditional risk factors and biomarkers and found $D_F = 2.02$ and $D_M = 1.74$.

Coronary heart disease (CHD) events

These events generally included definite or probable myocardial infarction (MI), silent MI indicated by electrocardiograms, definite CHD death, or coronary revascularisation procedures.

Selvin et al. (2010) included traditional risk factors only and found $D = 1.48$. Several papers sought to increase prognostic ability by adding new factors to traditional risk factors. Wood and Greenland (2009) and Polonsky et al. (2010) both added coronary calcium scores and found $D = 2.10$ and $D = 2.01$ respectively. Nambi et al. (2010) included measures of arterial thickness and found $D_F = 1.69$, $D_M = 1.14$, and $D = 1.57$ for all sexes. Rutten et al. (2010) added protein biomarkers and found $D_F = 1.62$ and $D_M = 1.12$. Rodondi et al. (2010) added a different selection of biomarkers and found $D = 0.76$.

CHD death

Thesis datasets WHI2, WHI3, and WHI4 all sought models predicting CHD death in male civil servants at three different job grades. They found $D = 1.36$, $D = 1.21$ and $D = 0.77$ respectively.

Stroke

Two papers attempted to predict occurrence of stroke. One used traditional risk factors and found $D = 1.63$ (Selvin et al., 2010); the other added protein biomarkers and found $D_F = 1.42$ and $D_M = 1.28$ (Rutten et al., 2010).

Patients with hypertension

Two papers reported models for hypertensive patients without other pre-existing CV conditions. Nelson et al. (2010) were concerned with predicting various outcomes for a mixed sex group of elderly patients with hypertension. For most outcomes the Framingham Risk Score produced the highest D (although values of D reported were very low): $D = 0.06$ for stroke; $D = 0.44$ for myocardial infarction; $D = 0.17$ for CHD death; $D = 0.39$ for CHD events; and $D = 0.62$ for CV disease. For cardiac / CVD death the Pocock Algorithm was the best model, with $D = 0.44$.

Weiss et al. (2010) also considered a group of hypertensive patients of both genders and found $D = 1.86$ for predicting cardiac death.

E.1.2 Incidence of diabetes

There were three papers in this category, looking at the risk of diabetes in apparently healthy subjects. One prospectively derived and validated the QDScore, a new diabetes risk algorithm for estimating the 10 year risk of acquiring type 2 diabetes (Hippisley-Cox et al., 2009). The model was developed using a very large general practice dataset which was split into development and validation datasets. The paper reported $D_F = 2.11$ and $D_M = 1.97$. The second paper externally validated the QDScore in a different general practice dataset and found $D_F = 1.83$ and $D_M = 1.76$ (Collins and Altman, 2011). Additionally, Selvin et al. (2010) reported a variety of endpoints including 6- and 15- year incident diabetes ($D = 2.56$ and $D = 1.62$ respectively).

E.1.3 Incidence of bone fracture

Four papers considered models for incidence of fracture. Two looked at both hip and all osteoporotic fractures; one only looked at hip fracture, and one just at any fracture (excluding skull, fingers & toes). The first hip fracture paper reported $D_F = 2.73$ and $D_M = 2.68$ (Hippisley-Cox and Coupland, 2009), the second reported $D = 1.77$ (Moayyeri et al., 2009). The latter paper also looked at how well the model predicted osteoporotic fracture of the vertebrae, radius or hip; for this endpoint $D_F = 1.85$ and $D_M = 1.34$.

Collins et al. (2011) externally validated the results of Hippisley-Cox and Coupland (2009) and found similar results: for hip fracture $D_F = 2.66$ and $D_M = 2.53$. For any osteoporotic fracture, they found $D_F = 2.02$ and $D_M = 1.60$.

The single paper looking at any fracture reported $D = 1.81$ (Kaptoge et al., 2008).

E.1.4 Chronic kidney disease

Hippisley-Cox and Coupland (2010b) used large general practice datasets to develop and validate a risk prediction model for chronic kidney disease. Their final model showed $D_F = 2.32$ and $D_M = 2.38$.

E.1.5 Predicting side effects of statins

One paper sought to develop a model to predict the side effects of statins which users were likely to experience, from a large general practice dataset. Hippisley-Cox and Cou-

pland (2010a) developed different models for various side effects separately for men and women and their best models for predicting each were: acute renal failure, $D_F = 2.49$, $D_M = 2.49$; cataracts, $D_F = 2.46$, $D_M = 2.48$; myopathy, $D_F = 1.75$, $D_M = 1.53$; liver dysfunction, $D_F = 0.89$, $D_M = 0.71$.

E.1.6 Colorectal / colon cancer

Ma et al. (2010) sought to develop and validate a model to predict 10 year risk of colorectal, colon only and rectal only cancer in Japanese men. Their best models showed $D = 1.20$ for colorectal cancer, $D = 1.25$ for colon cancer and $D = 1.17$ for rectal cancer.

E.1.7 Dental caries

One paper looked at prediction of dental caries in adolescents undergoing orthodontic treatment. They found a best model with $D = 1.32$ Chaussain et al. (2010).

E.2 Prognostic models

This section covers papers which developed and / or validated models in patients with a disease, and attempt to predict some disease-related event, or death – these are 'true' prognostic models. In these sections the subsection headings refer to the condition or disease which the patients had. The endpoint for each study is described in the text.

E.2.1 Cardiac patients

13 papers were found which considered survival models in patients after cardiovascular (CV) disease or intervention of some description.

Heart failure

Five papers in total analysed datasets of patients with varying degrees of heart failure.

Guazzi et al. (2010) included patients with 'stable heart failure', and aimed to develop a cardiopulmonary exercise prognostic score to predict cardiac death. Their model did not better the current best prognostic model, which showed $D = 1.54$ in this group.

Advanced heart failure Two papers included patients with advanced heart failure. The first sought to establish the additional value of echocardiography over the Seattle Heart Failure Model (SHFM), which predicts death, urgent cardiac transplantation or left ventricular assist device (LVAD) support (Agha et al., 2009). The authors found that adding echocardiography improved prognostic ability of the model, with $D = 2.65$ in this dataset. The second paper used the same endpoints, and aimed to validate the SHFM in patients with advanced heart failure, estimating $D = 1.38$ (Kalogeropoulos et al., 2009). They also looked at the value of the model in predicting death alone, which had $D = 1.39$.

Acute heart failure The first of two papers considering patients with acute heart failure looked at the additional prognostic value of uric acid for predicting all-cause mortality and found $D = 1.81$ (Alimonda et al., 2009). The second considered the value of carbohydrate antigen and used the same endpoint; $D = 2.07$ (Núñez et al., 2007).

Chronic heart failure One publication involved patients with chronic heart failure, using an endpoint of cardiac death. They considered the value of serial vs one-time imaging and found $D = 1.73$ (Kasama et al., 2010).

After CABG

Two papers considered the utility of various cardiac biomarkers in predicting all cause mortality in patients following coronary artery bypass graft (CABG). The first considered a fatty acid protein and found the best model to be $D = 1.71$ (Muehlschlegel et al., 2010b). The second found that adding a chromosome variant to the model improved prediction, $D = 1.78$ (Muehlschlegel et al., 2010a).

Pulmonary arterial hypertension

Benza et al. (2010) included patients with pulmonary arterial hypertension with endpoint all-cause mortality, finding the best model had $D = 1.49$.

After percutaneous coronary intervention

Two papers looked at all cause mortality in patients after a percutaneous coronary intervention procedure. Damman et al. (2011) looked exclusively at post-MI patients under-

going the procedure and found $D = 2.01$ when additional biomarkers were used. Singh et al. (2010) looked more generally at patients having this procedure (39% had previously had an MI) and found $D = 1.82$. Additionally this paper considered another endpoint of all-cause mortality plus MI and found $D = 1.37$.

Coronary heart disease

Benderly et al. (2010) included patients with CHD and looked at the endpoint of all-cause mortality, finding the best model had $D = 1.05$.

Ischaemic stroke

One paper tried to predict the time to recurrent stroke in patients who had a first ischaemic stroke. Their best model was based on clinical and imaging parameters and in the validation dataset showed $D = 1.61$ (Ay et al., 2010).

MI

The thesis dataset HOS included post-MI patients and predicted overall survival, with a $D = 1.98$ (model found using MFP with $p = 0.05$).

Atherosclerosis

One of the thesis datasets (STE, the SMART study), recruited patients with clinical manifestations of atherosclerosis and developed a model for predicting fatal and non-fatal vascular events ((non-)fatal ischaemic stroke, (non-)fatal MI, vascular death). The best model found using MFP with $p = 0.05$ had $D = 1.30$.

E.2.2 Cancer

Most of the papers in our literature search were in cancer; 99 in total. We divide discussion by cancer site, and sometimes further by early and advanced disease or other characteristics where appropriate.

Breast Cancer

Mixed group of breast cancer Two papers looked at a wide group of breast cancer patients. Naderi et al. (2006) sought to develop a gene-expression signature to predict OS in breast cancer patients; their best model showed $D_F = 0.92$ and $D_F = 0.59$ in two independent validation datasets. Haibe-Kains et al. (2010) used a fuzzy gene expression-based computational approach to develop a new prognostic model for breast cancer, and the resulting model (GENIUS) predicted distant metastasis or relapse free survival with $D_F = 2.10$ in a group of patients with all types & stages of breast cancer.

Several thesis datasets were breast cancer datasets. The 9 SEER datasets included patients with any type of breast cancer, and the best models for predicting OS found D_F values ranging from $1.59 - 2.07$. The FBC and RBC datasets used the outcome PFS and the best models found $D_F = 1.26$ and $D_F = 1.09$ respectively.

Early breast cancer Three papers in the breast cancer category considered early disease only. The first one sought to predict RFS in ER+ patients and found a best model with $D_F = 1.32$ (externally validated) (Campbell et al., 2010). The second used the endpoint of locoregional recurrence and found a model with $D_F = 0.82$ amongst all patients, $D_F = 0.86$ amongst mastectomy patients and $D_F = 0.62$ amongst patients who had breast-conserving surgery (van Nes et al., 2010). The third searched for a gene-expression score to predict distant metastasis-free survival and their best model showed $D_F = 2.01$ (Sánchez-Navarro et al., 2010).

Hormone-receptor specific breast cancer Two papers concentrated solely on patients with oestrogen receptor positive (ER+) breast cancer. In the first, the authors compared many models, separately in patients with N0 and N+ disease (Nielsen et al., 2010). Amongst N0 disease the best model showed $D_F = 0.98$ for an endpoint of RFS, and $D_F = 1.11$ for cancer-specific survival (CSS). Amongst N+ patients the Adjuvant!Online model performed best with $D_F = 0.62$ for RFS and $D_F = 0.74$ for CSS. In the second paper, the patients were all ER+ and had been treated with tamoxifen. The authors developed a biomarker model to predict RFS ($D_F = 1.85$) and RFS while still on tamoxifen treatment ($D_F = 1.77$) (Baneshi et al., 2010). Additionally, the GENIUS model developed by Haibe-

Kains et al. (2010) showed $D_F = 2.01$ amongst ER+/HER2- patients (endpoint distant metastasis or relapse free survival).

The GENIUS model also showed $D_F = 1.61$ amongst ER-/HER2- patients, and $D_F = 1.93$ amongst HER2+ patients (endpoint distant metastasis or relapse free survival) (Haibe-Kains et al., 2010).

Prostate Cancer

The eight papers reporting on prostate cancer patients fell into various subcategories.

Post - radical prostatectomy Most of the papers concerned patients with organ-confined disease, who had been treated with radical prostatectomy (RP). All the models in these papers sought to predict biochemical recurrence.

The first paper only considered patients with cancer confined to the central zone of the prostate. They found $D_M = 1.91$ when the location of the tumour was taken into account along with other predictors (Cohen et al., 2008). The second aimed to look at the prognostic value of microvessel density and found $D_M = 1.87$ (Erbersdobler et al., 2010). Ahyai et al. (2010) found $D_M = 2.09$, including a term for surgical margin. Cao et al. (2010) looked at whether the Gleason score of the tumour margin was more useful for predicting biochemical recurrence than the Gleason score of the main tumour and found that it was, with $D_M = 1.38$ for all patients and $D_M = 1.20$ for patients with a Gleason score of 7. The next paper looked at biomarkers (chromosome deletions) but found that these did not improve the prognostic ability of their base model ($D_M = 1.98$) (El Gammal et al., 2010). The final paper performed a head-to-head comparison of the three most commonly used preoperative models after RP and found the same model performed best for both 3 and 5 years biochemical recurrence-free survival ($D_M = 1.55$ and $D_M = 1.43$ respectively) (Lughezzani et al., 2010a).

Radiation therapy Williams et al. (2006) included patients who had external-beam radiation as their only treatment. As well as developing a new model to predict biochemical recurrence using recursive partitioning, the authors also sought to evaluate an existing model on two independent datasets, with 864 and 271 events respectively. The model showed $D_M = 1.50$ ($R^2_D = 35\%$) in the smaller dataset and $D_M = 1.02$ (20%) in the larger,

showing that a validated model can produce quite different estimates of D when applied to different datasets. The new model showed $D_M = 1.15$ in the Australian data and $D_M = 1.57$ in the Canadian.

One paper included patients treated with permanent prostate brachytherapy (radiotherapy implants). They developed a postoperative nomogram predicting the 9-year probability of prostate cancer recurrence and found $D_M = 1.25$ (Potters et al., 2010).

Advanced prostate cancer The thesis dataset APC reported on patients with advanced prostate cancer, with the endpoint OS. The best model found using MFP with $p = 0.05$ had $D_M = 0.85$.

Renal Cancer

The papers in this section can be divided by disease stage and/or pathology of the patients involved.

All pathology types These papers included all pathology types and also a mix of disease stages (including 80–85% non-metastatic patients). Isbarn et al. (2010) used disease-specific survival (DSS) as endpoint and found $D = 2.48$ for its best model (this was repeated for clear cell patients only; see below). Bigot et al. (2010) also used DSS and found $D = 2.34$ for its best model.

Metastatic One paper reported purely on patients with metastatic disease but with all pathology types, and with the endpoint of OS found $D = 0.72$ for the best model found (Royston et al., 2006).

Thesis paper KCA included patients with metastatic renal cancer and found $D = 1.2$ in a model predicting OS, selected by MFP with $p = 0.05$.

Clear cell pathology Tan et al. (2010) included non-metastatic clear cell patients. They looked at three endpoints and for the best model reported $D = 0.98$ for OS, $D = 1.46$ for DSS and $D = 1.18$ for DFS. Another paper also looked at clear cell patients only, to externally validate a previously published model. In this report a mix of patients were used but only 5% were metastatic; the endpoint was DSS and the authors found $D = 2.13$

(Zigeuner et al., 2010a). Isbarn et al. (2010) repeated their analyses for clear cell patients only and found $D = 2.36$ (endpoint DSS).

Papillary pathology Klatte et al. (2010) reported exclusively on patients with papillary renal cell carcinoma but included all stages of disease. The model they found for DSS gave $D = 3.34$ in the validation cohort. This seems very high and the estimate may include optimism due to a small number of events in the dataset.

Upper Urinary Tract Cancers

All these papers concern patients treated with radical nephro-ureterectomy. The first paper looked at the endpoint of DSS and found a best model with $D = 1.57$ (Jeldres et al., 2010). The next paper looked at RFS and DSS and found best models with $D = 1.90$ and $D = 1.96$ respectively (Raman et al., 2010). This paper also considered the same models in the subgroup of patients with T2 or worse disease and found both endpoints had $D = 1.39$. The final paper in this category again looked at RFS and DSS and found best models with $D = 2.00$ and $D = 2.05$ respectively (Zigeuner et al., 2010b).

One paper looked at OS and DSS, however instead of multivariate models this paper considered only the predictive power of age, and so cannot really be compared with the other models in this category. Prediction was consequently worse, with $D = 0.51$ for OS and $D = 0.33$ for DSS (Shariat et al., 2010c).

Bladder Cancers

The papers concerned with bladder cancer fall into two broad camps; those in non-muscle invasive cancer (tumour stage T_{IS} or T_{1a}), and those in patients with higher grade tumours (T_1 or worse).

Non-muscle invasive cancer The first paper in this category used the same model to predict recurrence-free survival (RFS) ($D = 0.92$), PFS ($D = 1.85$) and DSS ($D = 2.56$) (Pan et al., 2010). The second paper found their model had poor performance in RFS ($D = 0.39$) but was better at predicting PFS ($D = 0.98$) (Yamada et al., 2010).

Higher grade bladder cancer The first paper in this category looked at DSS only and found a best model with $D = 1.35$; they also looked at the model in patients with T2

or worse disease, and found $D = 1.40$ (Svatek et al., 2010). Another paper considered disease recurrence in various sites and found $D = 1.08$ for predicting recurrence in the upper urinary tract, and $D = 1.30$ for predicting recurrence in bone (Umbreit et al., 2010). Shariat et al. (2010d) looked at 2 models in node negative, node positive or all patients; for predicting both DFS and DSS. Amongst all patients, the best model for DFS had a $D = 1.40$, and the best for DSS had $D = 1.74$. The node positive patients generally showed lower D than node negative patients, and the models were better at predicting DSS than DFS in all cases. Shariat et al. (2010e) considered DFS and DSS, the best models for predicting these endpoints had $D = 1.48$ and $D = 1.42$ respectively. Shariat et al. (2010a) considered RFS and DSS with the best models having $D = 0.96$ and $D = 1.05$ respectively. The last paper again considered RFS and DSS and found best models having $D = 1.17$ and $D = 1.26$ respectively (Shariat et al., 2010b).

Liver Cancer

Two papers considered hepatocellular carcinoma, both in the palliative setting. The first paper assessed different scoring systems and found the best had $D = 1.01$ (Collette et al., 2008). The second paper sought to externally validate the results of the first paper and interestingly found discrepancies in D for the same scoring systems (Tournoux-Facon et al., 2011). For example the Okuda model showed $D = 1.01$ in the first paper and $D = 0.44$ in the second; the BCLC model had $D = 0.79$ and $D = 0.53$ respectively; while the CLIP score showed more consistency with $D = 0.81$ and $D = 0.78$. The best model found in Tournoux-Facon et al. (2011) had $D = 1.01$.

Pancreatic Neuro Endocrine Tumours

Two papers from the search considered this relatively rare cancer. One used SEER data to develop a staging system to predict OS, and found a D of 1.69 (Martin et al., 2011). The other attempted to validate a proposed TNM staging system and add extra factors, they found a best model for predicting DSS which had $D = 2.10$ (but note low sample size of $e = 85$) (Scarpa et al., 2010).

Gynaecological Cancer

In this category we found three papers each including different diseases in the gynaecological area. One paper in endometrial cancer considered OS as its endpoint and found a best model with $D = 1.51$ (Abu-Rustum et al., 2010). (Zivanovic et al., 2009) considered patients with uterine leiomyosarcoma and found the AJCC STS staging system provided the best prediction for OS ($D = 0.74$) and PFS ($D = 0.56$). A paper in advanced ovarian cancer considered only 30 day postoperative morbidity and found $D = 1.05$ (Gerestein et al., 2010). Additionally, thesis dataset OVA included patients with advanced ovarian cancer, and found the best model for OS (using MFP with $p = 0.05$) had $D = 0.91$.

Lung Cancer

The first paper in this category aimed to validate previously published pre- and post-operative models in non-small cell lung cancer (NSCLC) patients (van der Pijl et al., 2010). The pre-operative models showed $D = 1.05, 1.05, 0.92$ for 1, 2 and 3 year OS and the equivalent values for the post-operative model were $D = 1.32, 1.61$ and 1.68 (this model included factors relating to type of resection and pathological stage).

The second paper in this category included all types of lung cancer and aimed to externally validate the 7th edition TNM system for predicting OS (Strand et al., 2010). They found D values of 1.05 for all patients and 0.92 for NSCLC patients, which is comparable to the results of the pre-operative model in the first paper. Additionally, thesis dataset LVA included patients with any type of lung cancer; the best model for OS found using MFP with $\alpha = 0.05$ had $D = 1.48$.

Finally, Nowak et al. (2010) considered patients with mesothelioma. They developed models (incorporating PET imaging) for predicting OS in different patient subgroups. The best models for the various subgroups were: sarcomatoid pathology $D = 0.68$, non-sarcomatoid pleurodesis $D = 1.25$, nonsarcomatoid nonpleurodesis $D = 0.98$, and finally a model in all patients showed $D = 0.87$.

Advanced cancer

Three papers covered advanced cancer with the endpoint of overall survival (OS), but each included different selections of disease areas, which is probably why the three showed

quite different results. The best models found in the three papers were respectively $D = 0.75$ (Chow et al., 2009), $D = 1.27$ (Trédan et al., 2011) and $D = 2.66$ (Martin et al., 2010).

Stomach cancer

Two papers covered this type of cancer, however they may be of different enough types that they shouldn't be considered together. Firstly Woodall et al. (2009) found a prediction model for overall survival in gastro-intestinal stromal tumours (GIST) with $D = 1.27$. The second paper (Wang et al., 2009), looking to predict disease-specific survival in patients with gastric cancer, showed a best model with $D = 1.56$.

Head and neck cancer

The tumours covered under this broad umbrella category are probably disparate. One paper (Dorward et al., 2010) looking at pediatric astrocytomas (a brain tumour) found $D = 1.77$ for their best model predicting recurrence free survival. Another paper (van der Schroeff et al., 2010) looking at models to predict overall survival (OS) in salivary gland carcinoma patients, found $D = 1.69$ when pre-treatment factors were used, and $D = 1.77$ with post-treatment factors. Thesis dataset GLI included patients with malignant glioma; the best model for OS found using MFP with $p = 0.05$ had $D = 1.15$.

Leukemia

Three papers reported on outcomes in leukemia.

Acute lymphoblastic leukaemia (ALL) De Lorenzo et al. (2009) found $D = 1.04$ for their model predicting event-free survival (EFS).

Chronic lymphocytic leukemia (CLL) One paper (Molica et al., 2010) included patients with early CLL and looked at the endpoint of time to first treatment, finding $D = 2.47$.

Chronic myeloid leukemia (CML) Dickinson et al. (2010) included CML patients after stem cell transplant, with OS as the endpoint. The best model found reported $D = 0.98$.

Lymphoma

Two papers and one of the thesis datasets involved lymphoma patients.

Diffuse large B-cell lymphoma Bari et al. (2010) considered the endpoint of OS and found $D = 1.25$.

Follicular lymphoma. One paper and one of the thesis datasets concerned follicular lymphoma. Arcaini et al. (2010) used the endpoint of progression free survival (PFS) and reported $D = 0.80$. The best model for OS found with thesis dataset FOL (using MFP with $p = 0.05$) had $D = 1.23$.

Myeloma

The thesis dataset MYE consists of data from two trials in patients with myeloma. The best model (found using MFP with $\alpha = 0.05$) had $D = 0.77$.

IgG and IgA Monoclonal Gammopathies Rossi et al. (2009) used the endpoint PFS, finding $D = 1.37$.

Asymptomatic Multiple Myeloma Rossi et al. (2010) included patients with asymptomatic multiple myeloma and attempted to predict time to symptoms. $D = 1.61$.

Other Cancers

Various other cancers not falling into any of the previously mentioned broad categories were seen in just one single paper. The best models in these disease areas are given below for interest.

Melanoma Endpoint OS. $D = 0.50$ (Ben-Porat et al., 2006)

Germ Cell Tumour Patients with metastatic disease, endpoint PFS. $D = 0.93$ (Lorch et al., 2010)

Colorectal Cancer Patients with liver metastases, endpoint OS. $D = 0.44$ (Nathan et al., 2010)

Soft Tissues Sarcoma Primary retroperitoneal sarcoma, endpoint OS. $D = 1.61$ (Ardoino et al., 2010)

Adrenal tumour Adreno-cortical carcinoma, endpoint disease-specific survival (DSS). $D = 2.49$ (Lughezzani et al., 2010b)

E.2.3 HIV

Two papers report models in HIV patients; with slightly different endpoints. The first paper used the endpoint of all cause mortality within the first year of antiretroviral treatment; its best model had $D = 1.30$ (May et al., 2010). The second paper used the endpoint of new AIDS event or death (any cause) and reported $D = 1.51$ (May et al., 2004).

E.2.4 Liver disease

Primary biliary cirrhosis

Two thesis datasets included patients with primary biliary cirrhosis and both used OS as the endpoint. For the best models found using MFP with $p = 0.05$, PBC had $D = 2.70$ and PBC2 had $D = 2.55$.

Liver transplant

Aloia et al. (2010) include post-liver transplant patients and aimed to develop a model for prediction of OS. Their best model reported $D = 1.46$.

Hepatitis B

Yang et al. (2010) aimed to develop a model for predicting hepatocellular carcinoma in patients with hepatitis B. The best model found had $D = 2.32$.

E.2.5 Respiratory disease

Moran et al. (2008) included patients with acute lung injury (ALI) and acute respiratory distress syndrome (ARDS), and developed a model for predicting time to death within 28 days. The best model showed $D = 2.24$.

Swallow et al. (2007) considered patients with chronic obstructive pulmonary disorder (COPD) and used the outcomes death or lung transplant, finding $D = 1.02$.

E.2.6 Chagas disease

One paper (Lima-Costa et al., 2010) attempted to develop a model for OS in Chagas disease; the best model found had $D = 1.24$.

E.2.7 Raynaud's phenomenon

Ingegnoli et al. (2010) considered time to systemic sclerosis in patients with this condition. $D = 2.47$ was the best model found.

E.2.8 Epilepsy

Patients with early epilepsy were included in one paper, which aimed to predict time to next seizure. $D = 0.77$ (Kim et al., 2006)

E.2.9 Leg ulcer

A model predicting time to complete healing was developed on thesis dataset LVA, which included patients with leg ulcers. The best model had $D = 2.07$.

Bibliography

Altman D.G., 2009. Prognostic models: a methodological framework and review of models for breast cancer. *Cancer Investigation*, 27(3):235–243.

Altman D.G., Lyman G.H., 1998. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Research and Treatment*, 52(1-3):289–303.

Altman D.G., Vergouwe Y., Royston P., Moons K.G., 2009. Prognosis and prognostic research: validating a prognostic model. *BMJ*, 338:1432–1435.

Armitage P., Berry G., Matthews J.N.S., 2001. *Statistical Methods in Medical Research*. Blackwell Science, 4th edition.

Bender R., Augustin T., Blettner M., 2005. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

Bernardo M.V.P., Lipsitz S.R., Harrington D.P., Catalano P.J., 2000. Sample size calculations for failure time random variables in non-randomized studies. *Journal of the Royal Statistical Society, (Series D): The Statistician*, 49(1):31–40.

Burton A., Altman D.G., Royston P., Holder R.L., 2006. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292.

Byar D.P., Green S.B., 1980. The choice of treatment for cancer patients based on covariate information: application to prostate cancer. *Bulletin du Cancer*, 67:477–490.

Casella G., Berger R.L., 2001. *Statistical Inference*. Duxbury Press, 2 edition.

Choodari-Oskooei B., 2008. Summarising predictive ability of a survival model and applications in medical research. Ph.D. thesis, University College London.

Choodari-Oskooei B., Royston P., Parmar M.K., 2011. A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine*, pages doi: 10.1002+.

Christensen E., Neuberger J., Crowe J., Altman D.G., Popper H., Portmann B., Doniach D., Ranek L., Tygstrup N., Williams R., 1985. Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis. Final results of an international trial. *Gastroenterology*, 89(5):1084–1091.

Cianfrocca M., Goldstein L.J., 2004. Prognostic and predictive factors in early-stage breast cancer. *The Oncologist*, 9(6):606–616.

Concato J., Peduzzi P., Holford T.R., Feinstein A.R., 1995. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. *Journal of Clinical Epidemiology*, 48(12):1495–1501.

Copas J.B., 1983. Regression, Prediction and Shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):311–354.

Copas J.B., Long T., 1991. Estimating the residual variance in orthogonal regression with variable selection. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 40(1):51–59.

D'Agostino R.B., Vasan R.S., Pencina M.J., Wolf P.A., Cobain M., Massaro J.M., Kannel W.B., 2008. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, 117(6):743–753.

Demidenko E., 2007. Sample size determination for logistic regression revisited. *Statistics in Medicine*, 26(18):3385–3397.

Efron B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331.

Fleming T.R., Harrington D.P., 1991. *Counting Processes and Survival Analysis*. Wiley, 1st edition.

Freedman L.S., 1982. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine*, 1(2):121–129.

Gasparini G., Pozza F., Harris A.L., 1993. Evaluating the potential usefulness of new prognostic and predictive indicators in node-negative breast cancer patients. *Journal of the National Cancer Institute*, 85(15):1206–1219.

Gonen M., Heller G., 2005. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970.

Graf E., Schmoor C., Sauerbrei W., Schumacher M., 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Harrell F.E., 2001. *Regression Modeling Strategies*. Springer, 1st edition.

Harrell F.E., Lee K.L., Califf R.M., Pryor D.B., Rosati R.A., 1984. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 3(2):143–152.

Harrell F.E., Lee K.L., Mark D.B., 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

Henderson I.C., Patek A.J., 1998. The relationship between prognostic and predictive factors in the management of breast cancer. *Breast Cancer Research and Treatment*, 52(1-3):261–288.

Hermanek P., 1999. Prognostic factor research in oncology. *Journal of Clinical Epidemiology*, 52(4):371–374.

Hosmer D.W., Lemeshow S., 2000. *Applied Logistic Regression (Wiley Series in Probability and Statistics)*. Wiley-Interscience Publication, 2 edition.

Hosmer D.W., Lemeshow S., May S., 2008. *Applied Survival Analysis: Regression Modeling of Time to Event Data (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2nd edition.

Hsieh F., Lavori P.W., 2000. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials*, 21(6):552–560.

Hsieh F.Y., Bloch D.A., Larsen M.D., 1998. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, 17(14):1623–1634.

Kalbfleisch J.D., Prentice R.L., 1980. *The Statistical Analysis of Failure Time Data (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 1st edition.

Karapetis C.S., Khambata-Ford S., Jonker D.J., O'Callaghan C.J., Tu D., Tebbutt N.C., Simes R.J., Chalchal H., Shapiro J.D., Robitaille S., Price T.J., Shepherd L., Au H.J., Langer C., Moore M.J., Zalcberg J.R., 2008. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine*, 359(17):1757–1765.

Kent J.T., O'Quigley J., 1988. Measures of dependence for censored survival data. *Biometrika*, 75(3):525–534.

Lakatos E., Lan K.K.G., 1992. A comparison of sample size methods for the logrank statistic. *Statistics in Medicine*, 11(2):179–191.

MacLennan I.C.M., Kelly K., Crockson R.A., Cooper E.H., Cuzick J., Chapman C., 1988. Results of the MRC myelomatosis trials for patients entered since 1980. *Hematological Oncology*, 6(2):145–158.

Mahambrey T., Fowler R., Pinto R., Smith T., Callum J., Pisani N., Rizoli S., Adhikari N., 2009. Early massive transfusion in trauma patients: Canadian single-centre retrospective cohort study. *Canadian Journal of Anesthesia / Journal Canadien d'Anesthésie*, 56(10):740–750.

Mallett S., Royston P., Dutton S., Waters R., Altman D.G., 2010. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Medicine*, 8:20+.

Marmot M.G., Rose G., Shipley M., Hamilton P.J., 1978. Employment grade and coronary heart disease in British civil servants. *Journal of Epidemiology and Community Health*, 32(4):244–249.

McGuire W.L., 1991. Breast cancer prognostic factors: evaluation guidelines. *Journal of the National Cancer Institute*, 83(3):154–155.

McShane L.M., Altman D.G., Sauerbrei W., Taube S.E., Gion M., Clark G.M., Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics, 2005. REporting recommendations for tumor MARKer prognostic studies (REMARK). *Nature Clinical Practice Oncology*, 2(8):416–422.

Moher D., Schulz K., Altman D., 2001. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Medical Research Methodology*, 1(1):2+.

Moons K.G., Royston P., Vergouwe Y., Grobbee D.E., Altman D.G., 2009. Prognosis and prognostic research: what, why, and how? *BMJ*, 338:1317–1320.

O'Quigley J., Flandre P., 1994. Predictive capability of proportional hazards regression. *Proceedings of the National Academy of Sciences of the United States of America*, 91(6):2310–2314.

O'Quigley J., Xu R., 2001. *Handbook of statistics in clinical oncology*. Marcel Dekker, New York.

O'Quigley J., Xu R., Stare J., 2005. Explained randomness in proportional hazards models. *Statistics in Medicine*, 24(3):479–489.

Peduzzi P., Concato J., Feinstein A.R., Holford T.R., 1995. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology*, 48(12):1503–1510.

Perel P., Edwards P., Wentz R., Roberts I., 2006. Systematic review of prognostic models in traumatic brain injury. *BMC Medical Informatics and Decision Making*, 6(1):38+.

Putman L.M., van Gameren M., Meijboom F.J., de Jong P.L., Roos-Hesselink J.W., Witsenburg M., Takkenberg J.J.M., Bogers A.J.J.C., 2009. Seventeen years of adult congenital heart surgery: a single centre experience. *European Journal of Cardio-Thoracic Surgery*, 36(1):96–104.

Royston P., 2006a. Explained variation for survival models. *The Stata Journal*, 6(1):1–14.

Royston P., 2006b. Explained variation for survival models. *The Stata Journal*, 6(1):83–96.

Royston P., Moons K.G., Altman D.G., Vergouwe Y., 2009. Prognosis and prognostic research: Developing a prognostic model. *BMJ*, 338:1373–1377.

Royston P., Parmar M.K.B., 2002. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197.

Royston P., Parmar M.K.B., Sylvester R., 2004. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Statistics in Medicine*, 23(6):907–926.

Royston P., Sauerbrei W., 2004. A new measure of prognostic separation in survival data. *Statistics in Medicine*, 23(5):723–748.

Royston P., Sauerbrei W., 2008. *Multivariable Model - Building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables (Wiley Series in Probability and Statistics)*. Wiley, 1st edition.

Sauerbrei W., Hübner K., Schmoor C., Schumacher M., 1997. Validation of existing and development of new prognostic classification schemes in node negative breast cancer. *Breast Cancer Research and Treatment*, 42(2):149–163.

Sauerbrei W., Royston P., 1999. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162(1):71–94.

Sauerbrei W., Schumacher M., 1992. A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine*, 11(16):2093–2109.

Schemper M., Henderson R., 2000. Predictive accuracy and explained variation in Cox regression. *Biometrics*, 56(1):249–255.

Schemper M., Stare J., 1996. Explained variation in survival analysis. *Statistics in Medicine*, 15(19):1999–2012.

Schmoor C., Sauerbrei W., Schumacher M., 2000. Sample size considerations for the evaluation of prognostic factors in survival analysis. *Statistics in Medicine*, 19(4):441–452.

Schoenfeld D., Borenstein M., 2005. Calculating the power or sample size for the logistic and proportional hazards models. *Journal of Statistical Computation and Simulation*, 75(10):771–785.

Schoenfeld D.A., 1983. Sample-size formula for the proportional-hazards regression model. *Biometrics*, 39(2):499–503.

SEER, 2000. Surveillance, Epidemiology, and End Results Program (SEER), 2000.

Simon R., Altman D.G., 1994. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer*, 69(6):979–985.

Smith J.M., Dore C., Charlett A., Lewis J.D., 1992a. A randomised trial of biofilm dressing for venous leg ulcer. *Phlebology*, 7:107–113.

Smith L.R., Harrell F.E., Muhlbaier L.H., 1992b. Problems and potentials in modeling survival. In M.L. Grady, H.A. Schwartz, editors, *Medical Effectiveness Research Data Methods (Summary Report) AHCPR publication, no. 92-0056*, pages 151–159. US Dept of Health and Human Services, Agency for Health Care Policy and Research.

StataCorp, 2000. StataCorp. 2007. Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

Steyerberg E.W., 2008. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health)*. Springer, 1st edition.

Valsecchi M.G., Silvestri D., Sasieni P., 1996. Evaluation of long-term survival: use of diagnostics and robust estimators with Cox's proportional hazards model. *Statistics in Medicine*, 15(24):2763–2780.

Vittinghoff E., McCulloch C.E., 2007. Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, 165(6):710–718.

Volinsky C.T., Raftery A.E., 2000. Bayesian Information Criterion for Censored Survival Models. *Biometrics*, 56(1):256–262.

White I., 2011. Personal Communication.

White I.R., 2010. simsum: Analyses of simulation studies including Monte Carlo error. *The Stata Journal*, 10(3):369–385.

Xu R., O'Quigley J., 1999. A R^2 type measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics*, 12(1):89–107.

D Library Bibliography

Abu-Rustum N.R., Zhou Q., Gomez J.D., Alektiar K.M., Hensley M.L., Soslow R.A., Levine D.A., Chi D.S., Barakat R.R., Iasonos A., 2010. A nomogram for predicting overall survival of women with endometrial cancer following primary therapy: toward improving individualized cancer care. *Gynecologic Oncology*, 116(3):399–403.

Agha S.A., Kalogeropoulos A.P., Shih J., Georgiopoulou V.V., Giamouzis G., Anarado P., Mangalat D., Hussain I., Book W., Laskar S., Smith A.L., Martin R., Butler J., 2009. Echocardiography and risk prediction in advanced heart failure: incremental value over clinical markers. *Journal of Cardiac Failure*, 15(7):586–592.

Ahyai S.A., Zacharias M., Isbarn H., Steuber T., Eichelberg C., Köllermann J., Fisch M., Karakiewicz P.I., Huland H., Graefen M., Chun F.K.H., 2010. Prognostic significance of a positive surgical margin in pathologically organ-confined prostate cancer. *BJU International*, 106(4):478–483.

Alimonda A.L., Núñez J., Núñez E., Husser O., Sanchis J., Bodí V., Miñana G., Robles R., Mainar L., Merlos P., 2009. Hyperuricemia in acute heart failure. More than a simple spectator? *European Journal of Internal Medicine*, 20(1):74–79.

Aloia T.A., Knight R., Gaber A.O., Ghobrial R.M., Goss J.A., 2010. Analysis of liver transplant outcomes for United Network for Organ Sharing recipients 60 years old or older identifies multiple model for end-stage liver disease âĂŞ independent prognostic factors. *Liver Transplant*, 16(8):950–959.

Araujo A.B., Hall S.A., Ganz P., Chiu G.R., Rosen R.C., Kupelian V., Travison T.G., McKinlay J.B., 2010. Does erectile dysfunction contribute to cardiovascular disease risk prediction beyond the Framingham risk score? *Journal of the American College of Cardiology*, 55(4):350–356.

Arcaini L., Merli M., Passamonti F., Rizzi S., Ferretti V., Rattotti S., Pascutto C., Paulli M., Lazzarino M., 2010. Validation of follicular lymphoma international prognostic index 2 (FLIPI2) score in an independent series of follicular lymphoma patients. *British journal of haematology*, 149(3):455–457.

Ardoino I., Miceli R., Berselli M., Mariani L., Biganzoli E., Fiore M., Collini P., Stacchiotti S., Casali P.G.G., Gronchi A., 2010. Histology-specific nomogram for primary retroperitoneal soft tissue sarcoma. *Cancer*, 116(10):2429–2436.

Ay H., Gungor L., Arsava E.M., Rosand J., Vangel M., Benner T., Schwamm L.H., Furie K.L., Koroshetz W.J., Sorensen A.G., 2010. A score to predict early risk of recurrence after ischemic stroke. *Neurology*, 74(2):128–135.

Baneshi M.R., Warner P., Anderson N., Edwards J., Cooke T.G., Bartlett J.M., 2010. Tamoxifen resistance in early breast cancer: statistical modelling of tissue markers to improve risk prediction. *British Journal of Cancer*, 102(10):1503–1510.

Bari A., Marcheselli L., Sacchi S., Marcheselli R., Pozzi S., Ferri P., Balleari E., Musto P., Neri S., Aloe Spiriti M.A., Cox M.C., 2010. Prognostic models for diffuse large B-cell lymphoma in the rituximab era: a never-ending story. *Annals of Oncology*, 21(7):1486–1491.

Ben-Porat L., Panageas K.S., Hanlon C., Patel A., Halpern A., Houghton A.N., Coit D., 2006. Estimates of stage-specific survival are altered by changes in the 2002 American Joint Committee on Cancer staging system for melanoma. *Cancer*, 106(1):163–171.

Benderly M., Boyko V., Goldbourt U., 2010. Relation of body mass index to mortality among men with coronary heart disease. *The American Journal of Cardiology*, 106(3):297–304.

Benza R.L., Miller D.P., Gomberg-Maitland M., Frantz R.P., Foreman A.J., Coffey C.S., Frost A., Barst R.J., Badesch D.B., Elliott C.G., Liou T.G., McGoon M.D., 2010. Predicting survival in pulmonary arterial hypertension: insights from the Registry to Evaluate Early and Long-Term Pulmonary Arterial Hypertension Disease Management (REVEAL). *Circulation*, 122(2):164–172.

Bigot P., Lughezzani G., Karakiewicz P., Perrotte P., Rioux-Leclercq N., Catros-Quemener V., Bouet F., Moulinoux J.P.P., Cipolla B., Jacques J., 2010. The prognostic value of erythrocyte polyamine in the post-nephrectomy stratification of renal cell carcinoma specific mortality. *The Journal of Urology*, 183(2):486–491.

Bozorgmanesh M., Hadaegh F., Azizi F., 2010. Predictive performances of lipid accumulation product vs. adiposity measures for cardiovascular diseases and all-cause mortality, 8.6-year follow-up: Tehran lipid and glucose study. *Lipids in Health and Disease*, 9:100–112.

Butler J., Kalogeropoulos A., Georgiopoulou V., Belue R., Rodondi N., Garcia M., Bauer D.C., Satterfield S., Smith A.L., Vaccarino V., Newman A.B., Harris T.B., Wilson P.W., Kritchevsky S.B., Health ABC Study, 2008. Incident heart failure prediction in the elderly: the health ABC heart failure score. *Circulation: Heart Failure*, 1(2):125–133.

Campbell H.E., Gray A.M., Harris A.L., Briggs A.H., Taylor M.A., 2010. Estimation and external validation of a new prognostic model for predicting recurrence-free survival for early breast cancer patients in the UK. *British Journal of Cancer*, 103(6):776–786.

Cao D., Kibel A.S., Gao F., Tao Y., Humphrey P.A., 2010. The Gleason score of tumor at the margin in radical prostatectomy is predictive of biochemical recurrence. *The American Journal of Surgical Pathology*, 34(7):994–1001.

Chaussain C., Vital S.O., Viallon V., Vermelin L., Haignere C., Sixou M., Lasfargues J.J., 2010. Interest in a new test for caries risk in adolescents undergoing orthodontic treatment. *Clinical Oral Investigations*, 14(2):177–185.

Chow E., Abdolell M., Panzarella T., Harris K., Bezjak A., Warde P., Tannock I., 2009. Recursive partitioning analysis of prognostic factors for survival in patients with advanced cancer. *International Journal of Radiation Oncology Biology Physics*, 73(4):1169–1176.

Cohen R.J., Shannon B.A., Phillips M., Moorin R.E., Wheeler T.M., Garrett K.L., 2008. Central zone carcinoma of the prostate gland: A distinct tumor type with poor prognostic features. *The Journal of Urology*, 179(5):1762–1767.

Collette S., Bonnetain F., Paoletti X., Doffoel M., Bouché O., Raoul J.L., Rougier P., Masskouri F., Bedenne L., Barbare J.C., 2008. Prognosis of advanced hepatocellular carcinoma: comparison of three staging systems in two French clinical trials. *Annals of Oncology*, 19(6):1117–1126.

Collins G.S., Altman D.G., 2009. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ*, 339:b2584+.

Collins G.S., Altman D.G., 2010. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ*, 340:c2442+.

Collins G.S., Altman D.G., 2011. External validation of QDSCORE for predicting the 10-year risk of developing Type 2 diabetes. *Diabetic Medicine*, 28(5):599–607.

Collins G.S., Mallett S., Altman D.G., 2011. Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores. *BMJ*, 342:d3651+.

Cook N.R., Buring J.E., Ridker P.M., 2006. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Annals of Internal Medicine*, 145(1):21–29.

Damman P., Beijk M.A.M., Kuijt W.J., Verouden N.J.W., van Geloven N., Henriques J.P.S., Baan J., Vis M.M., Meuwissen M., van Straalen J.P., Fischer J., Koch K.T., Piek J.J., Tijssen J.G.P., de Winter R.J., 2011. Multiple biomarkers at admission significantly improve the prediction of mortality in patients undergoing primary percutaneous coronary intervention for acute ST-segment elevation myocardial infarction. *Journal of the American College of Cardiology*, 57(1):29–36.

de la Iglesia B., Potter J.F., Poulter N.R., Robins M.M., Skinner J., 2011. Performance of the ASSIGN cardiovascular disease risk score on a UK cohort of patients from general practice. *Heart*, 97(6):491–499.

De Lorenzo P., Antolini L., Valsecchi M.G., 2009. Evaluation of alternative prognostic stratifications by prediction accuracy measures on individual survival with application to childhood leukaemia. *European Journal of Cancer*, 45(8):1432–1437.

Dickinson A.M., Pearce K.F., Norden J., O'Brien S.G., Holler E., Bickeböller H., Balavarca Y., Rocha V., Kolb H.J.J., Hromadnikova I., Sedlacek P., Niederwieser D., Brand R., Ruutu T., Apperley J., Szydlo R., Goulmy E., Siegert W., de Witte T., Gratwohl A., 2010. Impact of genomic risk factors on outcome after hematopoietic stem cell transplantation for patients with chronic myeloid leukemia. *Haematologica*, 95(6):922–927.

Dorward I.G., Luo J., Perry A., Gutmann D.H., Mansur D.B., Rubin J.B., Leonard J.R., 2010. Postoperative imaging surveillance in pediatric pilocytic astrocytomas. *Journal of Neurosurgery: Pediatrics*, 6(4):346–352.

El Gammal A.T., Brüchmann M., Zustin J., Isbarn H., Hellwinkel O.J.C., Köllermann J., Sauter G., Simon R., Wilczak W., Schwarz J., Bokemeyer C., Brümmendorf T.H., Izbicki J.R., Yekebas E., Fisch M., Huland H., Graefen M., Schlomm T., 2010. Chromosome 8p deletions and 8q gains are associated with tumor progression and poor prognosis in prostate cancer. *Clinical Cancer Research*, 16(1):56–64.

Erbersdobler A., Isbarn H., Dix K., Steiner I., Schlomm T., Mirlacher M., Sauter G., Haese A., 2010. Prognostic value of microvessel density in prostate cancer: a tissue microarray study. *World Journal of Urology*, 28(6):687–692.

Gerestein C.G., Boer G.M.N.d., Eijkemans M.J., Kooi G.S., Burger C.W., 2010. Prediction of 30-day morbidity after primary cytoreductive surgery for advanced stage ovarian cancer. *European Journal of Cancer*, 46(1):102–109.

Guazzi M., Boracchi P., Arena R., Myers J., Vicenzi M., Peberdy M.A., Bensimhon D., Chase P., Reina G., 2010. Development of a cardiopulmonary exercise prognostic score for optimizing risk stratification in heart failure: the (P)e(R)i(O)dic (B)reathing During (E)xercise (PROBE) study. *Journal of Cardiac Failure*, 16(10):799–805.

Haibe-Kains B., Desmedt C., Rothé F., Piccart M., Sotiriou C., Bontempi G., 2010. A fuzzy gene expression-based computational approach improves breast cancer prognostication. *Genome Biology*, 11(2):R18+.

Hippisley-Cox J., Coupland C., 2009. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ*, 339:b4229+.

Hippisley-Cox J., Coupland C., 2010a. Individualising the risks of statins in men and women in England and Wales: population-based cohort study. *Heart*, 96(12):939–947.

Hippisley-Cox J., Coupland C., 2010b. Predicting the risk of chronic Kidney Disease in men and women in England and Wales: prospective derivation and external validation of the QKidney Scores. *BMC Family Practice*, 11:49–61.

Hippisley-Cox J., Coupland C., Robson J., Brindle P., 2010. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *BMJ*, 341:c6624+.

Hippisley-Cox J., Coupland C., Robson J., Sheikh A., Brindle P., 2009. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QD-Score. *BMJ*, 338:b880+.

Hippisley-Cox J., Coupland C., Vinogradova Y., Robson J., Brindle P., 2008a. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart (British Cardiac Society)*, 94(1):34–39.

Hippisley-Cox J., Coupland C., Vinogradova Y., Robson J., May M., Brindle P., 2007. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*, 335(7611):136–147.

Hippisley-Cox J., Coupland C., Vinogradova Y., Robson J., Minhas R., Sheikh A., Brindle P., 2008b. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*, 336(7659):1475–1482.

Hurley L.P., Dickinson M.M., Estacio R.O., Steiner J.F., Havranek E.P., 2010. Prediction of cardiovascular death in racial/ethnic minorities using Framingham risk factors. *Circulation: Cardiovascular Quality and Outcomes*, 3(2):181–187.

Ingegnoli F., Boracchi P., Gualtierotti R., Biganzoli E.M., Zeni S., Lubatti C., Fantini F., 2010. Improving outcome prediction of systemic sclerosis from isolated Raynaud's phenomenon: role of autoantibodies and nail-fold capillaroscopy. *Rheumatology*, 49(4):797–805.

Isbarn H., Patard J.J., Lughezzani G., Rioux-Leclercq N., Crépel M., Cindolo L., de la Taille A., Zini L., Villers A., Shariat S.F., Bertini R., Karakiewicz P.I., 2010. Limited prognostic value of tumor necrosis in patients with renal cell carcinoma. *Urology*, 75(6):1378–1384.

Jeldres C., Sun M., Lughezzani G., Isbarn H., Shariat S.F., Widmer H., Graefen M., Montorsi F., Perrotte P., Karakiewicz P.I., 2010. Highly predictive survival nomogram after upper urinary tract urothelial carcinoma. *Cancer*, 116(16):3774–3784.

Kalogeropoulos A., Psaty B.M., Vasan R.S., Georgiopoulou V., Smith A.L., Smith N.L., Kritchevsky S.B., Wilson P.W., Newman A.B., Harris T.B., Butler J., Cardiovascular Health Study, 2010. Validation of the health ABC heart failure model for incident heart failure risk prediction: the Cardiovascular Health Study. *Circulation: Heart Failure*, 3(4):495–502.

Kalogeropoulos A.P., Georgiopoulou V.V., Giamouzis G., Smith A.L., Agha S.A., Waheed S., Laskar S., Puskas J., Dunbar S., Vega D., Levy W.C., Butler J., 2009. Utility of the Seattle Heart Failure Model in patients with advanced heart failure. *Journal of the American College of Cardiology*, 53(4):334–342.

Kaptoge S., Beck T.J., Reeve J., Stone K.L., Hillier T.A., Cauley J.A., Cummings S.R., 2008. Prediction of incident hip fracture risk by femur geometry variables measured by hip structural analysis in the study of osteoporotic fractures. *Journal of Bone and Mineral Research*, 23(12):1892–1904.

Kasama S., Toyama T., Sumino H., Kumakura H., Takayama Y., Minami K., Ichikawa S., Matsumoto N., Sato Y., Kurabayashi M., 2010. Serial cardiac 123I-metaiodobenzylguanidine scintigraphic studies are more useful for predicting cardiac death than one-time scan in patients with chronic heart failure: sub-analysis of our previous report. *Nuclear Medicine Communications*, 31(9):807–813.

Kim L.G., Johnson T.L., Marson A.G., Chadwick D.W., 2006. Prediction of risk of seizure recurrence after a single seizure and early epilepsy: further results from the MESS trial. *The Lancet Neurology*, 5(4):317–322.

Klatte T., Remzi M., Zigeuner R.E., Mannweiler S., Said J.W., Kabbinavar F.F., Haitel A., Waldert M., de Martino M., Marberger M., Belldegrun A.S., Pantuck A.J., 2010. Development and external validation of a nomogram predicting disease specific survival

after nephrectomy for papillary renal cell carcinoma. *The Journal of Urology*, 184(1):53–58.

Lima-Costa M.F., Cesar C.C., Peixoto S.V., Ribeiro A.L., 2010. Plasma Îš-type natriuretic peptide as a predictor of mortality in community-dwelling older adults with Chagas disease: 10-year follow-up of the Bambuí Cohort Study of Aging. *American Journal of Epidemiology*, 172(2):190–196.

Lorch A., Beyer J., Bascoul-Mollevi C., Kramar A., Einhorn L.H., Necchi A., Massard C., De Giorgi U., Fléchon A., Margolin K.A., Lotz J.P., Germa Lluch J.R., Powles T., Kollmannsberger C.K., Group T.I.P.F.S., 2010. Prognostic factors in patients with metastatic germ cell tumors who experienced treatment failure with cisplatin-based first-line chemotherapy. *Journal of Clinical Oncology*, 28(33):4906–4911.

Lughezzani G., Budäus L., Isbarn H., Sun M., Perrotte P., Haese A., Chun F.K., Schlomm T., Steuber T., Heinzer H., Huland H., Montorsi F., Graefen M., Karakiewicz P.I., 2010a. Head-to-head comparison of the three most commonly used preoperative models for prediction of biochemical recurrence after radical prostatectomy. *European Urology*, 57(4):562–568.

Lughezzani G., Sun M., Perrotte P., Jeldres C., Alasker A., Isbarn H., Budäus L., Shariat S.F., Guazzoni G., Montorsi F., Karakiewicz P.I., 2010b. The European Network for the Study of Adrenal Tumors staging system is prognostically superior to the international union against cancer-staging system: a North American validation. *European Journal of Cancer*, 46(4):713–719.

Ma E., Sasazuki S., Iwasaki M., Sawada N., Inoue M., 2010. 10-Year risk of colorectal cancer: Development and validation of a prediction model in middle-aged Japanese men. *Cancer Epidemiology*, 34(5):534–541.

Martin L., Watanabe S., Fainsinger R., Lau F., Ghosh S., Quan H., Atkins M., Fassbender K., Downing G.M., Baracos V., 2010. Prognostic factors in patients with advanced cancer: Use of the patient-generated subjective global assessment in survival prediction. *Journal of Clinical Oncology*, 28(28):4376–4383.

Martin R.C., Kooby D.A., Weber S.M., Merchant N.B., Parikh A.A., Cho C.S., Ahmad S.A., Jin H., Hawkins W., Scoggins C.R., 2011. Analysis of 6,747 pancreatic neuroendocrine tumors for a proposed staging system. *Journal of Gastrointestinal Surgery*, 15(1):175–183.

May M., Boulle A., Phiri S., Messou E., Myer L., Wood R., Keiser O., Sterne J.A.C., Dabis F., Egger M., 2010. Prognosis of patients with HIV-1 infection starting antiretroviral therapy in sub-Saharan Africa: a collaborative analysis of scale-up programmes. *The Lancet*, 376(9739):449–457.

May M., Royston P., Egger M., Justice A.C., Sterne J.A.C., 2004. Development and validation of a prognostic model for survival time data: application to prognosis of HIV positive patients treated with antiretroviral therapy. *Statistics in Medicine*, 23(15):2375–2398.

Moayyeri A., Kaptoge S., Dalzell N., Luben R.N., Wareham N.J., Bingham S., Reeve J., Khaw K.T., 2009. The effect of including quantitative heel ultrasound in models for estimation of 10-year absolute risk of fracture. *Bone*, 45(2):180–184.

Molica S., Digiesi G., Battaglia C., Cutrona G., Antenucci A., Molica M., Giannarelli D., Sperduti I., Gentile M., Morabito F., Ferrarini M., 2010. Baff serum level predicts time to first treatment in early chronic lymphocytic leukemia. *European Journal of Haematology*, 85(4):314–320.

Moran J.L., Bersten A.D., Solomon P.J., Edibam C., Hunt T., Australian T., Group N.Z.I.C.S.C.T., 2008. Modelling survival in acute severe illness: Cox versus accelerated failure time models. *Journal of Evaluation in Clinical Practice*, 14(1):83–93.

Muehlschlegel J.D., Liu K.Y.Y., Perry T.E., Fox A.A., Collard C.D., Shernan S.K., Body S.C., CABG Genomics Investigators, 2010a. Chromosome 9p21 variant predicts mortality after coronary artery bypass graft surgery. *Circulation*, 122(11 Suppl).

Muehlschlegel J.D., Perry T.E., Liu K.Y., Fox A.A., Collard C.D., Shernan S.K., Body S.C., 2010b. Heart-type fatty acid binding protein is an independent predictor of death and ventricular dysfunction after coronary artery bypass graft surgery. *Anesthesia & Analgesia*, 111(5):1101–1109.

Naderi A., Teschendorff A.E., Barbosa-Morais N.L., Pinder S.E., Green A.R., Powe D.G., Robertson J.F.R., Aparicio S., Ellis I.O., Brenton J.D., Caldas C., 2006. A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, 26(10):1507–1516.

Nambi V., Chambless L., Folsom A.R., He M., Hu Y., Mosley T., Volcik K., Boerwinkle E., Ballantyne C.M., 2010. Carotid intima-media thickness and presence or absence of plaque improves prediction of coronary heart disease risk: the ARIC (Atherosclerosis Risk In Communities) study. *Journal of the American College of Cardiology*, 55(15):1600–1607.

Nathan H., de Jong M.C., Pulitano C., Ribero D., Strub J., Mentha G., Gigot J.F.F., Schulick R.D., Choti M.A., Aldrighetti L., Capussotti L., Pawlik T.M., 2010. Conditional survival after surgical resection of colorectal liver metastasis: an international multi-institutional analysis of 949 patients. *Journal of the American College of Surgeons*, 210(5):755–764.

Nelson M.R., Ryan P., Tonkin A.M., Ramsay E., Willson K., Wing L.W.H., Reid C.M., on behalf of the Second Australian National Blood Pressure Study Management Committee, 2010. Prediction of cardiovascular events in subjects in the Second Australian National Blood Pressure Study. *Hypertension*, 56(1):44–48.

Nielsen T.O., Parker J.S., Leung S., Voduc D., Ebbert M., Vickery T., Davies S.R., Snider J., Stijleman I.J., Reed J., Cheang M.C.U., Mardis E.R., Perou C.M., Bernard P.S., Ellis M.J., 2010. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptorâĂŞpositive breast cancer. *Clinical Cancer Research*, 16(21):5222–5232.

Nowak A.K., Francis R.J., Phillips M.J., Millward M.J., van der Schaaf A.A., Boucek J., Musk A.W., McCoy M.J., Segal A., Robins P., Byrne M.J., 2010. A novel prognostic model for malignant mesothelioma incorporating quantitative FDG-PET imaging with clinical parameters. *Clinical Cancer Research*, 16(8):2409–2417.

Núñez J., Núñez E., Consuegra L., Sanchis J., Bodí V., Martínez-Brotons A., Bertomeu-González V., Robles R., Bosch M.J., Fácila L., Darmofal H., Llàcer A., 2007. Carbohy-

drate antigen 125: an emerging prognostic risk factor in acute heart failure? *Heart*, 93(6):716–721.

Pan C.C., Chang Y.H., Chen K.K., Yu H.J., Sun C.H., Ho D.M.T., 2010. Constructing prognostic model incorporating the 2004 WHO/ISUP classification for patients with non-muscle-invasive urothelial tumours of the urinary bladder. *Journal of Clinical Pathology*, 63(10):910–915.

Paolo G., Maruyama S., Ozaki T., Taguchi A., Meigs J., Dimmeler S., Zeiher A.M., de Kreutzenberg S., Avogaro A., Nickenig G., Schmidt-Lucke C., Werner N., 2010. Circulating progenitor cell count for cardiovascular risk stratification: a pooled analysis. *PLoS one*, 5(7):e11488+.

Paynter N.P., Chasman D.I., Paré G., Buring J.E., Cook N.R., Miletich J.P., Ridker P.M., 2010. Association between a literature-based genetic risk score and cardiovascular events in women. *Journal of the American Medical Association*, 303(7):631–637.

Polonsky T.S., McClelland R.L., Jorgensen N.W., Bild D.E., Burke G.L., Guerci A.D., Greenland P., 2010. Coronary artery calcium score and risk classification for coronary heart disease prediction. *Journal of the American Medical Association*, 303(16):1610–1616.

Potters L., Roach M., Davis B.J., Stock R.G., Ciezki J.P., Zelefsky M.J., Stone N.N., Fearn P.A., Yu C., Shinohara K., Kattan M.W., 2010. Postoperative nomogram predicting the 9-year probability of prostate cancer recurrence after permanent prostate brachytherapy using radiation dose as a prognostic variable. *International Journal of Radiation Oncology Biology Physics*, 76(4):1061–1065.

Raman J.D., Ng C.K., Scherr D.S., Margulis V., Lotan Y., Bensalah K., Patard J.J.J., Kikuchi E., Montorsi F., Zigeuner R., Weizer A., Bolenz C., Koppie T.M., Isbarn H., Jeldres C., Kabbani W., Remzi M., Waldert M., Wood C.G., Roscigno M., Oya M., Langner C., Wolf S.S., Ströbel P., Fernández M., Karakiewcz P., Shariat S.F., 2010. Impact of tumor location on prognosis for patients with upper tract urothelial carcinoma managed by radical nephroureterectomy. *European Urology*, 57(6):1072–1079.

Rodondi N., Marques-Vidal P., Butler J., Sutton-Tyrrell K., Cornuz J., Satterfield S., Harris T., Bauer D.C., Ferrucci L., Vittinghoff E., Newman A.B., Health, Aging, and Body

Composition Study, 2010. Markers of atherosclerosis and inflammation for prediction of coronary heart disease in older adults. *American Journal of Epidemiology*, 171(5):540–549.

Rossi D., Fangazio M., De Paoli L., Puma A., Riccomagno P., Pinto V., Zigrossi P., Ramponi A., Monga G., Gaidano G., 2010. Beta-2-microglobulin is an independent predictor of progression in asymptomatic multiple myeloma. *Cancer*, 116(9):2188–2200.

Rossi F., Petrucci M.T., Guffanti A., Marcheselli L., Rossi D., Callea V., Vincenzo F., De Muro M., Baraldi A., Villani O., Musto P., Bacigalupo A., Gaidano G., Avvisati G., Goldaniga M., DePaoli L., Baldini L., 2009. Proposal and validation of prognostic scoring systems for IgG and IgA monoclonal gammopathies of undetermined significance. *Clinical Cancer Research*, 15(13):4439–4445.

Royston P., Reitz M., Atzpodien J., 2006. An approach to estimating prognosis using fractional polynomials in metastatic renal carcinoma. *British Journal of Cancer*, 94(12):1785–1788.

Rubinshtein R., Kuvin J.T., Soffler M., Lennon R.J., Lavi S., Nelson R.E., Pumper G.M., Lerman L.O., Lerman A., 2010. Assessment of endothelial function by non-invasive peripheral arterial tonometry predicts late cardiovascular adverse events. *European Heart Journal*, 31(9):1142–1148.

Rutten J.H., Mattace-Raso F.U., Steyerberg E.W., Lindemans J., Hofman A., Wieberdink R.G., Breteler M.M., Witteman J.C., van den Meiracker A.H., 2010. Amino-terminal pro-B-type natriuretic peptide improves cardiovascular and cerebrovascular risk prediction in the population: the Rotterdam study. *Hypertension*, 55(3):785–791.

Sánchez-Navarro I., Gámez-Pozo A., Pinto A., Hardisson D., Madero R., López R., San José B., Zamora P., Redondo A., Feliu J., Cejas P., Barón M.G., Vara J.A.F., Espinosa E., 2010. An 8-gene qRT-PCR-based gene expression score that has prognostic value in early breast cancer. *BMC Cancer*, 10.

Scarpa A., Mantovani W., Capelli P., Beghelli S., Boninsegna L., Bettini R., Panzuto F., Pederzoli P., Fave G.D., Falconi M., 2010. Pancreatic endocrine tumors: improved TNM staging and histopathological grading permit a clinically efficient prognostic stratification of patients. *Modern Pathology*, 23(6):824–833.

Sehestedt T., Jeppesen J., Hansen T.W., Wachtell K., Ibsen H., Torp-Pedersen C., Torp-Petersen C., Hildebrandt P., Olsen M.H., 2010. Risk prediction is improved by adding markers of subclinical organ damage to SCORE. *European Heart Journal*, 31(7):883–891.

Selvin E., Steffes M.W., Zhu H., Matsushita K., Wagenknecht L., Pankow J., Coresh J., Brancati F.L., 2010. Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults. *The New England Journal of Medicine*, 362(9):800–811.

Shariat S.F., Bolenz C., Karakiewicz P.I., Fradet Y., Ashfaq R., Bastian P.J., Nielsen M.E., Capitanio U., Jeldres C., Rigaud J., Müller S.C., Lerner S.P., Montorsi F., Sagalowsky A.I., Cote R.J., Lotan Y., 2010a. p53 expression in patients with advanced urothelial cancer of the urinary bladder. *BJU International*, 105(4):489–495.

Shariat S.F., Chade D.C., Karakiewicz P.I., Ashfaq R., Isbarn H., Fradet Y., Bastian P.J., Nielsen M.E., Capitanio U., Jeldres C., Montorsi F., Lerner S.P., Sagalowsky A.I., Cote R.J., Lotan Y., 2010b. Combination of multiple molecular markers can improve prognostication in patients with locally advanced and lymph node positive bladder cancer. *The Journal of Urology*, 183(1):68–75.

Shariat S.F., Godoy G., Lotan Y., Droller M., Karakiewicz P.I., Raman J.D., Isbarn H., Weizer A., Remzi M., Roscigno M., Kikuchi E., Bolenz C., Bensalah K., Koppie T.M., Kassouf W., Wheat J.C., Zigeuner R., Langner C., Wood C.G., Margulis V., 2010c. Advanced patient age is associated with inferior cancer-specific survival after radical nephroureterectomy. *BJU International*, 105(12):1672–1677.

Shariat S.F., Svatek R.S., Tilki D., Skinner E., Karakiewicz P.I., Capitanio U., Bastian P.J., Volkmer B.G., Kassouf W., Novara G., Fritsche H.M.M., Izawa J.I., Ficarra V., Lerner S.P., Sagalowsky A.I., Schoenberg M.P., Kamat A.M., Dinney C.P., Lotan Y., Marberger M.J., Fradet Y., 2010d. International validation of the prognostic value of lymphovascular invasion in patients treated with radical cystectomy. *BJU International*, 105(10):1402–1412.

Shariat S.F., Youssef R.F., Gupta A., Chade D.C., Karakiewicz P.I., Isbarn H., Jeldres C., Sagalowsky A.I., Ashfaq R., Lotan Y., 2010e. Association of angiogenesis related markers with bladder cancer outcomes and other molecular markers. *The Journal of Urology*, 183(5):1744–1750.

Singh M., Holmes D.R., Lennon R.J., Rihal C.S., 2010. Development and validation of risk adjustment models for long-term mortality and myocardial infarction following percutaneous coronary interventions. *Circulation: Cardiovascular Interventions*, 3(5):423–430.

Strand T.E., Rostad H., Wentzel-Larsen T., Von Plessen C., 2010. A population-based evaluation of the seventh edition of the TNM system for lung cancer. *European Respiratory Journal*, 36(2):401–407.

Svatek R.S., Shah J.B., Xing J., Chang D., Lin J., McConkey D.J., Wu X., Dinney C.P., 2010. A multiplexed, particle-based flow cytometric assay identified plasma matrix metalloproteinase-7 to be associated with cancer-related death among patients with bladder cancer. *Cancer*, 116(19):4513–4519.

Swallow E.B., Reyes D., Hopkinson N.S., Man W.D.C.D., Porcher R., Cetti E.J., Moore A.J., Moxham J., Polkey M.I., 2007. Quadriceps strength predicts mortality in patients with moderate to severe chronic obstructive pulmonary disease. *Thorax*, 62(2):115–120.

Tan M.H., Kanesvaran R., Li H., Tan H.L., Tan P.H., Wong C.F., Chia K.S., Teh B.T., Yuen J., Chong T.W., 2010. Comparison of the UCLA Integrated Staging System and the Leibovich Score in survival prediction for patients with nonmetastatic clear cell renal cell carcinoma. *Urology*, 75(6):1365–1370.e3.

Tournoux-Facon C., Paoletti X., Barbare J.C., Bouché O., Rougier P., Dahan L., Lombard-Bohas C., Faroux R., Raoul J.L., Bedenne L., 2011. Development and validation of a new prognostic score of death for patients with hepatocellular carcinoma in palliative setting. *Journal of Hepatology*, 54(1):108–114.

Trédan O., Ray-Coquard I., Chvetzoff G., Rebattu P., Bajard A., Chabaud S., Pérol D., Saba C., Quiblier F., Blay J.Y.Y., Bachelot T., 2011. Validation of prognostic scores for survival in cancer patients beyond first-line therapy. *BMC Cancer*, 11:95+.

Umbreit E.C., Crispen P.L., Shimko M.S., Farmer S.A., Blute M.L., Frank I., 2010. Multifactorial, site-specific recurrence model after radical cystectomy for urothelial carcinoma. *Cancer*, 116(14):3399–3407.

van der Pijl L.L.R., Birim O., van Gameren M., Kappetein A.P., Maat A.P.W.M., Steyerberg E.W., Bogers A.J.J.C., 2010. Validation of a prognostic model to predict survival after non-small-cell lung cancer surgery. *European Journal of Cardio-Thoracic Surgery*, 38(5):615–619.

van der Schroeff M.P., Terhaard C.H.J., Wieringa M.H., Datema F.R., Baatenburg de Jong R.J., 2010. Cytology and histology have limited added value in prognostic models for salivary gland carcinomas. *Oral Oncology*, 46(9):662–666.

van Nes J.G.H., Putter H., van Hezewijk M., Hille E.T.M., Bartelink H., Collette L., van de Velde C.J.H., 2010. Tailored follow-up for early breast cancer patients: A prognostic index that predicts locoregional recurrence. *European Journal of Surgical Oncology*, 36(7):617–624.

Wang X., Wan F., Pan J., Yu G.Z., Chen Y., Wang J.J., 2009. Prognostic value of the ratio of metastatic lymph nodes in gastric cancer: An analysis based on a Chinese population. *Journal of Surgical Oncology*, 99(6):329–334.

Weiss S.A., Blumenthal R.S., Sharrett A.R., Redberg R.F., Mora S., 2010. Exercise blood pressure and future cardiovascular death in asymptomatic individuals. *Circulation*, 121(19):2109–2116.

Williams S., Duchesne G., Gogna N., Millar J., Pickles T., Pratt G., Turner S., 2006. An international multicenter study evaluating the impact of an alternative biochemical failure definition on the judgment of prostate cancer risk. *International Journal of Radiation Oncology Biology Physics*, 65(2):351–357.

Wood A.M., Greenland P., 2009. Evaluating the prognostic value of new cardiovascular biomarkers. *Disease Markers*, 26(5):199–207.

Woodall C.E., Brock G.N., Fan J., Byam J.A., Scoggins C.R., McMasters K.M., Martin R.C.G., 2009. An evaluation of 2537 gastrointestinal stromal tumors for a proposed clinical staging system. *Arch Surg*, 144(7):670–678.

Yamada T., Tsuchiya K., Kato S., Kamei S., Taniguchi M., Takeuchi T., Yamamoto N., Ehara H., Deguchi T., 2010. A pretreatment nomogram predicting recurrence- and

progression-free survival for nonmuscle invasive bladder cancer in Japanese patients. *International Journal of Clinical Oncology*, 15(3):271–279.

Yang H.I.I., Sherman M., Su J., Chen P.J.J., Liaw Y.F.F., Iloeje U.H., Chen C.J.J., 2010. Nomograms for risk of hepatocellular carcinoma in patients with chronic hepatitis B virus infection. *Journal of Clinical Oncology*, 28(14):2437–2444.

Zigeuner R., Hutterer G., Chromecki T., Imamovic A., Kampel-Kettner K., Rehak P., Langner C., Pummer K., 2010a. External validation of the Mayo Clinic stage, size, grade, and necrosis (SSIGN) score for clear-cell renal cell carcinoma in a single European centre applying routine pathology. *European Urology*, 57(1):102–109.

Zigeuner R., Shariat S.F., Margulis V., Karakiewicz P.I., Roscigno M., Weizer A., Kikuchi E., Remzi M., Raman J.D., Bolenz C., Bensalah K., Capitanio U., Koppie T.M., Kassouf W., Sircar K., Patard J.J.J., Fernández M.I., Wood C.G., Montorsi F., Ströbel P., Wheat J.C., Haitel A., Oya M., Guo C.C., Ng C., Chade D.C., Sagalowsky A., Langner C., 2010b. Tumour necrosis is an indicator of aggressive biology in patients with urothelial carcinoma of the upper urinary tract. *European Urology*, 57(4):575–581.

Zivanovic O., Leitao M.M., Iasonos A., Jacks L.M., Zhou Q., Abu-Rustum N.R., Soslow R.A., Juretzka M.M., Chi D.S., Barakat R.R., Brennan M.F., Hensley M.L., 2009. Stage-specific outcomes of patients with uterine leiomyosarcoma: A comparison of the International Federation of Gynecology and Obstetrics and American Joint Committee on Cancer staging systems. *Journal of Clinical Oncology*, 27(12):2066–2072.