
THE UNIVERSITY
*of* EDINBURGH

# Statistical modelling of biomarkers incorporating non-proportional effects for survival data: with illustration by application to two residual risk models for predicting risk in early breast cancer.

Jacqueline Stephen

Presented for the Degree of Doctor of Philosophy

University of Edinburgh

2015

# Declaration

I, Jacqueline Stephen, declare that the following thesis has been composed by me and has not been submitted for any other degree or professional qualification.

# Abstract

Personalised medicine is replacing the one-drug-fits-all approach with many prognostic models incorporating biomarkers available for risk stratifying patients. Evidence has been emerging that the effects of biomarkers change over time and therefore violate the assumption of proportional hazards when performing Cox regression. Analysis using the Cox model when the assumptions are invalid can result in misleading conclusions.

This thesis reviews existing approaches for the analysis of non-proportional effects with respect to survival data. A number of well-developed approaches were identified but to date their uptake in practice has been limited. There is a need for more widespread use of flexible modelling to move away from standard analysis using a Cox model when the assumption of proportional hazards is violated.

Two novel approaches were applied to investigate the impact of follow-up duration on two residual risk models, IHC4 and Mammostrat, for predicting risk in early breast cancers using two studies with different lengths of follow up; the Edinburgh Breast Conservation Series (BCS) and the Tamoxifen versus Exemestane Adjuvant Multinational (TEAM) trial.

Similar results were observed between the two approaches that were considered, the multivariable fractional polynomial time (MFPT) approach and Royston-Parmer flexible parametric models, with their respective advantages and disadvantages being discussed.

The analyses identified a strong time-varying effect of IHC4 score with the prognostic effect of IHC4 score on time-to distant recurrence decreasing with increasing follow-up time. Mammostrat score identified a group of patients with an increased risk of distant recurrence over full follow-up in the TEAM and Edinburgh BCS cohorts. The results suggest a combined IHC4 and Mammostrat risk score could provide information on the risk of recurrence and warrants further study.

# Lay Summary

Prognostic models are tools that have been developed which attempt to stratify patients into low or high risk groups to aid clinical decision making. These tools often incorporate biomarkers, measurements that are indicators of a particular disease state. The Cox model is the most commonly used approach for analysing survival time data in medical research. The underlying assumption of this type of modelling is that of proportional hazards, assuming the effect of the biomarkers are constant over time. Analysis using the Cox model when the assumptions are invalid can result in misleading conclusions.

This thesis reviews alternatives to the Cox model which allow effects to vary over time with respect to survival data. A number of well-developed approaches were identified but to date their uptake in practice has been limited. There is a need for more widespread use of flexible modelling to move away from standard analysis using a Cox model when the model assumptions are violated.

Two novel approaches were applied to investigate the impact of follow-up duration on two residual risk models, IHC4 and Mammostrat, for predicting risk in early breast cancers using two studies with different lengths of follow up.

Similar results were observed between the two approaches that were considered, the multivariable fractional polynomial time (MFPT) approach and Royston-Parmer flexible parametric models, with their respective advantages and disadvantages being discussed.

The analyses identified a strong time-varying effect of IHC4 score with the prognostic effect of IHC4 score on time-to distant recurrence decreasing with increasing follow-up time. Mammostrat score identified a group of patients with an increased risk of distant recurrence over full follow-up in the TEAM and Edinburgh BCS cohorts. The results suggest a combined

IHC4 and Mammostrat risk score could provide information on the risk of recurrence and warrants further study.

# Publications and awards relating to the work of this thesis

## Paper in peer-reviewed journal

J. Stephen, G. Murray, D. A. Cameron, J. Thomas, I. H. Kunkler, W. Jack, G. R. Kerr, T. Piper, C. L. Brookes, D. Rea, C. H. Van de Velde, A. Hasenberg, C. Markopoulos, L. Dirix, C. Seynaeve and J. Bartlett. Time dependence of biomarkers: Non-proportional effects of immunohistochemical panels predicting relapse risk in early breast cancer. *British Journal of Cancer*. 2014; 111: 2242-7.

## Oral presentation at conference

J. Stephen, G. Murray, D. A. Cameron and J. Bartlett. Statistical modelling of biomarkers incorporating non-proportional effects for survival data. *RSS 2014 International Conference*, 1-4 September 2014, Sheffield, UK.

## Poster presentations at conference

J. Stephen, G. Murray, D.A. Cameron, J. Thomas, I.H. Kunkler, W. Jack, G.R. Kerr, T. Piper, C.L. Brookes, D. W. Rea, C. J. H. van de Velde, A. Hasenburg, C. Markopoulos, L. Dirix ,C. Seynaeve and J. Bartlett. Comparison of immunohistochemical residual risk panels to predict risk in early breast cancers treated with endocrine therapy. *San Antonio Breast Cancer Symposium*, December 2014, San Antonio, Texas.

J. Stephen, G. Murray, D.A. Cameron, J. Thomas, I.H. Kunkler, W. Jack, G.R. Kerr, T. Piper, C.L. Brookes, D. W. Rea, C. J. H. van de Velde, A. Hasenburg, C. Markopoulos, L. Dirix ,C. Seynaeve and J. Bartlett. Time dependence of biomarkers: Non-proportional effects of immunohistochemical panels predicting relapse risk in early breast cancer. *San Antonio Breast Cancer Symposium*, December 2014, San Antonio, Texas.

J. Stephen, G. Murray, D. A. Cameron and J. Bartlett. Statistical modelling of biomarkers incorporating non-proportional effects for survival data. *International Society for Clinical Biostatistics 35$^{th}$ Annual Conference*, 24-28 August 2014, Vienna, Austria.

J. Stephen, G. Murray, D. A. Cameron and J. Bartlett. Impact of length of follow-up on the evaluation of prognostic scores with an example using two breast cancer studies. *International Society for Clinical Biostatistics 35$^{th}$ Annual Conference*, 24-28 August 2014, Vienna, Austria.

J. Stephen, G. Murray, D. A. Cameron and J. Bartlett. Statistical modelling of biomarkers incorporating non-proportional effects for survival data. *Clinical Trial Methodology Conference*, 18-19 Novemeber 2013, Edinburgh, UK.

# Acknowledgements

I would like to thank my supervisors, Professor Gordon Murray, Professor John Bartlett and Professor David Cameron for their excellent supervision, advice and guidance. I am very grateful for the many opportunities they have provided.

A special thanks to Morag Leitch for a warm welcome to the Centre for Population Health Sciences, who has been an invaluable support throughout the years and was a large contributor in making a three month work placement in Toronto possible.

Thank you to all my friends and family for their encouragement and support, with a special mention to my grandparents. Huge love and appreciation to Sean Ford for his constant support, patience and for keeping me sane over the last few months.

# Contents

**Chapter 1: Introduction and Outline of Thesis ..... 1**

**Chapter 2: Clinical Background ..... 3**

2.1 Breast Cancer ..... 3
2.2 Biomarkers ..... 4
2.2.1 Definition ..... 4
2.2.2 Established Biomarkers in Breast Cancer ..... 6
2.3 Prognostic Models ..... 7
2.4 Time-Dependency of Effects ..... 11
2.5 Discussion ..... 14

**Chapter 3: Survival Analysis ..... 15**

3.1 Introduction ..... 15
3.2 General Principles ..... 15
3.3 The Survival Function and Hazard Rate ..... 16
3.4 Non-Parametric Estimates ..... 17
3.4.1 Kaplan-Meier ..... 17
3.4.2 Nelson-Aalen Estimator ..... 18
3.5 Proportional Hazards Models ..... 19
3.5.1 The Exponential Model ..... 19
3.5.2 The Weibull Model ..... 19
3.5.3 The Cox Model ..... 20

**Chapter 4: Literature Review ..... 27**

4.1 Introduction ..... 27
4.2 Methods ..... 27
4.3 Results ..... 30
4.4 Multiplicative Hazards Models ..... 30
4.4.1 The Stratified Cox Model ..... 33
4.4.2 The Weighted Cox Model ..... 33
4.4.3 Smoothed Schoenfeld Residuals ..... 34
4.4.4 The Extended Cox Model ..... 35
4.5 Additive Hazards Models ..... 49
4.6 Additive-Multiplicative Hazards Models ..... 53

4.7 Parametric Models ..... 54
4.8 Other Models ..... 57
4.8.1 Cure and Frailty Models ..... 58
4.8.2 Tree-Based Method ..... 59
4.8.3 Bayesian Approaches ..... 61
4.9 Application of novel methods to data analysis ..... 62
4.9.1 Fractional Polynomials ..... 63
4.9.2 Regression Splines ..... 64
4.9.3 Additive and Additive-Multiplicative Models ..... 64
4.9.4 Parametric Models ..... 65
4.9.5 Source of Citations ..... 66
4.10 Reviews of the approaches for modelling non-proportional effects ..... 66
4.10.1 Data Sets and Approaches Compared ..... 66
4.10.2 Summary of Findings ..... 67
4.11 Discussion ..... 71
4.11.1 Limitations ..... 73

**Chapter 5: IHC4 and Mammostrat ..... 75**

5.1 Introduction ..... 75
5.2 Model Development and First Validation ..... 75
5.2.1 IHC4 ..... 75
5.2.2 Mammostrat Score ..... 77
5.3 Independent Evaluations of IHC4 Score ..... 78
5.3.1 Contribution of the IHC4+C score to decision making in clinical practice… ..... 79
5.3.2 Comparison of PAM50 Risk of Recurrence Score with Oncotype DX and IHC4… ..... 79
5.3.3 Comparison of the breast-cancer index (BCI) assay, Oncotype Dx, and IHC4… ..... 80
5.3.4 Clinical Impact of using the IHC4 Score ..... 81
5.4 Independent Evaluations of Mammostrat Score ..... 82
5.4.1 Chemo-sensitivity and Stratification of Mammostrat in the NSABP B14 and B20 Trials ..... 82
5.4.2 Mammostrat as a tool to stratify breast cancer patients at risk of recurrence during endocrine therapy ..... 83
5.4.3 Mammostrat as a predictor of Early Relapse Risk in the TEAM Study ..... 84
5.5 Time-dependent analysis of other prognostic scores ..... 85
5.5.1 Validation and Clinical Utility of MammaPrint ..... 85
5.5.2 Strong Time Dependence of the 76-Gene Prognostic Signature ..... 85

5.5.3 Comparison of prognostic gene expression signatures for breast cancer86
5.5.4 EndoPredict prognostic of late distant metastases ..... 87
5.5.5 Breast Cancer Index Identifies Breast Cancer Patients at Risk for Early- and Late-Distant Recurrence..... 88
5.5.6 Long-term impact of MammaPrint on breast cancer outcome..... 89
5.6 Discussion ..... 90

**Chapter 6: The Data ..... 93**

6.1 Materials ..... 93
6.1.1 The Edinburgh Breast Conservation Series ..... 93
6.1.2 The Tamoxifen Exemestane Adjuvant Multinational Trial ..... 93
6.2 Variables ..... 94
6.2.1 Clinical Variables ..... 94
6.2.2 Biomarkers ..... 94
6.2.3 Survival Outcome ..... 97
6.3 Patient Demographics ..... 98
6.4 Missing Data ..... 101
6.4.1 Edinburgh BCS Cohort ..... 101
6.4.2 TEAM Cohort ..... 111
6.5 Software ..... 111

**Chapter 7: Analysis of IHC4 and Mammostrat ..... 113**

7.1 Introduction ..... 113
7.2 Methods ..... 113
7.2.1 Materials ..... 113
7.2.2 Statistical Analysis ..... 119
7.3 Results ..... 120
7.3.1 Kaplan-Meier Estimates of TTDR ..... 121
7.3.2 Classification of patients according to IHC4 and Mammostrat score .. 124
7.3.3 Assumption of Linearity ..... 125
7.3.4 Comparison of IHC4 and Mammostrat ..... 127
7.3.5 Assumption of Proportional Hazards ..... 135
7.4 Discussion ..... 138

**Chapter 8: Analysis of IHC4 and Mammostrat Incorporating Non-Proportional Effects ..... 141**

8.1 Introduction ..... 141
8.2 Piece-wise constant effects ..... 141

8.2.1 Results ..... 142
8.2.2 Summary ..... 164
8.3 MFPT Approach ..... 165
8.3.1 Splitting the data ..... 165
8.3.2 Results ..... 166
8.3.3 Summary ..... 182
8.4 RP Flexible Parametric Models ..... 186
8.4.1 Results ..... 186
8.4.2 Summary ..... 200
8.5 Comparison of Approaches ..... 202
8.5.1 Extended Cox Model versus Proportional Hazards ..... 202
8.5.2 RP Models versus Extended Cox Models ..... 204
8.6 Conclusions from various subgroup analyses ..... 206
8.7 Discussion ..... 209

**Chapter 9: Analysis of Individual Biomarkers ..... 213**

9.1 Introduction ..... 213
9.2 Materials ..... 213
9.3 Biomarkers ..... 213
9.4 Correlation ..... 214
9.5 Assumptions of Linearity and Proportional Hazards ..... 214
9.5.1 Linearity on the Log-hazard ..... 214
9.5.2 Proportional Hazards ..... 214
9.6 Multivariable Model ..... 221
9.6.1 10 year follow-up ..... 222
9.6.2 Full follow-up ..... 237
9.7 RP Flexible Parametric Models ..... 243
9.7.1 Choice of scale ..... 244
9.7.2 Selection of variables and functional forms ..... 246
9.7.3 Comparison of Estimated Effects ..... 251
9.7.4 Restricted Mean Survival Time ..... 263
9.8 Discussion ..... 268

**Chapter 10: Discussion and Recommendations ..... 273**

10.1 Summary of Results ..... 273
10.1.1 Statistical Methods for the Analysis of Non-Proportional Hazards ..... 273
10.1.2 Evaluation of IHC4 and Mammostrat Residual Risk Panels ..... 278

10.2 Limitations ..... 282
10.3 Future Directions ..... 283
10.4 Conclusion ..... 284

**References.. ..... 287**

# List of Tables

Table 4.1 Multiplicative Hazard Models ..... 31

Table 4.2 Advantages/Disadvantages of Multiplicative Hazards Models ..... 32

Table 4.3 Approaches for Estimation of $\boldsymbol{\beta}(\boldsymbol{t})$ in the Extended Cox Model ..... 36

Table 4.4 Advantages/Disadvantages of the Approaches for Estimation of $\boldsymbol{\beta}(\boldsymbol{t})$ in the Extended Cox Model ..... 38

Table 4.5 Additive Hazard Models and Multiplicative-Additive Hazard Models ..... 51

Table 4.6 Advantages\Disadvantages of Additive Hazards Models and Additive-Multiplicative Hazards Models ..... 52

Table 4.7 Parametric Models and Others ..... 55

Table 4.8 Advantages/Disadvantages associated with Parametric and Other Models ..... 56

Table 4.9 Summary of Review Papers of Approaches for Dealing with Non-Proportional Hazards ..... 68

Table 4.10 Key Points of Review Papers of Approaches for Dealing with Non-Proportional Hazards ..... 69

Table 6.1 Comparison of scores as categorical or continuous predictors as assessed by AIC from a univariate Cox regression model. ..... 95

Table 6.2 Baseline Characteristics of Patients ..... 100

Table 6.3 Information on Completeness of Data ..... 103

Table 6.4 Statistics that can and connot be combined using Rubin's rules ..... 103

Table 6.5 Summary of variables in the imputation model ..... 105

Table 6.6 Estimated coefficients from a univariate Cox regression of IHC4 and Mammostrat score from a complete case analysis and different imputation procedures. ..... 111

Table 7.1 Baseline characteristics of patients in the analysis set ..... 118

Table 7.2 Information on number of patients, number of events and median survival time for patient subgroups ..... 121

Table 7.3 Five and Fifteen-year freedom from distant recurrence estimates ..... 124
Table 7.4 Frequencies of patients according to their classification of IHC4 and Mammostrat scores ..... 125
Table 7.5 Multivariate Cox regression of IHC4 score and Mammostrat score ..... 129
Table 7.6 Wald tests and performance data for assessing the amount of information added by the IHC4 score or the Mammostrat score or both to clinical factors ..... 130
Table 7.7 Multivariate Cox regression of IHC4 score and Mammostrat score for Exploratory Subgroups ..... 131
Table 7.8 Wald tests and performance data for assessing the amount of information added by IHC4 score or Mammostrat score or both to clinical factors for exploratory subgroups. .... 132
Table 7.9 Grambsch-Therneau test of non-proportional hazards ..... 137
Table 8.1 Performance data on IHC4 and Mammostrat score in addition to clinical factors ..... 143
Table 8.2 Period-Specific Multivariate Cox Regression of IHC4 and Mammostrat score .. 149
Table 8.3 Period-Specific Multivariate Cox Regression of IHC4 and Mammostrat score for Exploratory Subgroup Analysis ..... 150
Table 8.4 Wald tests and performance data for assessing the amount of information added by the IHC4 score or the Mammostrat score or both to clinical factors in the first 5 years of follow-up. ..... 151
Table 8.5 Period-Specific Multivariate Cox Regression of IHC4 and Mammostrat score Combined ..... 152
Table 8.6 Period-Specific Multivariate Cox Regression of IHC4 and Mammostrat score Combined for Exploratory Subgroup Analysis ..... 153
Table 8.7 Period Specific Multivariate Cox Regression of the Combined Score ..... 157
Table 8.8 Period-Specific Multivariate Cox Regression of the Combined Score for Exploratory Subgroup Analysis ..... 158

Table 8.9 Significant predictors of TTDR in the 0-5 and 5-10 year time intervals by patient subgroup ..... 164
Table 8.10 Example of splitting data in half year intervals ..... 166
Table 8.11 Number of observations created when categorising the time interval ..... 166
Table 8.12 Multivariate Cox regression of Mammostrat score up to 10-years follow-up ... 169
Table 8.13 Multivariate Cox regression of Mammostrat score after adjustment for time-varying IHC4 score up to 10-years follow-up ..... 171
Table 8.14 Multivariate Cox regression of the Combined score for the exploratory subgroups in the TEAM cohort ..... 178
Table 8.15 Wald tests and performance data for assessing the amount of information added by the time-varying IHC4 score or the Mammostrat score or both to clinical factors in the first 10 years of follow-up. ..... 180
Table 8.16 Wald tests and performance data for assessing the amount of information added by the time-varying IHC4 score or the Mammostrat score or both to clinical factors for full-follow-up in the Edinburgh BCS cohort. ..... 182
Table 8.17 Time-varying effects identified by the FPT algorithm ..... 183
Table 8.18 Significant predictors of TTDR after identification of time-varying effects ..... 185
Table 8.19 Flexible parametric regression of Mammostrat score up to 10-years follow-up 190
Table 8.20 Flexible parametric regression of Mammostrat score over full follow-up in the Edinburgh BCS cohort ..... 192
Table 8.21 5 and 10-year restricted-mean survival estimates for the Edinburgh BCS and TEAM cohorts ..... 197
Table 8.22 15 and 20-year restricted-mean survival estimates for the Edinburgh BCS cohort ..... 198
Table 8.23 Flexible parametric regression of Combined score up to 10-years follow-up ... 199
Table 8.24 Comparison of hazard ratio estimates for Mammostrat score from an extended Cox or RP model ..... 205

Table 8.25 Conclusions on whether IHC4 and/or Mammostrat were prognostic of TTDR in various subgroups. 207

Table 9.1 Associations between biomarkers and clinical variables in the Edinburgh BCS cohort 215

Table 9.2 Associations between biomarkers and clinical variables in the TEAM cohort. 216

Table 9.3 Grambsch-Therneau test of non-proportional hazards 221

Table 9.4 MFPT Results for the Edinburgh BCS Cohort (10 year follow-up) – biomarker only model 222

Table 9.5 MFPT Results for the TEAM cohort – biomarker only model 224

Table 9.6 MFPT Results for the Edinburgh BCS cohort (10 year follow-up) – biomarker and clinical model 225

Table 9.7 MFPT Results for the TEAM cohort – biomarker and clinical model 226

Table 9.8 MFPT Results for the TEAM cohort restricting ER and Ki67 to be linear – biomarker and clinical model 227

Table 9.9 MFPT Results for the TEAM cohort restricting ER and Ki67 to be linear – biomarker only model 228

Table 9.10 MFPT Results for the Edinburgh BCS cohort (10 year follow-up) restricting PgR to be linear – biomarker only model 229

Table 9.11 MFPT Results for the Edinburgh BCS cohort (10 year follow-up) restricting PgR to be linear – biomarker and clinical model 230

Table 9.12 Multivariate Cox Regression of biomarkers identified by the MFPT approach in the Edinburgh BCS Cohort (10 year follow-up) 231

Table 9.13 Multivariate Cox Regression of biomarkers identified by the MFPT approach in the TEAM Cohort. 232

Table 9.14 Edinburgh BCS cohort (10-year follow-up). Comparison of discrimination of model identified in step 1 of the MFPT approach and the final model. 236

Table 9.15 TEAM cohort. Comparison of discrimination of model identified in step 1 of the MFPT approach and the final model..... 237
Table 9.16 MFPT Results for the Edinburgh BCS Cohort with Full follow-up – Biomarker only model ..... 238
Table 9.17 MFPT Results for the Edinburgh BCS Cohort with Full follow-up ..... 239
Table 9.18 Multivariate Cox Regression of biomarkers identified by the MFPT approach in the Edinburgh BCS Cohort (full follow-up) ..... 240
Table 9.19 Edinburgh BCS cohort (full follow-up). Comparison of discrimination of model identified in step 1 of the MFPT approach and the final model..... 242
Table 9.20 Edinburgh BCS cohort (10-year follow-up). Choice of scale and baseline complexity for multivariable model containing the 9 biomarkers. ..... 245
Table 9.21 Edinburgh BCS cohort (full follow-up). Choice of scale and baseline complexity for multivariable model containing the 9 biomarkers. ..... 245
Table 9.22 TEAM cohort. Choice of scale and baseline complexity for multivariable model containing the 9 biomarkers. ..... 246
Table 9.23 Final Biomarker Only Model from the MFPT approach fitting using RP models with different scales for the Edinburgh BCS cohort (10-year follow-up)..... 247
Table 9.24 Final Biomarker Only Model from the MFPT approach fitted using RP models with different scales for the TEAM cohort. ..... 249
Table 9.25 Final Biomarker Only Model from the MFPT approach fitting using RP models with different scales for the Edinburgh BCS cohort (full follow-up) ..... 250
Table 9.26 Comparison of estimated coefficients from a RP model with different choice of scales..... 251
Table 9.27 5 and 10-year restricted-mean survival estimates for the Edinburgh BCS cohort (10-year follow-up) from a RP model with hazards and Aranda-Ordaz scales ..... 265
Table 9.28 5 and 10-year restricted-mean survival estimates for the TEAM cohort from a RP model with hazard and Aranda-Ordaz scales..... 266

Table 9.29 5, 10 and 15-year restricted-mean survival estimates for the Edinburgh BCS cohort (full follow-up) from a RP model with hazards and Aranda-Ordaz scales ..... 267

# List of Figures

Figure 2.1 Breast Cancer European Age-Standardised Incidence and Mortality Rates..... 4

Figure 2.2 Annual Hazard Rates for Breast Cancer Death ..... 12

Figure 4.1 Literature Review Search Strategy ..... 28

Figure 4.2 Flowchart of Eligibility Assessment and Inclusion ..... 29

Figure 6.1 Distribution of Mammostrat Score ..... 96

Figure 6.2 Kaplan-Meier curves for time to distant recurrence. ..... 99

Figure 6.3 Distribution of IHC4 Score..... 101

Figure 6.4 Distribution of the continuous covariates. ..... 104

Figure 6.5 Plots of martingale residuals versus continuous covariates..... 107

Figure 6.6 Comparison of linear and non-linear effects. ..... 107

Figure 6.7 Comparison of Observed (0) and Imputed (1-42) data for continuous variables. ..... 109

Figure 6.8 Comparison of observed (0) versus imputed (1-42) data using predictive mean matching..... 109

Figure 6.9 Comparison of observed (0) versus imputed (1-42) data for categorical variables. ..... 110

Figure 7.1 Clinical subgroups of interest ..... 115

Figure 7.2 Kaplan-Meier curves for time to distant recurrence for clinical subgroups. ..... 117

Figure 7.3 Kaplan-Meier curves for time-to distant recurrence..... 122

Figure 7.4 Kaplan-Meier curves for time-to distant recurrence (exploratory subgroups) ... 123

Figure 7.5 Kaplan-Meier curves for TTDR for IHC4 score vs Mammostrat score. ..... 125

Figure 7.6 Results of the MFP algorithm..... 126

Figure 7.7 Plot of martingale residuals versus number of positive nodes..... 127

Figure 7.8 Calibration slope estimates of observed versus predicted 5 year risk of distant recurrence..... 134

Figure 7.9 Plots of Schoenfeld residuals for each explanatory variable ..... 136
Figure 8.1 All ER-positive patients. Observed versus predicted 5 year risk of distant recurrence (DR)..... 144
Figure 8.2 ER-positive patients treated with endocrine therapy only. Observed versus predicted 5 year risk of distant recurrence ..... 145
Figure 8.3 All ER-positive patients. Period-specific Kaplan-Meier curves ..... 147
Figure 8.4 ER-positive patients treated with endocrine therapy only. Period-specific Kaplan-Meier curves..... 148
Figure 8.5 Period-specific Kaplan-Meier curves for time to distant recurrence for the combined score ..... 156
Figure 8.6 All ER-positive patients. Two-yearly adjusted hazard ratio estimates ..... 160
Figure 8.7 ER-positive patients treated with endocrine therapy only. Two-yearly adjusted hazard ratio estimates..... 161
Figure 8.8 Combined Score. Two-yearly adjusted hazard ratio estimates ..... 163
Figure 8.9 Results of the FPT algorithm for ICH4 score ..... 167
Figure 8.10 All ER-positive patients, IHC4 Score. Time-dependent adjusted hazard ratio estimate (up to 10 years) ..... 167
Figure 8.11 Results of the FPT algorithm for Mammostrat high risk ..... 169
Figure 8.12 IHC4 Score – Simultaneous Analysis. Time-dependent adjusted hazard ratio estimate (up to 10 years) ..... 170
Figure 8.13 IHC4 Score. Time-dependent adjusted hazard ratio estimate (full follow-up) . 172
Figure 8.14 Mammostrat Score. Time-dependent adjusted hazard ratio estimate (full follow-up)..... 172
Figure 8.15 IHC4 Score – Simultaneous Analysis. Time-dependent adjusted hazard ratio (full follow-up)..... 173
Figure 8.16 IHC4 Score – Exploratory Subgroups. Time-dependent adjusted hazard ratio 174

Figure 8.17 IHC4 Score – Simultaneous Analysis (Exploratory Subgroups). Time-dependent adjusted hazard ratio estimate ..... 176
Figure 8.18 Combined Score. Time-dependent adjusted hazard ratio estimate (full follow-up) ..... 177
Figure 8.19 Combined Score (exploratory subgroups). Time-dependent adjusted hazard ratio estimate (full follow-up) ..... 178
Figure 8.20 Comparison of adjusted effect and 95% CIs (dashed lines) for a ten-unit increase in IHC4. ..... 184
Figure 8.21 Comparison of estimated effects between subgroups. ..... 185
Figure 8.22 Modelling the baseline hazard ..... 187
Figure 8.23 Comparison of baseline hazard rates ..... 188
Figure 8.24 Different degrees of freedom (df) for the time-varying ICH4 score ..... 189
Figure 8.25 All ER-positive patients. Time-dependent adjusted hazard ratio estimate (up to 10 years) ..... 189
Figure 8.26 ER-positive patients treated with endocrine therapy. Time-varying effect of IHC4 score ..... 189
Figure 8.27 Time-dependent adjusted hazard ratio estimate (full follow-up) ..... 191
Figure 8.28 Different degrees of freedom (df) for time-varying Mammostrat ..... 192
Figure 8.29 Predicted distant recurrence rates ..... 194
Figure 8.30 Predicted difference in survival probability ..... 194
Figure 8.31 Predicted difference in hazard rates and survival probability ..... 195
Figure 8.32 Combined Score. Time-dependent adjusted hazard ratio estimate (full follow-up) ..... 199
Figure 8.33 Comparison of time-constant (PH) and time-varying (TV) effect ..... 201
Figure 8.34 Comparison of extended Cox versus Cox PH model. ..... 202
Figure 8.35 Comparison of extended Cox model and piece-wise constant effects ..... 203
Figure 8.36 Comparison of extended Cox model and two-yearly intervals ..... 203

Figure 8.37 Comparison of extended Cox and RP models. ..... 205
Figure 9.1 Plots of martingale residuals versus continuous covariates. ..... 217
Figure 9.2 Comparison of non-linear (dashed line) and linear (solid line) effects ..... 218
Figure 9.3 Plots of Schoenfeld residuals for each explanatory variable versus survival time in the Edinburgh BCS cohort. ..... 219
Figure 9.4 Plots of Schoenfeld residuals for each explanatory variable versus survival time in the TEAM cohort. ..... 220
Figure 9.5 Illustration of different fractional polynomials to model interaction ..... 224
Figure 9.6 Comparison of non-linear (dashed line) and linear (solid line) effects of PgR ... 229
Figure 9.7 Edinburgh BCS cohort. Time-varying adjusted hazard ratios ..... 233
Figure 9.8 TEAM cohort. Time-varying adjusted hazard ratios ..... 234
Figure 9.9 Edinburgh BCS cohort – full follow-up. Time-varying adjusted hazard ratios .. 241
Figure 9.10 Time-varying (solid black line) and time-constant (dash-dot blue line) adjusted hazard ratio ..... 247
Figure 9.11 Time-varying adjusted hazard ratio for Ki67 for those p53-negative (solid line) and p53-positive (dashed line) ..... 249
Figure 9.12 Comparison of the effect of Ki67 from a univariate RP model with different choice of scales. ..... 253
Figure 9.13 Comparison of estimated effects from a multivariable RP model with different scales ..... 254
Figure 9.14 Edinburgh BCS Cohort. Estimated hazard ratios and 95% confidence intervals (dashed lines) from a multivariable RP model with Aranda-Ordaz (AO) scale ..... 256
Figure 9.15 TEAM Cohort. Comparison of estimated effects from a multivariable RP model with different scales (hazards (H), odds (O), probit and Aranda-Ordaz (AO)) ..... 257
Figure 9.16 Time-varying adjusted hazard ratio for p53 from a Cox model with a time-varying (TV) effect, a RP model with hazards (H) scale and Aranda-Ordaz (AO) scale ..... 259

Figure 9.17 Estimated hazard ratios and 95% confidence intervals (dashed lines) from a multivariable RP model with Aranda-Ordaz (AO) scale for all ER-positive patients in the TEAM cohort. ..... 259

Figure 9.18 Comparison of estimated effects from a multivariable RP model with different scales (hazards (H), odds (O), probit and Aranda-Ordaz (AO)) for all ER-positive patients in the Edinburgh BCS cohort with full follow-up. ..... 261

Figure 9.19 Comparison of estimated effects from a univariate RP model of (A) PgR and (B) NDRG1 with different scales (hazards (H), odds (O), probit and Aranda-Ordaz (AO)) for all ER-positive patients in the Edinburgh BCS cohort with full follow-up. ..... 262

Figure 9.20 Estimated effects and 95% confidence intervals (dashed lines) for multivariable RP model with Aranda-Ordaz (AO) scale for all ER-positive patients in the Edinburgh BCS cohort with full follow-up. ..... 263

# Abbreviations

| | |
|---|---|
| 95% CI | 95% Confidence Interval |
| AIC | Akaike's Information Criteria |
| AHR | Average Hazard Ratio |
| AI | Aromatase Inhibitors |
| ANN | Artificial Neural Network |
| AO | Aranda-Ordaz |
| AoL | Adjuvant! Online |
| ATAC | Arimidex, Tamoxifen, Alone or in Combination |
| ATLAS | Adjuvant Tamoxifen: Longer Against Shorter |
| aTTom | Adjuvant Tamoxifen—To Offer More |
| BCCA | British Columbia Cancer Agency |
| BCI | Breast-Cancer Index |
| BCS | Breast Conservation Series |
| BDSM | Bayesian Dynamic Survival Model |
| BIC | Bayes information criterion |
| CART | Classification And Regression Trees |
| CCIH | Comprehensive Cancer Institute of Huntsville |
| CCF | Cleveland Clinic Foundation |
| CDF | Cumulative Distribution Function |
| CRP | C-Reactive Protein |
| DP | Dirichlet Process |
| DR | Distant Recurrence |
| DRFS | Distant Recurrence Free Survival |
| DMFS | Distant Metastases Free Survival |
| EP | EndoPredict |
| HR | Hazard Ratio |
| ER | Estrogen Receptor |
| FISH | Fluorescence In Situ Hybridization |
| FP | Fractional Polynomial |

| | |
|---|---|
| FPT | Fractional Polynomial-Time |
| GGI | Gene expression Grade Index |
| GORH | Generalised Odds-Rate Hazards |
| IHC | Immunohistochemistry |
| LN | Lymph-node |
| MAR | Missing At Random |
| MCMC | Markov Chain Monte Carlo |
| MICE | Multiple Imputation using Chained Equations |
| MINDACT | Microarray In Node-negative and 1–3 node positive Disease may Avoid ChemoTherapy |
| MDT | Multidisciplinary Team |
| MFP | Multivariable Fractional Polynomial |
| MFPT | Multivariable Fractional Polynomial-Time |
| NA | Not Applicable |
| NICE | National Institute for Health and Care Excellence |
| NPI | Nottingham Prognostic Index |
| NSABP | National Surgical Adjuvant Breast and Bowel Project |
| OS | Overall Survival |
| PCR | Polymerase Chain Reaction |
| PDF | Probability Density Function |
| PgR | Progesterone Receptor |
| PH | Proportional Hazards |
| PL | Partial Likelihood |
| PMM | Predictive Mean Matching |
| RFI | Recurrence-Free Interval |
| RMST | Restricted Mean Survival Time |
| RoR | Risk of Recurrence |
| RP | Royston-Parmer |
| RS | Recurrence Score |
| TEAM | Tamoxifen Exemestane Adjuvant Multinational |
| TNM | Tumour size, Nodes, Metastases |

| | |
|---|---|
| TMA | Tissue Microarray |
| TTDR | Time-To Distant Recurrence |
| TV | Time-Varying |

# Chapter 1: Introduction and Outline of Thesis

Breast cancer is a worldwide public health problem and is an extremely heterogeneous disease, with varied genotypic and phenotypic features, behaviour and response to therapy. Whilst the mortality rate has decreased, the incidence of breast cancer is on the rise and investigation into long-term effects of adjuvant therapy is becoming more important as people are surviving longer. Personalised medicine is replacing the one-drug-fits-all approach with many prognostic models incorporating biomarkers available for risk stratifying patients.

A standard approach to model the time to a certain event such as relapse or death is the Cox proportional hazards model. Evidence has been emerging that the effects of biomarkers change over time and therefore violate the assumption of proportional hazards when performing Cox regression. Analysis using the Cox model when the assumptions are invalid can result in misleading conclusions.

Compared with other clinical subgroups, ER-positive patients have the best overall prognosis, however many recurrences occur in patients who remain disease-free for 5 years. Assessment of risk of recurrence over time is clinically important to identify patients with a low risk of recurrence for avoidance of chemotherapy and those with high risk of recurrence beyond 5 years for the decision of whether to extend adjuvant endocrine therapy.

The first aim of this thesis is to review existing methods for the analysis of survival data when there is evidence of non-proportional effects. The second aim is a comparison of two residual risk models, IHC4 and Mammostrat, to determine which provides more information on the risk of recurrence in the context of additional clinical factors or whether combining both approaches would increase the information available to patients and clinicians. The main approach of this thesis is the application of novel methods to real data.

The thesis is arranged in the following order. First, the clinical background is introduced in chapter 2. A brief overview of the statistical methods used when analysing time-to-event data is given in chapter 3 with an emphasis on the assumptions in the Cox proportional hazards model.

Chapter 4 is a detailed review of existing methods for the analysis of survival data when the proportional hazards assumption does not hold. Advantages and disadvantages are summarised with an interest in whether the methods have been applied in practice.

The two residual risk models of interest in this thesis are described in chapter 5, with detail on the development and previous evaluation of the models. The two data sets used for analysis throughout the thesis are described in chapter 6, with detail on the multiple imputation performed on one of the cohorts.

The first data analysis, in Chapter 7, involves the comparison of the two residual risk models, with a focus on fitting the standard Cox proportional hazards model and assessing the validity of the model assumptions.

Chapter 8 applies two of the methods identified in the literature review chapter for the analysis of IHC4 and Mammostrat incorporating non-proportional effects. The individual markers from the two residual risk models are then analysed in Chapter 9 to determine which have a prognostic effect on the outcome and which, if any, show evidence of a non-proportional effect.

In the final chapter (Chapter 10), the main findings of the thesis are discussed and suggestions are made for possible further areas of interest.

# Chapter 2: Clinical Background

## 2.1 Breast Cancer

Breast cancer is a worldwide public health problem and is an extremely heterogeneous disease, with varied genotypic and phenotypic features, behaviour and response to therapy (Rakha et al., 2012). It is the most frequently diagnosed cancer among women in the western world and is a leading cause of mortality, second to lung cancer (Cancer Research UK, 2012). In 2011 there were 50,285 new cases of breast cancer diagnosed in the UK (Cancer Research UK, 2015). The large numbers of cases and complexity of treatment options makes decisions about the correct treatment to give to patients difficult.

Whilst the mortality rate has decreased, the incidence of breast cancer is on the rise (Figure 2.1). Ten year relative survival rate for women has increased from 41% for those diagnosed with breast cancer in 1971-1975 in England and Wales to 73% for those diagnosed between 1996 and 2000 (Cancer Research UK, 2012). The incidence rate of breast cancer increased from 74.2 per 100,000 in 1975 to 124.1 per 100,000 in 2008, whereas mortality rate decreased from 38.6 per 100,000 in 1975 to 26.1 per 100,000 in 2008 (Figure 2.1). Investigation into long-term effects of adjuvant therapy is therefore becoming more and more important as people are surviving longer.

As a result of improvements in routine breast cancer screening, a shift toward the detection of early-stage node-negative breast cancer with better prognosis has occurred (Weigel and Dowsett, 2010). This creates challenges for clinicians in determining the best choice of adjuvant treatment.

Adjuvant therapy is defined by oncologists as ‘treatment for presumed microscopic disease’, to reduce the subsequent risk of relapse and death (Bartlett et al., 2013). Standard practice in

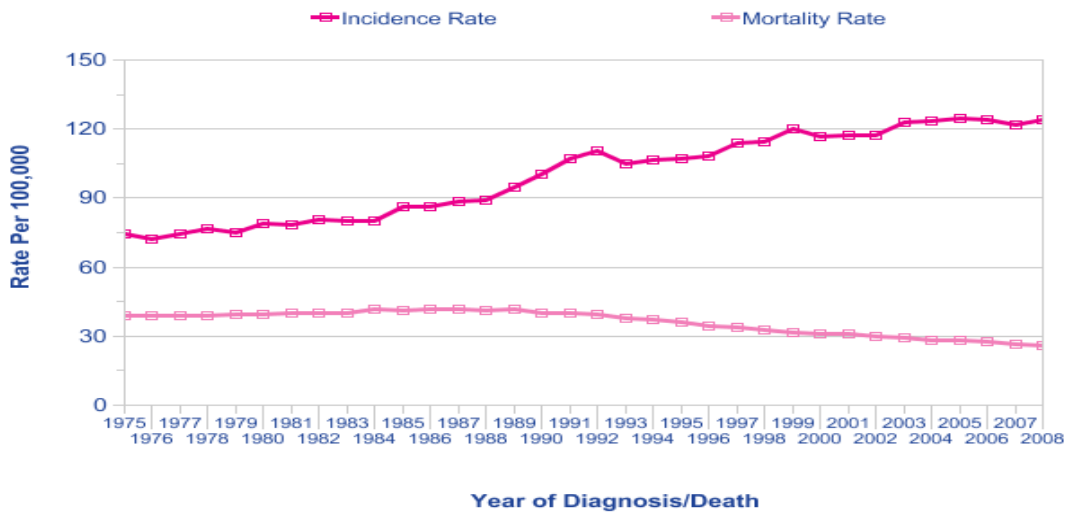



**Figure 2.1** Breast Cancer European Age-Standardised Incidence and Mortality Rates for Females in Great Britain between 1975 and 2008 (Cancer Research UK, 2012).

the UK is to offer chemotherapy to most postmenopausal women with axillary node involvement (NICE, 2015). However, chemotherapy has serious side effects and is also expensive. Therefore it is of benefit to determine which patient characteristics may optimize the treatment selection process.

## 2.2 Biomarkers

Research investigating biomarkers for early detection, prognosis and the prediction of treatment responses in breast cancer is rapidly expanding (Pultz et al., 2014).

### 2.2.1 Definition

Biological markers, more commonly known as biomarkers, are often defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Atkinson et al., 2001). Biomarkers theoretically have the advantage of improved specificity and reduction of heterogeneity that is present in grouping the population by a phenotypic

marker (European Medicines Agency, 2011). They can be used in multiple research phases, such as non-clinical laboratory development, preclinical animal studies and all phases of a clinical trial.

Biomarkers can generally be distinguished into three types: classifier, prognostic or predictive. A classifier marker is used to select an appropriate target population, such as a DNA marker that does not change over the course of the study (Cosmatos and Chow, 2008).

Prognostic markers have been defined as markers that "classically identify patients with differing risks of a specific outcome, such as progression or death" (Sargent et al., 2005). They inform about the clinical outcomes, independent of treatment, and can be used as a tool for staging of the disease or classification of the extent of the disease. For example, patients with smaller tumours have a better long-term survival than those with larger tumours (Elston et al., 1999) thus making tumour size a prognostic marker.

A predictive marker has been defined as a marker that "predicts the differential efficacy (benefit) of a particular therapy based on marker status". They inform about the treatment effect on the clinical endpoint and predict how individual patients will respond to specific treatments. An example of this is hormone receptor status predicts response to endocrine therapies in breast cancer (Elston et al., 1999). Personalised medicine is replacing the one-drug-fits-all approach with the use of predictive biomarkers to distinguish those who will benefit from a treatment and those who will not and even those who may potentially be harmed (Califf et al., 2010). Improving the ability to predict which patients will benefit from a treatment could also help control medical costs and improve the success rate of clinical drug development.

### 2.2.2 Established Biomarkers in Breast Cancer

*Classic clinicopathological features*

Classical clinicopathological features which indicate a patient's prognosis include tumour size, histological subtype and grade, lymph node metastases, and lymphovascular invasion, which are derived from careful histological analysis of primary breast cancer samples (Weigel and Dowsett, 2010). The TNM (tumour size, nodes, metastasis) system integrates these into tumour stages that have major prognostic value defined by four broad categories:

Stage 1 - early stage breast cancer where the tumour is less than 2 cm across and has not spread beyond the breast.

Stage II – early stage breast cancer where the tumour is either less than 2 cm across and has spread to the lymph nodes under the arm; or the tumour is between 2 and 5 cm (with or without spread to the lymph nodes under the arm); or the tumour is greater than 5 cm and has not spread outside the breast.

Stage III – locally advanced breast cancer where the tumour is greater than 5 cm across and has spread to the lymph nodes under the arm; or the cancer is extensive in the underarm lymph nodes; or the cancer has spread to lymph nodes near the breastbone or to other tissues near the breast.

Stage IV – metastatic breast cancer where the cancer has spread outside the breast to other organs in the body.

*Estrogen Receptor*

One of the most established and important biomarkers in breast cancer is Estrogen Receptor (ER) expression. It is a prognostic marker with ER-positive tumours associated with better survival than ER-negative tumours in the short-term (Fisher et al., 1988). It is also a recognised

predictive marker with ER-positive tumours indicating benefit from endocrine therapy (Goldhirsch et al., 2005).

*Progesterone Receptor*

Estrogen and progesterone receptors (PgR) are strongly linked, with tumours expressing PgR but not ER being uncommon. PgR is also a prognostic factor with higher levels of PgR associated with better outcome compared to lower levels of PgR (Ravdin et al., 1992). However there is no evidence to demonstrate PgR as a predictive marker of treatment with endocrine therapy.

*HER2*

HER2 (human epidermal growth factor receptor 2) is a gene that can play a role in the development of breast cancer (oncogene). It is another prognostic marker, where over expression of HER2 (HER2-positive) has been associated with an increase in the chance of relapsing and tend to have a shorter overall survival (Slamon et al., 1987). It is also a highly predictive marker and HER2-positive breast cancer has been shown to benefit from treatment with trastuzumab (Herceptin)(Romond et al., 2005).

## 2.3 Prognostic Models

Prognostic models are widely used in cancer for investigating patient outcome in relation to multiple patient and disease characteristics (Altman, 2009). Such a model may allow the (reasonably) reliable classification of patients into two or more groups with different prognoses. It may be of particular interest to identify patients with a good prognosis that adjuvant therapy would not be (cost-)beneficial, or a group with a poor prognosis that more aggressive adjuvant therapy would be justified (Clark, 1994).

Defining a prognostic model as a combination of at least two different variables to predict patient outcome, Altman in 2009 identified 54 studies presenting one or more new prognostic

models in a review of published prognostic models in breast cancer (Altman, 2009). However, very few of the prognostic models had been independently evaluated in separate data sets.

In the past 10 years new diagnostic assays have emerged that promise to allow the identification of some women with invasive breast cancer who are at low risk of recurrence and for whom chemotherapy offers toxicity without a clinically meaningful benefit (Bartlett et al., 2013). These diagnostic assays can be put into two groups based on their respective methodology: multiparametric gene expression measurements or extended immunohistochemistry (IHC) testing. The former being more complex and more expensive compared to standard IHC testing.

*Nottingham Prognostic Index (NPI)*

The NPI is one of the oldest prognostic models for breast cancer patients and is often used in clinical practice (Haybittle et al., 1982). It is a simple combination of histopathological examination of tumour size, lymph-node (LN) stage and tumour grading assembled into a prognostic index formula. Several studies have been performed on NPI and confirmed its' prognostic ability (Galea et al., 1992, Balslev et al., 1994, Todd et al., 1987, Guerra et al., 2003, Sundquist et al., 1999).

*Adjuvant! Online*

Adjuvant! Online is a tool for assessing the risks of an individual patient developing recurrent disease and/or dying within 10 years and the magnitude of benefit to be gained by adjuvant therapy for individual patients (Ravdin et al., 2001). It uses the same factors as the NPI as well as ER status, age and comorbidity. Several studies have assessed the validity of the Adjuvant! Online tool (Campbell et al., 2009, Mook et al., 2009, Olivotto et al., 2005, Hajage et al., 2011, Bhoo-Pathy et al., 2012) with limitations including overoptimistic predictions in other populations, overoptimistic results in young and high grade patients, and the need to consider new predictors such as Ki67, HER2 and Mitotic Index (Hajage et al., 2011).

*Oncotype Dx*

Oncotype DX is a PCR (polymerase chain reaction)-based expression assay testing 21 genes found in breast cancer (Paik et al., 2004). The results are reported as a 'recurrence score', ranked between 0 and 100, estimating the risk of recurrence after tamoxifen treatment in node-negative breast cancers. The recurrence score has been proven to be a predictor of outcome and response to adjuvant chemotherapy (Paik et al., 2006, Albain et al., 2010, Dowsett et al., 2009, Mamounas et al., 2010) and has been shown to provide additional prognostic information to Adjuvant! Online (Tang et al., 2011). An American Society of Clinical Oncology Expert Panel reviewed the evidence and recommended the use of Oncotype DX in routine care in 2007 (Harris et al., 2007) and in 2013 NICE (National Institute for Health and Care Excellence) issued guidance for use of Oncotype DX in England and Wales as an option to help doctors decide whether to offer chemotherapy (NICE, 2013). Oncotype DX is being assessed in a randomised trial called TAILORx (Trial Assigning IndividuaLised Options for Treatment)(Zujewski and Kamin, 2008) to evaluate the ability of the recurrence score to guide therapeutic decisions in patients with an intermediate recurrence score. Over 10,000 patients have been recruited into the trial with recruitment complete. Oncotype DX is also being assessed in another trial, the RxPONDER Trial (Rx for Positive Node, Endocrine Responsive Breast Cancer) to determine whether node positive breast cancer patients with low to intermediate recurrence score benefit from chemotherapy. Recruitment began in 2011 with an aim to recruit 4000 women.

*MammaPrint*

MammaPrint is a 70-gene microarray-based expression signature used to estimate the recurrence risk for early-stage breast cancer (van't Veer et al., 2002). It has been shown to provide prognostic information and response to adjuvant chemotherapy (van't Veer et al., 2002, Buyse et al., 2006, Drukker et al., 2013, Drukker et al., 2014, Bueno-de-Mesquita et al.,

2009, van de Vijver et al., 2002). The accuracy of MammaPrint in selecting patients for adjuvant chemotherapy is being compared to Adjuvant! Online in a randomized trial called MINDACT (Microarray In Node-negative and 1–3 node positive Disease may Avoid ChemoTherapy)(Bogaerts et al., 2006) where over 6600 patients have been enrolled and recruitment is complete.

*PAM50*

PAM50 is a multiplex PCR assay that translates expression array data into a clinically viable diagnostic assay (Nielsen et al., 2010, Chia et al., 2012, Parker et al., 2009), using 50 genes to identify molecular subtypes of early breast cancer. Similar to Oncotype DX, this assay also generates a numerical risk score (risk of recurrence), which has been shown to provide more prognostic information in endocrine-treated patients with ER-positive, node-negative disease compared with Oncotype Dx in the trans-ATAC cohort (Dowsett et al., 2013).

*Mammostrat*

The Mammostrat assay relies on the IHC analysis of five markers (p53, NDRG1, SLC7A5, CEACAM5 and HTF9C)(Ring et al., 2006, Bartlett et al., 2010) and classifies patients into three risk groups. First described in 2006, this assay was validated across multiple retrospective institutional and clinical trial cohorts, including the NSABP B-14 and NSABP B-20 trials (Ross et al., 2008). Recent evidence from the TEAM trial suggests that this assay also provides information on residual risk in patients treated with aromatase inhibitors (AI)(Bartlett et al., 2012). After Federal Drug Administration approval of this test as a marker of residual risk in early breast cancer, the assay is available on a commercial basis within the USA.

*IHC4*

The IHC4 residual risk model is an extension of longstanding evidence on the ability of four conventional IHC markers, ER, PgR, HER2 and Ki67 to select patients at increased residual risk after adjuvant endocrine therapy. An algorithm has been developed that integrates these data into a predictor of risk, which has been claimed to provide information equivalent to that from the more complex and expensive Oncotype DX assay (Cuzick et al., 2011). The IHC4+C (IHC markers plus clinicopathologic parameters) score also provides additional information on residual risk of distant recurrence to ER-positive primary breast cancer patients receiving adjuvant endocrine therapy, supplementary to that provided by the Adjuvant! Online and NPI intermediate-risk groups (Barton et al., 2012).

## 2.4 Time-Dependency of Effects

Long-term treatment decisions such as extended adjuvant endocrine therapy beyond 5 years may still rely on biomarker data from the primary diagnostic specimen. This assumes the effects of the markers over long periods of time are constant with no time-dependent effects but there is evidence to suggest this is not the case.

The annual hazard rates for breast cancer deaths (percentage per year) after initial diagnosis among women in the National Cancer Institute's Surveillance, Epidemiology, and End Results 13 Registries database (SEER, 2010) show the risk of breast cancer recurrence and death varies over time, i.e. it is non-proportional according to prognostic and predictive factors (Figure 2.2). The hazard curve for breast cancer deaths peaks between 2 and 3 years after initial diagnosis and then declines sharply suggesting the biological mechanisms responsible for early and late cancer-specific events are fundamentally different (Jatoi et al., 2011).

The effects of estrogen receptor is a well-established marker proven to change over time (Hilsenbeck et al., 1998, Mulligan et al., 2008, Coradini et al., 2000). Patients with ER-negative tumours are recognised as having a higher risk of early relapse than those with ER-

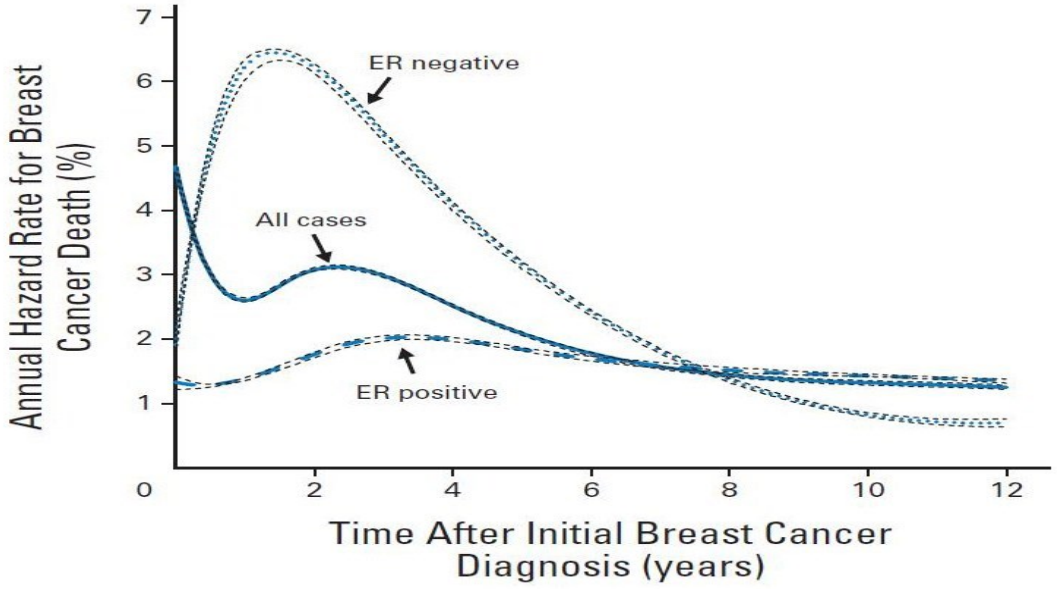



**Figure 2.2** Annual Hazard Rates for Breast Cancer Death : All cases combined, ER-negative and ER-positive breast cancers using the National Cancer Institute's Surveillance, Epidemiology, and End Results 13 Registries Databases (SEER, 2010, Jatoi et al., 2011)

positive tumours. However, overtime the risk of relapse in ER-negative patients decreases whilst the risk in ER-positive patients remains constant. This is displayed in Figure 2.2. The hazard rates for breast cancer death for patients with ER-negative and ER-positive tumours peak between the first and third years after diagnosis with more than a 3-fold difference in risk. However, after the seventh and eighth years their hazard curves cross, after which women with ER-negative tumours have a lower rate of breast cancer death than women with ER-positive tumours.

A classic example discussing time-varying effects is the paper by Gore and colleagues where it was stated that non-monotone convergent hazard functions are associated with most clinical covariates in breast cancer and contradicts the use of proportional hazards models (Gore et al., 1984). Models that assume proportional hazards (e.g. the Cox model) are a popular class of models for analysing survival data that have the assumption that the effects of the covariates do not change over time (introduced in chapter 3). The gene signature MammaPrint has already

been shown to have possible time-dependent effects with better prediction of patients at high risk of early relapse rather than those at risk of later disease progression(Buyse et al., 2006, Desmedt et al., 2007).

It is important to understand time-dependent effects as the different patterns in mortality over time suggests biological mechanisms responsible for early and late breast cancer events differ and may result in differences in responses to treatments. Time-dependence may suggest de novo resistance mechanisms and identify patients who would benefit from switching to another treatment. Subtypes that have time-dependent effects could then potentially be categorised further into two clinically distinctive groups: those with risk of relapse/death at an early stage and those expected to show long-term survival and help find preferential treatment options. Understanding time-dependent effects is also important for trial designs. To fully determine the long-term effects of new therapies, long-term follow up of patients enrolled in breast cancer clinical trials might be necessary. If effects are non-proportional, adaptive trial designs that allow early stoppage of trials may prevent finding important late effects. On the other hand some adjuvant treatments may only have early benefits and longer follow-up may dilute those effects. This will need to be considered in both the design and analysis of such trials.

In a clinical setting, time-dependency of risk-factors is relevant as it is important to identify patients who are at high risk of late recurrence. Clinicians must consider whether the long-term risk of recurrence of primary or secondary breast cancer is the same in all patients and therefore consider the best treatment options. Two studies, the Adjuvant Tamoxifen: Longer Against Shorter (ATLAS) trial and the Adjuvant Tamoxifen—To Offer More? (aTTom) study now support longer-term tamoxifen use for up to 10 years (Davies et al., 2013, Gray et al., 2013). It has also been shown that extended adjuvant endocrine therapy with letrozole improves disease-free survival of postmenopausal women with estrogen receptor or progesterone receptor positive tumours after 5 years of tamoxifen therapy (Goss et al., 2003).

## 2.5 Discussion

Compared with other clinical subgroups, ER-positive, lymph-node negative patients have the best overall prognosis, however many recurrences occur in patients who remain disease-free for 5 years (Early Breast Cancer Trialists' Collaborative, 2005, Blows et al., 2010, Cheng et al., 2012).

Assessment of risk of recurrence over time is clinically important to identify patients with a low risk of recurrence for avoidance of chemotherapy and those with high risk of recurrence beyond 5 years for the decision of whether to extend adjuvant endocrine therapy.

# Chapter 3: Survival Analysis

## 3.1 Introduction

This project involves analysis of time-to-event data and therefore requires the use of survival analysis, a class of statistical methods for studying the occurrence and timing of events (Ata and Sozer, 2007). It is widely used in oncology since we are often interested in studying a delay, such as the time from cancer diagnosis or treatment commencement to cancer recurrence or death (Bellera et al., 2010). In this chapter, the basic concepts and theories involved will be introduced with a focus on the Cox proportional hazards model.

## 3.2 General Principles

In survival analysis the term 'failure' is used to define the occurrence of the event of interest and 'survival time' is the length of time taken for the failure to occur.

Traditional statistical methods, such as ordinary least squares regression, cannot be used as not everyone experiences the event of interest (censoring) and the distribution of time-to-event outcome is usually not normally distributed. Throughout this project we will be concerned with right-censoring. Right-censoring is when it is known that the event of interest occurs sometime after the recorded follow-up period or the event might never occur. So at the time of observation (or end of follow-up) the relevant event has not yet occurred (the event-date is unknown) and so the total length of time between entry and exit is unknown. Given entry at time 0 and observation at time *t*, we only know that the completed spell is of length $T > t$, i.e. their time-to-event will be longer than their time in the study.

## 3.3 The Survival Function and Hazard Rate

The survival function, S(t), measures the proportion of subjects who have not experienced the event as a function of time. Let T be a continuous non-negative random variable denoting the time of occurrence of the event of interest. The survival function can then be written as

$$S(t) = P(T \geq t) \ , \qquad 3.1.$$

with probability density function (pdf) $f(t)$ and cumulative distribution function (cdf) $F(t)$ defined as

$$F(t) = P(T < t) \ , \qquad 3.2.$$

the probability that the event has occurred by time t, sometimes referred to as the failure function. The survival function is related to the cdf:

$$S(t) = 1 - F(t) = \int_t^{\infty} f(x)dx \ , \qquad 3.3.$$

the probability that the event of interest has not occurred by time t. The survival function can only decrease and at time 0 none of the subjects have experienced the event giving S(t)=1.

A very important concept in survival analysis is the hazard function, h(t), and is defined by

$$h(t) = \lim_{\partial \to 0} \frac{P(t \leq T \leq t + \partial | T \geq t)}{\partial} \ . \qquad 3.4.$$

The hazard is the instantaneous failure rate at time t which is the probability of the event occurring in the next instant given that the subject is currently alive. The hazard function can be written as

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = \frac{-d}{dt}\log\big(S(t)\big) \ , \qquad 3.5.$$

and the survival function can therefore be written in terms of the hazard function:

$$S(t) = \exp\left(-\int_0^t h(u)du\right) \quad . \qquad 3.6.$$

The hazard function is an estimate of the incidence rate (as a function of time) and is measured in events per unit time. It is not a probability and therefore can take values between 0 and infinity. A lower hazard rate implies a higher survival function.

Another quantity of interest is the cumulative hazard function, defined as

$$H(t) = \int_0^t h(u)du \quad . \qquad 3.7.$$

The cumulative hazard function is a measure of how much 'hazard' a subject has been exposed to at time *t* and is used as the basis for a number of residuals. The survivor function is related to the cumulative hazard function by

$$S(t) = \exp(-\mathrm{H(t)}) \quad . \qquad 3.8.$$

## 3.4 Non-Parametric Estimates

The survival function is often estimated non-parametrically, a simple approach which is free of assumptions.

### 3.4.1 Kaplan-Meier

The most commonly used method is the Kaplan-Meier (or product-limit) estimate (Kaplan and Meier, 1958) which involves estimating S(t) at discrete values of time *t*. It is based on conditional probabilities as we are interested in the probability of a patient surviving the next unit of time given that they have survived so far.

The Kaplan-Meier estimate only needs to be estimated at each unique failure time. Let $t_k$ be the survival time for the k[th] unique failure time. We then need to obtain an estimate of $p_k$, the probability a subject survives the interval (k-1,k) given they were alive at time (k-1):

$$\begin{aligned} p_k &= \frac{No.of\ subjects\ who\ survived\ the\ time\ interval\ (k-1,k)}{Number\ of\ subjects\ at\ risk\ at\ time\ k-1} \\ &= \frac{n_k - d_k}{n_k} \\ &= 1 - \frac{d_k}{n_k}, \end{aligned} \qquad 3.9.$$

where $n_k$ is the number of patients alive at the start if the interval of interest and $d_k$ is the number of patients dying in the interval.

The Kaplan-Meier survival function estimate at the k[th] time point is

$$\begin{aligned} \hat{S}(t_k) &= \prod_{i=1}^{k} 1 - \frac{d_i}{n_i} \quad , \\ &= \hat{S}(t_{k-1})\left(1 - \frac{d_k}{n_k}\right) \quad . \end{aligned} \qquad 3.10.$$

The Kaplan-Meier estimator is a useful descriptive tool but it has not been designed to incorporate several covariates simultaneously.

### 3.4.2 Nelson-Aalen Estimator

Another non-parametric method is to estimate the cumulative hazard H(t) using the Nelson-Aalen estimator:

$$\widehat{H}(t_i) = \sum_{j=1}^{i} \frac{d_j}{n_j} \quad , \qquad 3.11.$$

where this expression is estimating the hazard at each distinct time of death $t_j$ as the ratio of the number of deaths to the number exposed. The cumulative hazard up to time t is simply the

sum of the hazards at all death times up to t, and has an intuitive interpretation as the expected number of deaths in (0, t] per unit at risk (Rodriguez, 2005).

## 3.5 Proportional Hazards Models

Considering two types of proportional hazards models: parametric, which assume some functional form for the baseline hazard e.g. Exponential and Weibull models; and semi-parametric models, which make no assumption about the shape of the baseline hazard function, e.g. the Cox proportional hazards model.

### 3.5.1 The Exponential Model

The simplest assumption for the hazard is when we assume that it is constant over time. Thus

$$\begin{aligned} h(t) &= \lambda \quad , \\ \Rightarrow S(t) &= \exp(-\lambda t) \ , \\ \Rightarrow \ f(t) &= \lambda \exp(-\lambda t) \, . \end{aligned} \tag{3.12.}$$

This gives $f(t)$ to be the same as the pdf of the exponential distribution, therefore assuming the hazard rate is constant assumes the survival times have an exponential distribution.

### 3.5.2 The Weibull Model

An alternative and more flexible form of the hazard function is

$$\begin{aligned} h(t) &= \lambda\gamma t^{\gamma-1} \quad , \\ \Rightarrow S(t) &= \exp(-\lambda t^{\gamma}) \ , \\ \Rightarrow \ f(t) &= \ \lambda\gamma t^{\gamma-1} \exp(-\lambda t^{\gamma}) \quad . \end{aligned} \tag{3.13.}$$

This gives $f(t)$ to be the same as the pdf of the Weibull distribution and if $\gamma = 1$ then the Weibull model reverts to the exponential model. Although this model is more flexible in modelling the hazard function it restricts the function for the hazard to be monotonic.

### 3.5.3 The Cox Model

The Cox model introduced by David R. Cox in 1972 (Cox, 1972) is one of the most cited statistical papers (Ryan and Woodall, 2005) and is the most commonly used approach for analysing survival time data in medical research (Bradburn et al., 2003). The Cox proportional hazards model is a semi-parametric model as it makes no assumption about the shape of the underlying hazard. The result derives from innovative use of the proportional hazard assumption and a partial likelihood (PL) method of estimation rather than maximum likelihood.

Let $x_1, x_2 \cdots\cdots, x_p$ be the values of $p$ covariates $X_1, X_2 \cdots\cdots, X_p$. The Cox model models the hazard function and parameters are estimated on a log scale:

$$h(t; x) = h_0(t) exp\left(\sum_{j=1}^{p} \beta_j x_j\right), \qquad 3.14.$$

where $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$ is a 1 x p vector of regression parameters and $h_0(t)$ is the baseline hazard function. This model has the assumption of log-linearity i.e. the effect of the predictors act linearly on the log-hazard. The proportional hazards (PH) assumption corresponds to assuming that the hazard ratio comparing one level of a covariate to another is constant over time. The effect of a covariate $x_1$ can be expressed by the hazard ratio:

$$h(t, x_1 + 1) \Big/ h(t, x_1) = \exp(\beta_1)\,, \qquad 3.15.$$

and is a measure of the change in the hazard function with respect to a unit increase in $x_1$.

If we integrate both sides of equation 3.14 between 0 and t we get

$$H_i(t) = H_0(t) \exp(\beta^T X_i)\ \ , \qquad 3.16.$$

and we have proportional cumulative hazards. Changing signs and exponentiation gives

$$\exp(-H_i(t)) = \exp(-H_0(t)\exp(\beta^T X_i)) \quad ,$$
$$S_i(t) = S_0(t)^{\exp(\beta^T X_i)} \quad . \qquad 3.17.$$

This is a complex function and it is difficult to assess proportional hazards from a graph of survival curves.

Since $h_0(t)$ is free from parametric assumption, it is not possible to apply the full likelihood function for estimating β's. Cox suggested an estimation procedure in which the analysis concentrates only on the effect of covariates and leaving $h_0(t)$ completely unspecified (Cox, 1972, Cox, 1975),. The coefficients (β) in a Cox proportional hazards model can be estimated using the partial likelihood method, which is a function of observed survival times and unknown parameters:

$$\prod_{i=1}^{n}\left[\frac{\exp(\sum_{j=1}^{p}\beta_j x_j)}{\sum_{l\in R(t_{(i)})}\exp(\sum_{j=1}^{p}\beta_j x_j)}\right] . \qquad 3.18.$$

Where $i = 1,2\cdots n,$ denote the n ordered exact failure times for each subject i, and $R(t_{(i)})$ consists of all individuals whose survival times are at least $t_{(i)}$. The summation in the denominator of the likelihood is the sum of the values of $\exp\left(\sum_{j=1}^{p}\beta_j x_j\right)$ over all individuals who are at risk at this time. The likelihood depends only on the rank of failure times and the variance of the partial likelihood is larger than the variance of the complete likelihood. Tied survival times can be handled using the likelihood form of Breslow (Breslow, 1974), the default method in most statistical packages, however the bias can be substantial if the number of ties is large (HertPicciotto and Rockhill, 1997).

#### *3.5.3.1 Assessment of Log-Linearity Assumption*

The inclusion of non-linear covariate effects in Cox regression is common with many straight-forward options available to implement them.

A graphical approach involves the use of Martingale residuals (Therneau et al., 1990). Most residuals are based around each subject's cumulative hazard at their event (or censoring) time.

The interest is in whether some subjects have a large amount of hazard without having the event while others have the event quickly. The cumulative hazard for the i$^{th}$ subject up to the time of their death or censoring is estimated from a Cox PH model as

$$\hat{r}_{Ci} = \hat{H}_o(t_i) \exp(\hat{\beta}' x_i) \quad , \qquad 3.19.$$

known as Cox-Snell residuals but they have no easy interpretation. Martingale residuals can be defined by

$$r_{Mi} = d_i - r_{Ci} \, , \qquad 3.20.$$

where $d_i = 1$ if the i$^{th}$ subject had an event and $d_i = 0$ if the i$^{th}$ subject had a censored observation. These can be interpreted as the observed number of events for the i$^{th}$ individual minus the total amount of hazard experienced by that individual, a type of observed minus expected. Plotting the Martingale residuals against the continuous variables allows you to check the functional form of the covariate.

A simple method is to categorise a continuous variable into several groups, estimate a separate effect for each group and check if the estimates show a linear trend. However, the usual disadvantages of categorising a continuous variable arise such as discarding potentially important quantitative information, thus reducing the power to detect a true association with survival, the choice of interval is subjective and results in an un-realistic step function. Another common practice is to transform the variable using some common mathematical functions, such as logarithm, polynomial, fractional polynomial (Royston and Altman, 1994) etc. and determine if this provides a better fit to the model. Regression spline methods can also be used to characterise non-linear effects (Gray, 1992, Osullivan, 1988).

#### *3.5.3.2 Assessment of Proportional Hazards Assumption*

Assessment of the proportional hazards assumption can be done by many numerical or graphical approaches. There are a large variety of methods to check for non-PH and Ng'Andu

provides an overview of five test statistics to check the PH assumption of the Cox model (Ngandu, 1997).

Some of the methods, briefly, are as follows: In his original paper, Cox suggested testing the PH assumption by including a time-dependent covariate (interaction terms with time) in the model and determining whether this improved the fit of the model (Cox, 1972). A test was developed by Harrell based on the correlation between the Schoenfeld partial residuals (Schoenfeld, 1982) and the rank order of the failure time (Harrell, 1986). There are a set of Schoenfeld residuals for each covariate in the model and are defined as

$$r_{Sij} = d_i \left( x_{ij} - E\left(x_{ij} \middle| R_i\right)\right) , \qquad 3.21.$$

where $R_i$ is the set of individuals at risk at time $t_i$ and $E\left(x_{ij}\middle|R_i\right) = \frac{\sum_{l \epsilon R(t_i)} x_{lj} \exp(\beta^T x_{lj})}{\sum_{l \epsilon R(t_i)} \exp(\beta^T x_{lj})}$ . The Schoenfeld residual for a particular subject and particular covariate is the difference between the observed value of the covariate and the expected, conditional on the risk set $R_i$ of subjects at risk at time $t_i$ .

Grambsch & Therneau developed a score test which is a test of non-zero slopes in a generalised linear regression of the scaled Schoenfeld residuals on chosen function(s) of time (Grambsch and Therneau, 1994). Therneau and colleagues also considered a score process using Martingale residuals (Therneau et al., 1990). Finally omnibus goodness-of-fit tests, such as those developed by Moreau and colleagues have been developed to detect non- proportional hazards by approximating the functional form for the time-dependence with a piecewise linear function constant over *a priori* specified time intervals (Moreau et al., 1985, Moreau et al., 1986*)*,.

Graphical approaches have been criticised; interpreting graphical plots can be arbitrary and the conclusions from the graphs depend on the subjectivity of the researcher. Some of the graphical approaches are log-minus-log survival plots of survivor functions, comparison

between survival curves based on the Cox regression model and Kaplan-Meier estimates and a smoothed plot of scaled Schoenfeld residuals versus time (Ata and Sozer, 2007).

#### *3.5.3.3 Limitations of the Cox model*

The assumptions of the traditional Cox PH model are often violated when modelling prognostic factors in cancer studies. With classic examples given by Valsecchi in ovarian cancer and Gore in breast cancer (Valsecchi et al., 1996, Gore et al., 1984). Often the assumptions are not verified. In a 1995 review of cancer publications using a Cox model, Altman and colleagues reported that most studies did not report attempts to verify the PH assumption (Altman et al., 1995); similar findings were reported more recently in 2008 by Mathoulin-Pelissier and colleagues (Mathoulin-Pelissier et al., 2008).

When the assumptions of the Cox PH model are not met, it is possible that subsequent analyses and risk estimates will be biased.

Bellera et al. state

'for variables not satisfying the proportionality assumption, the power of the corresponding tests is reduced, that is, we are less likely to conclude for a significant effect when there is actually one. If the hazard ratio is increasing over time, the estimated coefficient assuming PH is overestimating at first and underestimating later on. For those variables of the model with a constant hazard ratio, the power of tests is also reduced as a consequence of an inferior fit of the model' (Bellera et al., 2010).

Spurious non-proportional effects may also be introduced by incorrectly modelling other parts of the data, such as omission of an important covariate, an incorrect functional form of a continuous covariate or an inappropriate survival model (Buchholz and Sauerbrei, 2011). Therefore it is important to consider both the assumption of linearity and proportional hazards.

As stated by Abrahamowicz et al.

‘[T]he proportional hazards (PH) assumption… implies that the impact of each covariate on hazard remains constant during the entire follow-up time. While testing the PH assumption is interesting in its own right, simultaneous modelling of nonlinear and time dependent effects of the exposure of interest may be necessary to avoid biased estimates and incorrect conclusions’ (Abrahamowicz et al., 2003).

The baseline hazard function is completely unspecified in the Cox PH model. This can be seen as a great advantage (one avoids potential problems from specifying the wrong shape), but some may also see it as a disadvantage if one is particularly interested in the shape of the baseline hazard function for its own sake. For instance, the behaviour of the hazard function is of potential medical interest because it is directly related to the time-course of an illness (Royston and Parmar, 2002b). Modelling the baseline hazard also allows estimation of absolute risks which can provide additional information when presented alongside relative risks.

# Chapter 4: Literature Review

## 4.1 Introduction

The limitations of the Cox model have been discussed in chapter 3, in particular the assumption of proportional hazards which is often invalid for current prognostic models and biomarkers. A review of existing approaches for the analysis of non-proportional effects with respect to survival data was performed. The aim was to identify available approaches for investigating non-proportional effects, summarise advantages and disadvantages and perform citation searching to determine which approaches have been used in practice.

## 4.2 Methods

A literature search was conducted on the 27/02/12 using EMBASE (1947 to present), MEDLINE (1946 to present) and Web of Science (WoS) (1899 to present), which are online information sources of published literature, using the search strategy in Figure 4.1.

The search strategy was developed based on the following inclusion and exclusion criteria:

| Inclusion | Exclusion |
|---|---|
| • Articles that discuss and propose **statistical** approaches (methodological papers) for the analysis of **time-dependent/non-proportional effects** in **survival** data. | • Studies investigating **time-dependent covariates** - i.e. covariates whose values change over time.<br>• Approaches based on different methodological assumptions – e.g. **case-control studies, relative survival** or **competing risks.** |

EMBASE and MEDLINE

1. A search for terms *time-dependent* or *time-varying* or *non-proportional* or *time-by-covariate interaction* or *time-by-treatment interaction* in titles or abstracts (non-hyphenated terms also used).
2. A search for *statistics as topic* (major focus of the article) mapped to: biostatistics (mesh) OR data interpretation, statistical (mesh) OR models, statistical (mesh, major focus) OR statistical distributions (mesh) OR statistics, nonparametric (mesh) OR survival analysis (mesh, major focus) OR survival (mesh, major focus) OR proportional hazards model (mesh, major focus).
3. A search for the term *flexible* in titles or abstracts and *survival analysis* or *survival* as a major focus of the article.
4. A combined search of 1 and 2 or 3.
5. Exclusion of studies found in 4 if they contained the terms *relative survival* or *case-control* or *competing risks* in titles or abstracts.
6. Further exclusion of remaining studies in 5 if they contained the terms *repeated measures* or *longitudinal data* or *longitudinal covariate* in the title or abstract.

Web of Science

7. A search for terms *time-dependent* or *time-varying* or *non-proportional* or *time-by-covariate interaction* or *time-by-treatment interaction* in titles (non-hyphenated terms also used).
8. A search for *survival* or *Cox* or *proportional* or *hazards* in titles and abstracts.
9. A search for the term *flexible* in titles.
10. A combined search of 7 and 8 or 8 and 9.
11. Exclusion of studies found in 10 if they contained the terms *relative survival* or *case-control* or *competing risks* in titles or abstracts.
12. Further exclusion of remaining studies in 11 if they contained the terms *repeated measures* or *longitudinal data* or *longitudinal covariate* in the title or abstract.

**Figure 4.1** Literature Review Search Strategy

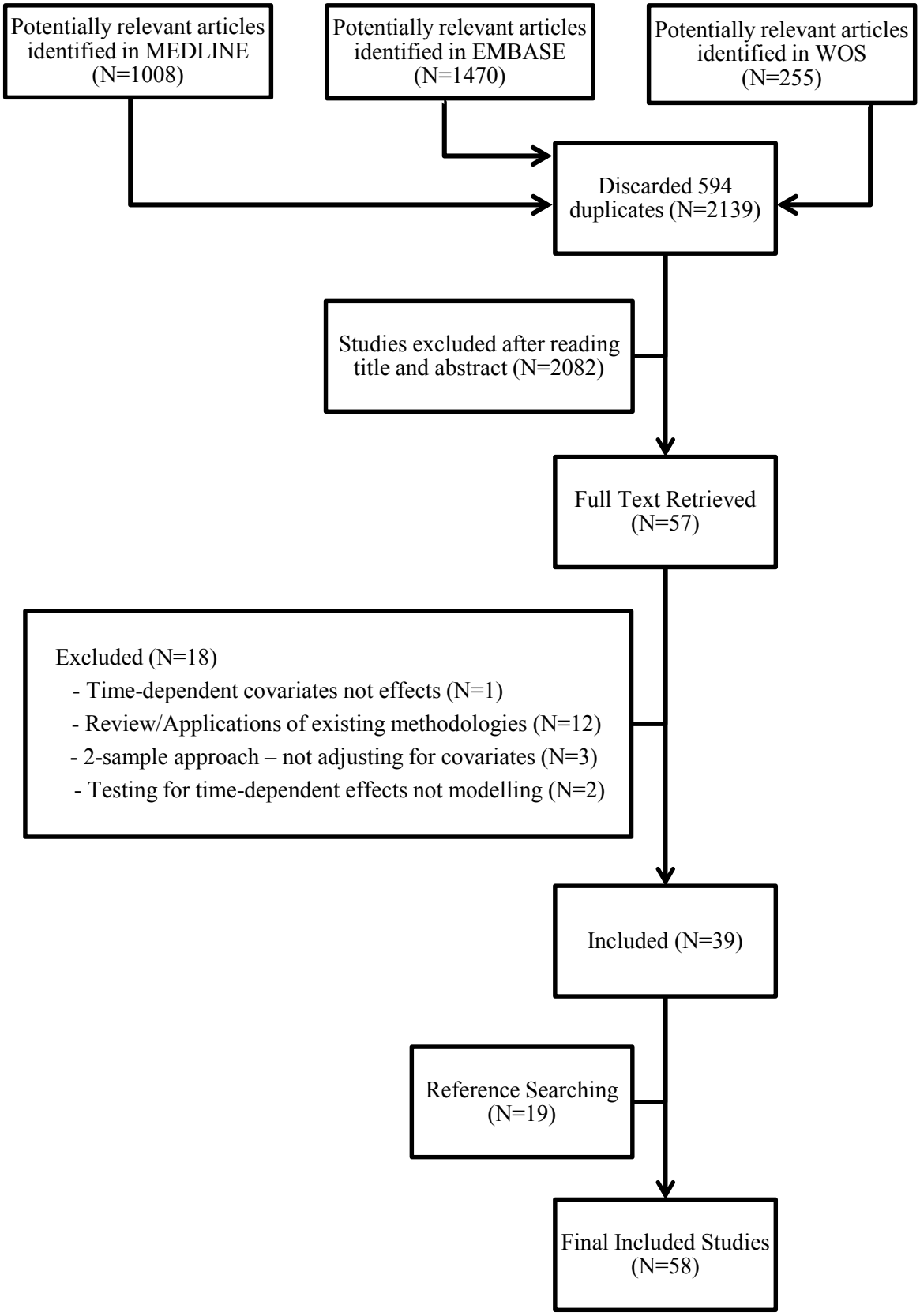



**Figure 4.2** Flowchart of Eligibility Assessment and Inclusion

## 4.3 Results

The search was carried out in February 2012 and returned 2,139 potentially relevant articles from MEDLINE, EMBASE and WoS after excluding duplicates (Figure 4.2). Titles and abstracts were screened and 57 full text papers were retrieved for further evaluation. A further 18 papers were excluded, mainly due to the articles being reviews or applications of existing methodologies. The references were screened for additional papers and this resulted in a further 19 papers being included in the final review.

The 58 papers included in the final review highlighted five broad classifications of models for investigating non-proportional effects: multiplicative hazards models (Table 4.1), additive hazards models (Table 4.5), additive-multiplicative hazards models (Table 4.5), parametric models and others (Table 4.7).

## 4.4 Multiplicative Hazards Models (Table 4.1)

Multiplicative hazard models are models where the effects of the covariates are modelled on a multiplicative scale. The Cox model introduced by Cox in 1972 (Cox, 1972), is an example of a multiplicative hazards model with the assumption that the hazard ratios are constant over time:

$$\lambda(t;x) = \lambda_0(t)exp\left(\sum_{i=1}^{p}\beta_i x_i\right). \qquad 4.1.$$

The literature search identified 4 multiplicative hazard models for incorporating non-proportional effects (Table 4.1) and the advantages and disadvantages of these approaches are given in Table 4.2.

**Table 4.1 Multiplicative Hazard Models**

| Model Name | Formula | Author (Year) | Citation Searching* |
|---|---|---|---|
| Stratified Cox Model | $\lambda(t;X,Z) = \lambda_{0k}(t)\exp\{X_1(t)\beta_2 + \cdots + X_p(t)\beta_p\}$ | Dabrowska (1997)<br>Wei & Schaubel (2008) | 47<br>9 |
| Weighted Cox Model | $AHR = \frac{\int\left(\frac{h_1(t)}{h(t)}\right)w(t)f(t)dt}{\int\left(\frac{h_0(t)}{h(t)}\right)w(t)f(t)dt}$ | Kalbfleisch & Prentice (1981)<br>Schemper (1992)<br>Schemper et al. (2009) | 36<br>60<br>29 |
| Smoothed Schoenfeld Residuals | $r_{Sij} = d_i\left(x_{ij} - E\left(x_{ij}\middle\vert R_i\right)\right)$ | Pettitt & Daud (1990)<br>Grambsch & Therneau (1994) | 36<br>1,518 |
| Extended Cox Model | $\lambda(t;x) = \lambda_0(t)exp\left(\sum_{i=1}^{p}\beta_i(t)x_i\right)$ | See Table 4.3 | See Table 4.3 |

* Refer to section 4.9

**Table 4.2 Advantages/Disadvantages of Multiplicative Hazards Models**

| Model Name | Advantages | Disadvantages |
|---|---|---|
| Stratified Cox Model | • Simple and straightforward approach<br>• Useful when covariate with non-proportional effect is not of direct interest | • Does not provide information on the non-proportional effect<br>• Loss of power due to categorisation of the data<br>• More than one covariate with a time-varying effect will further reduce the power due to stratifying across multiple variables |
| Weighted Cox Model | • Intuitive interpretability of the resulting average hazard ratio<br>• Reduces complexity of analysis and therefore more suitable for small sample sizes | • Does not provide information on the shape of the non-proportional effect |
| Smoothed Schoenfeld Residuals | • Use as either a diagnostic or estimation technique | • Uncertainty estimates that go along with these estimates are not easy to use in practice |
| Extended Cox Model | • See Table 4.4 | • See Table 4.4 |

### 4.4.1 The Stratified Cox Model

A simple option is to use a Cox model stratified by the variable suspected to have a time-varying effect. The variable needs to be categorical or if continuous it must be categorised. Each stratum *k* has a distinct baseline hazard but common values for the coefficient vector β. If a categorical covariate $Z_1$, with levels $k = 1, \ldots, K$, is responsible for the non-proportionality, the model can be extended to include K strata:

$$\lambda(t; X, Z) = \lambda_{0k}(t) \exp\{X_1(t)\beta_2 + \cdots + X_p(t)\beta_p\}. \qquad 4.2.$$

Stratifying assumes that the other covariates are acting in the same way in each stratum and therefore hazard ratios are similar across strata (Bellera et al., 2010). Dabrowska (1997) studied a general stratified Cox's regression model and Wei & Schaubel (2008) proposed an estimator for the aggregate treatment effect under non-proportional hazards based on treatment-specific baseline cumulative hazards estimated under a stratified Cox model (Dabrowska, 1997, Wei and Schaubel, 2008). The main disadvantage is it does not provide information on the time-dependent effect.

### 4.4.2 The Weighted Cox Model

When the proportional hazards assumption is violated, the average hazard ratio from a Cox model is often under- or overestimated (Schemper et al., 2009). One approach to deal with this is using weighted estimation in Cox regression (Schemper, 1992) and the methodology has been reviewed, including recent developments, by Schemper and colleagues (Schemper et al., 2009) They state their preferred definition of an average hazard ratio (AHR), introduced by Kalbfleisch & Prentice (Kalbfleisch and Prentice, 1981) , is

$$AHR = \frac{\int \left(\frac{h_1(t)}{h(t)}\right) w(t) f(t) dt}{\int \left(\frac{h_0(t)}{h(t)}\right) w(t) f(t) dt}, \qquad 4.3.$$

where $h_0(t)$ and $h_1(t)$ denote the hazards of groups $G_0$ and $G_1$ at time *t*, respectively, and $h(t) = h_0(t) + h_1(t)$ .The weight function w(t) is chosen to reflect the relative importance attached to the hazard ratios in different time periods, with the most basic choices being w(t)=1 and w(t)=S(t), the survival function. In the population definition of the AHR the density of the events in time (which result in contributions to the partial likelihood in Cox's model) is needed and symbolised by $f(t)$. This is the simple two-sample case without censoring and is extended to weighted estimation within Cox's model where a vector $\beta$ of *k* regression parameters is to be estimated. Weighted maximum likelihood estimates $\hat{\beta}_r$ of regression parameters $\beta_r$ , $1 \leq r \leq k$ , are derived as solutions to weighted score equations

$$\sum_{j=1}^{m} w(t_j) \frac{\partial l_j}{\partial \beta_r} = 0 \ , \qquad 4.4.$$

where $l_j$ is the contribution to the log-likelihood at failure time $t_j$. With censored samples, $w(t) = \hat{S}(t)\hat{G}(t)^{-1}$ is required for a weighted Cox regression where $\hat{G}(t)$ denotes the Kaplan-Meier estimator of the censoring or potential follow-up distribution. This weighting allows an intuitive interpretability of the resulting average hazard ratio $\exp(\hat{\beta})$, independent of proportional hazards.

### 4.4.3 Smoothed Schoenfeld Residuals

Pettitt & Daud and Grambsch & Therneau considered smoothing the Schoenfeld residuals to estimate the time-varying regression effects (Pettitt and Daud, 1990, Therneau et al., 1990, Grambsch and Therneau, 1994). Firstly a Cox model is fitted with the relevant predictors. There are a set of Schoenfeld residuals for each covariate in the model and are defined as

$$r_{Sij} = d_i \left( x_{ij} - E\left(x_{ij} \middle| R_i\right) \right) \ , \qquad 4.5.$$

where $d_i$ is the event indicator, $R_i$ is the set of individuals at risk at time $t_i$ and $E\left(x_{ij} \middle| R_i\right) = \frac{\sum_{l \epsilon R(t_i)} x_{lj} \exp(\beta^T x_{lj})}{\sum_{l \epsilon R(t_i)} \exp(\beta^T x_{lj})}$ . The Schoenfeld residual for a particular subject and particular covariate

is the difference between the observed value of the covariate and the expected, conditional on the risk set $R_i$ of subjects at risk at time $t_i$ . Grambsch & Therneau showed that $E(r^*_{Sij}) + \hat{\beta} \approx \beta(t_j)$, where $r^*_{Sij}$ is the scaled Schoenfeld residual (Grambsch and Therneau, 1994). The residuals are then smoothed, with Pettitt & Daud considering a class of smoothers which fit lines or curves to the points using local information and take into account the differing variances (Pettitt and Daud, 1990). A smoothed plot of the scaled Schoenfeld residuals versus respectively observed failure times and their ranks is then used to estimate the shape of the non-proportional effect. This approach is more advantageous as the use of a diagnostic plot to test the PH assumption.

### 4.4.4 The Extended Cox Model

An extension of the Cox model is to let the effects depend on time:

$$\lambda(t; x) = \lambda_0(t) exp\left(\sum_{i=1}^{p} \beta_i(t) x_i\right) \quad , \qquad 4.6.$$

where $\beta(t) = \left(\beta_1(t), \dots, \beta_p(t)\right)$ is a p-dimensional time-varying regression coefficient.

All covariates might not have non-proportional effects and this leads on to the semi-parametric version of the extended Cox model:

$$\lambda(t; x, z) = \lambda_0(t) \exp\left(\sum_{i=1}^{p} \beta_i(t) x_i + \sum_{j=1}^{q} \gamma_j z_i\right) \quad , \qquad 4.7.$$

where X and Z are $p$ and $q$ dimensional covariates respectively and the parameters of the model are the non-parametric $p$-dimensional $\beta(t)$ and $q$-dimensional regression parameter $\gamma$.

Various approaches for estimation of $\beta(t)$ are available (Table 4.3) with their respective advantages and disadvantages given in Table 4.4.

**Table 4.3** Approaches for Estimation of $\boldsymbol{\beta(t)}$ in the Extended Cox Model

| **Approach** | **Formula/Notes** | **Author (Year)** | **Citation Searching*** |
|---|---|---|---|
| Non-Parametric Estimation | $\lambda(t;x) = \lambda_0(t)exp\left(\sum_{i=1}^{p}\beta_i(t)x_i\right)$ | Murphy & Sen (1991)<br>Verweij & Houwelingen (1995)<br>Zucker & Karr (1990)<br>Hastie & Tibshirani (1993)<br>Martinussen et al. (2002)<br>Cai & Sun (2003)<br>Tian et al. (2005) | 99<br>64<br>120<br>903<br>59<br>67<br>62 |
| Piecewise Constant Effects | $\lambda(t;x) = \begin{cases}\lambda_0(t)\exp(\alpha^T x), & t \le \boldsymbol{B}\\ \lambda_1(t)\exp(\gamma^T x), & t > \boldsymbol{B}\end{cases}$ | Anderson & Senthilselvan (1982)<br>Moreau et al. (1985) | 73<br>100 |
| Simple Time-by-Covariate Interactions | $\lambda(t;x) = \lambda_0(t)exp\left(\sum_{i=1}^{p}\beta_i x_i f_i(t)\right)$ | Cox (1972) | 31,055 |
| Polynomials and Fractional Polynomials | $\lambda(t;X) = \begin{cases}\lambda_o(t)\exp[X(\beta_0+\beta_1 t^{q_1}+\beta_2 t^{q_2})], & q_1 \neq q_2\\ \lambda_o(t)\exp[X(\beta_0+\beta_1 t^{q}+\beta_2 t^{q}\log(t))], & q_1 = q_2 = q\end{cases}$ | Stablein et al. (1981)<br>Berger et al. (2003)<br>Sauerbrei et al. (2007) | 75<br>24<br>23 |
| Regression Splines | $\lambda(t;x) = \lambda_0(t)\exp\left(\sum_j f_j(t)x_{ij}\right)$ ,<br>where the functions $f_j(t) = \sum_k \theta_{jk}\beta_{jk}(t)$ are modelled with spline basis functions. | Gray (1992)<br>Hess (1994)<br>Abrahamowicz et al. (1996)<br>Brown et al. (2007)<br>Abrahamowicz & Mackenzie (2007) | 277<br>107<br>109<br>6<br>35 |

**Table 4.3 (Continued) Approaches for Estimation of $\boldsymbol{\beta(t)}$ in the Extended Cox Model**

| Approach | Formula/Notes | Author (Year) | Citation Searching* |
|---|---|---|---|
| Reduced Rank Regression Models | $\lambda(t;X) = \lambda_0(t)exp\left\{\sum_{k=1}^{r}(X\beta_k)(F(t)\gamma_k)\right\}$ | Perperoglou et al. (2006a) | 17 |
| Other Approaches | Neural Networks | Biganzoli et al. (1998) | 128 |
| Bayesian Approaches | Bayesian Dynamic Survival Model | Gamerman (1991) | 144 |
| | Hierarchical Cox Model | Sargent (1997) | 38 |
| | Hierarchical Bayes Model | Gustafson (1998) | 18 |
| | Nonparametric Bayesian Approach | Mckeague & Tighiouart (2000) | 19 |
| | Hierarchical Nonparametric Bayesian Approach | Haneuse et al. (2008) | 3 |
| | Geoadditive Survival models | | |
| | Fully Bayesian framework | Hennerfeind et al. (2006) | 66 |
| | Empirical Bayes approach | Kneib & Fahrmeir (2007) | 51 |
| | Bayesian penalised spline models | Costa & Shaw (2009) | 8 |
| | Linear Bayesian (LB) estimation method | He et al. (2010) | 3 |
| | Monotone hazard ratio using Cox's model | Kim et al. (2011) | 0 |

* Refer to section 4.9

**Table 4.4** Advantages/Disadvantages of the Approaches for Estimation of $\boldsymbol{\beta(t)}$ in the Extended Cox Model

| Model Name | Advantages | Disadvantages |
|---|---|---|
| Non-Parametric Estimation | • Flexible tools for testing the PH assumption | • Complicates investigation into shape of non-proportional effect<br>• Interpretation of cumulative coefficient functions are not straightforward |
| Piecewise Constant Effects | • Simple approach; represent as short and long term effects | • Need sufficient numbers of events in each step of the time scale<br>• Choice of cut-points are subjective<br>• Reduction of power due to categorisation of the data<br>• Proportional hazards still assumed within each step |
| Simple Time-By-Covariate Interactions | • Easy to implement | • Inference dependent on parametric function of time<br>• Simple options restrictive on available shape of non-proportional effect<br>• Inappropriate choice of function may lead to incorrect interpretation of the true effect. |
| Polynomials and Fractional Polynomials (FP) | • FPs allow a flexible family of shapes to model the non-proportional effects<br>• Implementation and inference is straightforward using standard estimation techniques as the approach preserves the linear structure of the predictor | • The global (non-local) definition of FPs still suffers from the same restrictions as other non-local definitions for smooth function estimation.<br>• Involves prior assumptions on the dynamic structure which are often data driven |

**Table 4.4 (Continued) Advantages/Disadvantages of the Approaches for Estimation of $\boldsymbol{\beta(t)}$ in the Extended Cox Model**

| Model Name | Advantages | Disadvantages |
|---|---|---|
| Splines | • Flexible functions<br>• Allows easily verifying the existence of time-variation | • They demand prespecifications which influence the shape of the non-proportional effect, such as the choice of the number and location of knots or a smoothing parameter which are often data driven |
| Reduced Rank Regression models | • Reduces the number of parameters in order to obtain more stable and parsimonious models depending on the rank of the model | • Does not allow for selection of covariates and non-linear effects<br>• All effects are modelled as time-varying based on the same time functions |
| Neural Networks | • Provides smoothed hazard function estimation and allows for non-linear covariate effects | • Requires discrete survival times<br>• Complex computer-intensive process<br>• Not straightforward |

#### *4.4.4.1 Non-Parametric Estimation of Time-Varying Coefficients.*

When $\beta(t)$ in the extended Cox model is assumed to be non-parametric, there exist several methods for estimation.

Murphy & Sen proposed a histogram sieve estimator procedure by assuming that the coefficient functions, $\beta(t)$, are piecewise constant (Murphy and Sen, 1991). Verweij & Houwelingen proposed estimating the coefficients at each event time using a penalised partial likelihood approach and smoothing parameters based on the Akaike's Information Criteria (AIC)(Verweij and van Houwelingen, 1995). However, the assumption that the baseline hazard function and coefficient functions are piecewise constant may not be appropriate in real applications.

The coefficient functions, $\beta(t)$, can be estimated smoothly: Zucker & Karr considered nonparametric penalised partial likelihood of time-dependent effects using spline fitting (see section 4.4.4.5 for an explanation of splines), leaving $\lambda_0(t)$ unspecified (Zucker and Karr, 1990). This was further developed by Hastie & Tibshirani in 1993 who used a smoothing spline partial likelihood method to estimate $\beta(t)$ based on natural cubic splines with knots at unique failure times (Hastie and Tibshirani, 1993).

More recently, Martinussen and colleagues developed an efficient estimation procedure for the cumulative parameter functions of the fully time-varying extended Cox model (Martinussen et al., 2002). They were the first to study in detail the semi-parametric version in which the influence of some of the covariates varies with time while the effects of the remaining covariates are constant.

Cai & Sun proposed estimating the time-dependent coefficient functions using a local partial likelihood technique and Tian and colleagues followed on from this proposing a simple estimation procedure based on a kernel-weighted partial likelihood approach (Cai and Sun, 2003, Tian et al., 2005).

#### 4.4.4.2 *Piecewise Constant Effects*

A simple adaptation of the Cox model to handle non-proportional effects is by partitioning of the time axis, called piecewise constant effects, resulting in a step function for $\beta(t)$. In 1982, Anderson & Senthilselvan proposed using a two-step regression model for hazard functions (Anderson and Senthilselvan, 1982). It is assumed that $\beta_i(t) = \alpha_i$ for $t \leq \boldsymbol{B}$ and $\beta_i(t) = \gamma_i$ for $t > \boldsymbol{B}$ , $i = 1, \dots, p$ , for some **B**, which is not assumed known a *priori*. The corresponding hazard function is

$$\lambda(t; x) = \begin{Bmatrix} \lambda_0(t)\exp(\alpha^T x), & t \leq \boldsymbol{B} \\ \lambda_1(t)\exp(\gamma^T x), & t > \boldsymbol{B} \end{Bmatrix} . \qquad 4.8.$$

Moreau and colleagues generalised the approach by Anderson & Senthilselvin to a larger number of intervals (Moreau et al., 1985). The assumption that the baseline hazard function and coefficient functions are piecewise constant may not be appropriate in real situations. However, it is a simple approach in terms of statistical application and interpretation and is widely used in practice.

#### 4.4.4.3 *Simple Time-By-Covariate Interactions*

Simple interactions with the covariate *x* and a pre-defined function of time *f(t)* was first introduced by Cox in 1972 which allowed smooth estimation of $\beta(t)$ (Cox, 1972):

$$\lambda(t; x) = \lambda_0(t) exp\left(\sum_{i=1}^{p} \beta_i x_i f_i(t)\right) . \qquad 4.9.$$

Monotonic functions for *f(t)* were initially used such as linear, logarithmic or rank functions of time. The main disadvantage is the simple options are restrictive on the available shape of the non-proportional effect.

#### 4.4.4.4 *Polynomials and Fractional Polynomials*

The simple monotonic functions of *t* can be extended to polynomials, e.g. $(t) = \gamma_1 t + \gamma_2 t^2 + \cdots$ , used by Stablein and colleagues (Stablein et al., 1981). Fractional polynomials (FP) allow a more flexible family of shapes to model the non-proportional effects compared with

polynomials as they allow negative and non-integer powers (Royston and Altman, 1994). An FP1 transformation of an argument $x > 0$ with power $q$ is defined as $x^q$, where $q$ belongs to the standard set of powers $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ with $x^0 = \log(x)$ rather than usual notation $x^0 = 1$. An FP2 transformation of $x$ with powers $\boldsymbol{q} = (q_1, q_2)$ is the vector $x^{\boldsymbol{q}}$ with

$$x^{\boldsymbol{q}} = x^{(q_1,q_2)} = \begin{cases} (x^{q_1}, x^{q_2}), & q_1 \neq q_2 \\ x^q, x^q \log(x), & q_1 = q_2 = q \end{cases} . \qquad 4.10.$$

This gives the time-varying Cox model based on untransformed X and an FP2 function of t with powers $\boldsymbol{q} = (q_1, q_2)$ :

$$\lambda(t; X) = \lambda_o(t) \exp[X\varphi_2(t; \boldsymbol{q})]$$
$$= \begin{cases} \lambda_o(t) \exp[X(\beta_0 + \beta_1 t^{q_1} + \beta_2 t^{q_2})], & q_1 \neq q_2 \\ \lambda_o(t) \exp[X(\beta_0 + \beta_1 t^q + \beta_2 t^q \log(t))], & q_1 = q_2 = q \end{cases} . \qquad 4.11.$$

The use of fractional polynomials in the time-varying Cox model have been considered by Berger and colleagues in 2003 and Sauerbrei and colleagues in 2007 (Berger et al., 2003, Sauerbrei et al., 2007). Berger introduced the time-dependent covariate model (or dynamic Cox model) as

$$\lambda(t; x) = \lambda_0(t) exp\left(\sum_{i=1}^{q} \beta_i(t) x_i\right), \qquad 4.12.$$

with $\beta_i(t) = \gamma_{i0} + \sum_{j=1}^{M} \gamma_{ij} t^{(q_j)}$ being an FP of maximum degree M=2 and $\gamma_{ij}$ are the regression coefficients. This approach allows detecting and simultaneously modelling dynamic effect structures as the PH model is nested within the dynamic Cox model and can therefore be tested using a likelihood ratio test.

It has been shown that a covariate may appear to have a time-dependent effect due to miss modelling another part of the model, for example the wrong functional form of a covariate or omission of an important covariate. Sauerbrei and colleagues introduced a method for

multivariable modelling of time-varying effects called a multivariable fractional polynomial-time (MFPT) approach which can address both the violation of the linearity and PH assumptions. It is a multivariable strategy to select variables which have an effect on the outcome, determine the functional form of continuous variables and determine whether the PH assumption is sensible or if time-varying effects are present. As described by Sauerbrei et al., it involves three stages:

- Stage 1. Determine time-fixed model $M_0$
    - Select model $M_0$ using the multivariable fractional polynomial (MFP) algorithm (Sauerbrei and Royston, 1999) assuming PH (full time-period).
- Stage 2. If necessary, add covariate with short-term effect
    - Start with model $M_0$.
    - Restrict time period to $(0, \tilde{t})$, e.g. $\tilde{t}$ defined by first half of events.
    - Run the MFP algorithm for $(0, \tilde{t})$ and add, if necessary, significant covariates to $M_0$. This gives model $M_1$.
- Stage 3. Investigate possible time-varying effects of variables in $M_1$
    - For each covariate (with selected FP transformation) of $M_1$ run the FPT algorithm to investigate time-varying effect adjusting for all covariates of $M_1$.
    - Use a forward selection procedure to add significant time varying effects to model $M_1$. This gives the final model $M_2$.

The MFP algorithm combines the determination of FP functions with backward elimination. The maximum complexity (degree) of FP for each continuous predictor must be specified with Sauerbrei et al. suggesting the use of degree 2. Initially, the model includes a linear term for each of the continuous variables and all the binary and categorical variables. The 'best' fitting FP for each of the continuous variables is found based on deviance differences, while adjusting for all continuous, binary and categorical predictors already in the model. Variables that are

not significant predictors of outcome are dropped from the model using backward elimination and a pre-defined significance level. For the continuous covariates the best fit second-degree FP (FP2) is tested at the α-level against the best fit first-degree FP (FP1). If this is non-significant, the FP1 is tested against a straight line. If the linear term is not significant the variable is eliminated. Convergence occurs when no further changes to selected variables and their FP transformations take place.

The FPT approach in stage 3 is based on a model of the type

$$\lambda(t;x) = \lambda_0(t)exp\left(\sum_{i=1}^{p}\beta_i(t)f_i(x_i)\right) , \qquad 4.13.$$

allowing non-linear functional forms of the covariates $f_i(x_i)$ and time-varying effects $\beta_i(t)$. The FPT algorithm mimics the MFP algorithm for selecting an FP function of a continuous covariate. The best fitting FP1 and FP2 functions to model the time-varying effects are determined based on the deviances of the models. The best fitting FP2 is compared with a constant effect. If non-significant, the PH assumption is accepted otherwise model reduction is performed to find the most parsimonious representation of the non-proportional effect. The best FP2 is compared with the default (logarithmic) function, a FP1 based on log(t). If non-significant, the logarithmic function is accepted. If significant, the best fitting FP2 is compared to the best fitting FP1. If significant the more complex FP2 is chosen, otherwise the simpler FP1 function is selected.

#### *4.4.4.5 Splines*

Another option to model time-by-covariate interactions smoothly is through the use of splines. Splines are flexible mathematical functions defined by piecewise polynomials. The points at which the polynomials join are called knots and certain constraints are imposed to ensure the function is smooth. The most common splines used in practice are cubic splines, in particular cubic B(asis)-splines, where the fitted function is forced to have continuous $0^{th}$,$1^{st}$ and $2^{nd}$

derivatives. B-splines can be evaluated in a numerically stable way by the de Boor algorithm (de Boor, 1978).

Defining K knots, $t_1, \dots, t_K$ the cubic spline function is

$$S(x) = \sum_{j=0}^{3} \beta_{oj} x^j + \sum_{i=1}^{K} \beta_{i3} (x^j - t_i)_+^3 \quad , \qquad 4.14.$$

where the "+" notation means that $u_+ = \begin{cases} u & if\ u > 0 \\ 0 & if\ u \leq 0 \end{cases}$ resulting in K+4 parameters needed in the linear predictor. Cubic splines can sometimes behave poorly in the tails where there is often a small amount of data. They can be extended to *restricted* cubic splines which are forced to be linear before the first knot and after the final knot.

The choice of the number and location of knots is often arbitrary; a better option is to use smoothing splines where knot selection is based on a mean squared criterion. The level of smoothness is chosen based on selecting the degrees of freedom for each spline fit and can be controlled using the Akaike's Information Criteria (AIC)(Akaike, 1974).

The choice of the number of knots has been subject to a large body of published research, with too few or too many knots resulting in under- or over-fitting of the data. One option is to use penalised splines (P-splines); splines with a penalty for the model complexity which can be seen as a compromise between smoothing and regression splines. The number of knots defining the spline function is larger than that justified by the data but the level of over-fitting is controlled by a roughness penalty over the curve. The most common choice is a penalty based on the integral of a squared derivative of the spline curve (Costa and Shaw, 2009).

Time-dependent effects are modelled simply by introducing an interaction between the covariate and the spline variables:

$$\lambda(t;x) = \lambda_0(t)\exp\left(\sum_j f_j(t)x_{ij}\right), \qquad 4.15.$$

where the functions $f_j(t) = \sum_k \theta_{jk}\beta_{jk}(t)$ are modelled with spline basis functions.

These various spline methods were used to model time-dependent effects by numerous authors (Gray, 1992, Hess, 1994, Abrahamowicz et al., 1996, Brown et al., 2007).

Similar to the approach by Sauerbrei and colleagues, except with the use of splines rather than fractional polynomials, Abrahamowicz & Mackenzie suggested a generalisation of the Cox model with relaxations of the assumptions of log-linearity and proportional hazards, called the product model (Abrahamowicz and MacKenzie, 2007):

$$\lambda(t;x) = \lambda_0(t)exp\sum_{m=1}^{M}\beta_m(t)r_m(x_m)\ , \qquad 4.16.$$

where *m* indexes the predictors $x_1, \ldots, x_M$. It is based on the product of two flexible estimates, the dose-response function $r_m(x_m)$ used to relax the log-linearity assumption and the time-dependent function $\beta_m(t)$ to relax the PH assumption, using a quadratic B-spline basis. The product model is a flexible regression spline-based model which allows simultaneously relaxing assumptions of PH and linearity regarding the impact of the continuous predictor on the log-hazard and helps avoid finding spurious non-proportional effects by miss modelling other parts of the data. When using flexible methods it is often difficult not to over-fit the model to the data and a trade-off between model parsimony and goodness-of-fit must be determined. To overcome this they suggest using model selection criteria based on AIC and the model can be reduced to the conventional Cox PH model if appropriate. Further work is needed to develop a general model selection strategy to identify the correct model for each of several, possibly correlated, predictors.

#### 4.4.4.6 *Reduced Rank Regression Models*

The problem of the choice of time functions being data driven was addressed by Perperoglou and colleagues who introduced a more general technique for fitting time-by-covariate interactions with the use of reduced rank regression models (Perperoglou et al., 2006a). A model with time-varying effects of the covariates can be described by the structure matrix Θ that contains the coefficients of the covariates and their interactions with the time functions. The structure matrix can be factorised in different ways, resulting in a matrix of lower rank *r* (i.e. the number of parameters to be estimated). Consider **B** a $p \times r$ matrix and **Γ** a $q \times r$ matrix of coefficients, that factorise the Θ matrix as $\Theta = B\Gamma^T$ . The rank *r* model is written as

$$\lambda(t; \boldsymbol{X}) = \lambda_0(t) \exp\left(\boldsymbol{X}\boldsymbol{B}\boldsymbol{\Gamma}^T\boldsymbol{F}^T(t)\right) \qquad 4.17.$$

$$= \lambda_0(t) exp\{\textstyle\sum_{k=1}^{r}(\boldsymbol{X}\boldsymbol{\beta}_k)(\boldsymbol{F}(t)\boldsymbol{\gamma}_k)\} \ ,$$

where $\boldsymbol{\beta}_k$ is the k[th] column of **B** and $\boldsymbol{\gamma}_k$ is the k[th] column of **Γ.** Thus, the parameters consist of a set of *r* linear combinations of time functions $\boldsymbol{F}(t)\boldsymbol{\gamma}_k$ and *r* linear combinations of covariates $\boldsymbol{\beta}_k$ . For flexible time-varying effects a full rank is most appropriate (equivalent to the extended Cox model), while for more parsimonious modelling and more rigid time effects a smaller rank should be preferred using AIC to determine the optimal rank. The main disadvantage of this approach is that all effects are modelled as time-varying.

#### 4.4.4.7 *Other Approaches*

Another approach suggested by Biganzoli and colleagues is the use of an artificial neural network (ANN) for the estimation of the functional relationships between covariates and time in survival data (Biganzoli et al., 1998). It provides smoothed hazard function estimation and allows for non-linear covariate effects however this method requires discrete survival times, it is a complex computer-intensive process and is not a straightforward model.

#### *4.4.4.8 Bayesian Inference*

Bayesian inference for approaches to analyse non-proportional effects in survival data have also been investigated, which makes inferences on the posterior distribution and allows making direct probability statements.

In 1991, Gammerman proposed a Bayesian dynamic survival model (BDSM) using a sequential analysis based on factorisation of the likelihood over the time intervals which the parameters are allowed to vary between (assuming a piecewise constant baseline hazard function) (Gamerman, 1991). In 1997, Sargent proposed a Bayesian approach defining a hierarchical Cox model with a state-space structure using Markov Chain Monte Carlo (MCMC) methods to estimate the time-varying coefficients (Sargent, 1997). Sargent's method is similar to the Cox model in that no assumption is made about the baseline hazard function. Gustafson proposed a hierarchical Bayes model which relaxes both assumptions of proportionality and the additive effects assumption of the Cox model in 1998; where the additivity assumption specifies that the effect associated with a particular explanatory variable does not depend on the levels of the other explanatory variables (Gustafson, 1998). Time-dependent effects are explicitly modelled and the additivity of the effects is relaxed through the use of a modified neural network structure.

Mckeague & Tighiouart proposed a nonparametric Bayesian approach to the extended Cox model where the baseline hazard function $\lambda_0(t)$ and the covariate effect $\beta(t)$ are specified as piecewise constant effects, where the number and position of jump times are taken as random, and modelled as independent stochastic processes with sample paths taking the form of step functions (McKeague and Tighiouart, 2000). A disadvantage to this method is that the model assumes the same time scale and jump times for both the baseline hazard function and the covariate effect. Haneuse and colleagues extended this approach to allow for separate timescales for the baseline hazard and the time-varying effect of a time-dependent covariate

(Haneuse et al., 2008). They use a hierarchical model for the hazard function with separate timescales being incorporated via conditionally independent stochastic processes.

New developments include geoadditive survival models where the common linear predictor in the Cox model is generalised to an additive predictor, including non-parametric components for the log-baseline hazard, time-varying effects and possibly non-linear effects of continuous covariates. Hennerfeind and colleagues proposed flexible continuous-time geoadditive models with inference developed within a fully Bayesian framework using penalised regression splines and Kneib & Fahrmeir develop inference based on an empirical Bayes approach (Hennerfeind et al., 2006, Kneib and Fahrmeir, 2007).

Recently Costa & Shaw estimated time-varying regression coefficients from the extended Cox model through penalised cubic spline functions using the partial likelihood and considered both empirical and full Bayesian inference methods (Costa and Shaw, 2009). He and colleagues further explored the BDSM introduced by Gamerman in 1991 and suggested the linear Bayesian (LB) estimation method; where the coefficients are modelled using first-order random walks only, meaning all coefficient functions are piecewise constant (He et al., 2010). The main difficulty with this method is that smoothing parameters must be pre-specified. Kim and colleagues proposed a Bayesian approach for estimating hazard functions under the constraint of a monotone hazard ratio using Cox's model with a monotone time-dependent coefficient (Kim et al., 2011). They used a signed gamma process prior for the time-dependent coefficient and the Bayesian bootstrap prior for the baseline hazard function and developed an efficient MCMC algorithm for estimation.

## 4.5 Additive Hazards Models (Table 4.5)

It is possible the covariates add to, rather than multiply, the baseline hazard resulting in additive hazards models. The additive hazards models found in the literature search are shown in Table 4.5 with their respective advantages and disadvantages given in Table 4.6.

The Aalen additive model introduced by Aalen in 1980 is less well known than the Cox model but it is useful when the interest lies in the risk difference rather than the relative risk and when modelling time-dependent effects (Aalen, 1980). It is a fully additive and non-parametric model whose regression coefficients are allowed to vary over time. The model is

$$\lambda(t; X) = X^T \beta(t)\,, \qquad 4.18.$$

where X is a $p$-dimensional covariate and $\beta(t)$ is a $p$-dimensional regression coefficient. The additive risk model specifies that the hazard function associated with a set of possibly time-varying covariates is the sum of, rather than the product of, the baseline hazard function and the regression function of the covariates (Lin and Ying, 1994). The model is non-parametric in the sense that no functional form is assumed for the baseline hazard or the regression functions. It was further studied by Aalen in 1989 and 1993 (Aalen, 1989, Aalen, 1993) and estimation properties of the model were given by McKeague and Huffer & McKeague (McKeague, 1988, Huffer and McKeague, 1991). In the Aalen additive hazard model all covariates are incorporated as time-varying and a more parsimonious approach is available through the partly-parametric version of Aalen's additive model.

Mckeague & Sasieni introduced the partly-parametric version of Aalen's model where some of the covariates are allowed to vary non-parametrically with time where as others remain constant (McKeague and Sasieni, 1994):

**Table 4.5** Additive Hazard Models and Multiplicative-Additive Hazard Models

| Model Name | Formula | Author (Year) | Citation Searching* |
|---|---|---|---|
| **Additive Hazard Models**<br>Aalen Additive Model | $\lambda(t;X) = X^T\beta(t)$ | Aalen (1980, 1989,1993)<br>Mckeague (1988)<br>Huffer & McKeague (1991) | NA,290,104<br>26<br>107 |
| Partly-Parametric Additive Risk Model | $\lambda(t;X,Z) = X^T\beta(t) + Z^T\gamma$ | Mckeague & Sasieni (1994)<br>Lin & Ying (1994)<br>Dunson & Herring (2005) | 136<br>256<br>8 |
| **Additive-Multiplicative Hazard Models**<br>Excess Hazard Models | $\lambda(t;X,Z) = X^T\alpha(t) + \lambda_0(t)\exp\{Z^T\beta\}$ | Martinussen & Scheike (2002) | 50 |
| Cox-Aalen Model | $\lambda(t;X,Z) = [X^T\alpha(t)]\exp(Z^T\beta)$ | Scheike & Zhang (2002) | 44 |

* Refer to section 4.9

**Table 4.6** Advantages\Disadvantages of Additive Hazards Models and Additive-Multiplicative Hazards Models

| Model Name | Advantages | Disadvantages |
|---|---|---|
| **Additive Hazard Models**<br>Aalen Additive Model | • Useful when interest lies in risk difference rather than relative risk<br>• Flexible model with easy interpretation of excess risk of the covariates<br>• Easier estimation of survival function<br>• Does not require use of smoothing parameters | • Does not provide parameters or formulae that are easily reported<br>• Non-parametric nature of the model results in<br>- large standard errors<br>- possibility of estimating partially negative hazard rates |
| Partly-Parametric Additive Risk Model | • Compromise between bias and variance compared to fully non-parametric model<br>• Semi-parametric method, leaving the baseline unspecified, resulted in an alternative to Cox PH model | • More complicated inference<br>• Possibility of estimating partially negative hazard rates |
| **Additive-Multiplicative Hazard Models**<br>Excess Hazard Models | • Useful when the baseline mortality expressed in the baseline hazard is unknown | • More complex interpretation of the model |
| Cox-Aalen Model | • Flexible class of models<br>• Easy estimation and interpretation of covariate effects on the baseline hazard<br>• Does not require smoothing parameters | • Interpretation of baseline hazard more complicated |

$$\lambda(t; X, Z) = X^T \beta(t) + Z^T \gamma \,, \qquad 4.19.$$

where $\beta(t)$ is a $p$-dimensional time-varying regression coefficient and $\gamma$ is a $q$-dimensional time-invariant coefficient. Therefore, allowing some covariates to be time-varying and others to remain constant, giving a good compromise between variance and bias compared to the fully non-parametric model (Cortese et al., 2010). Lin & Ying developed a simple semi-parametric method to make inference about the regression parameters with an unspecified baseline hazard function similar to the partial-likelihood-based methods used for the Cox proportional hazards model (Lin and Ying, 1994).

Dunson & Herring developed an approach for Bayesian inference in the additive hazards model and overcame the problem of negative hazards being produced (Dunson and Herring, 2005). They also proposed an approach that facilitates selection and formally accommodates uncertainty between proportional and additive hazards.

## 4.6 Additive-Multiplicative Hazards Models (Table 4.5)

The additive and multiplicative models discussed previously can be used simultaneously to complement each other to model covariates that lead to relative risk and covariates that need additional flexibility. The additive Aalen model and the multiplicative Cox model can be added together to give proportional excess hazards models or multiplied together to give the Cox-Aalen model with their respective advantages and disadvantages given in Table 4.6. Martinussen & Scheike studied a general class of additive-multiplicative hazard models (or excess hazards models) assumed to have the form

$$\lambda(t; X, Z) = X^T \alpha(t) + \lambda_0(t) \exp\{Z^T \beta\} \,, \qquad 4.20.$$

where $\alpha(t)$ is a q-vector of time-varying regression functions, $\lambda_0(t)$ is the baseline hazard of the excess risk term, and $\beta$ is a p-dimensional vector of relative-risk coefficients (Martinussen and Scheike, 2002). The background mortality is unknown and modelled by Aalen's additive model. The excess risk due to the covariates is described by the Cox PH model.

Scheike & Zhang introduced the Cox-Aalen model given by (Scheike and Zhang, 2002):

$$\lambda(t; X, Z) = [X^T \alpha(t)] \exp(Z^T \beta) \ . \qquad 4.21.$$

Thus, allowing some covariate effects to be non-parametric and time-varying and other effects to be included as time constant. It extends the Cox and Aalen models, providing a very flexible class of models with easy estimation and interpretation of covariate effects on the baseline hazard.

## 4.7 Parametric Models (Table 4.7)

The use of the Cox model and other semi-parametric methods involves ignoring the baseline hazard, treating it as a high-dimensional nuisance parameter. It has been suggested that not modelling the baseline hazard results in missing important information directly related to the time-course of an illness (Royston and Parmar, 2002a). Therefore parametric approaches, which model the baseline hazard, may provide additional information with advantages and disadvantages of the various parametric approaches given in Table 4.8.

Mackenzie suggested the use of a parametric family of continuous survival distributions for the analysis of non-proportional hazards: the generalised time-dependent logistic model (Mackenzie, 1996). The baseline hazard takes the functional form

$$\lambda_0(t|\lambda, \alpha, \gamma) = \frac{\lambda \exp(t\alpha + \gamma)}{1 + \exp(t\alpha + \gamma)} \ , \qquad 4.22.$$

where $\lambda > 0$ and $\alpha$ and $\gamma$ are scalars. This model differs from its standard parametric competitors, such as the exponential and Weibull models, as it allows the time-dependence to follow a more flexible, sigmoid pattern and allows the effects of the covariates to diminish with time in an intuitively appealing way.

**Table 4.7** Parametric Models and Others

| Model Name | Formula | Author (Year) | Citation Searching* |
|---|---|---|---|
| **Parametric Models** | | | |
| Generalised Time-Dependent Logistic Family | $\lambda_0(t\|\lambda,\alpha,\gamma) = \frac{\lambda\exp(t\alpha+\gamma)}{1+\exp(t\alpha+\gamma)}$ | Mackenzie (1996) | 19 |
| Flexible Parametric Models | $\ln\{H(t\|x_i)\} = s\{\ln(t) \mid \gamma, k_0\} + \sum_{j=1}^{D} s\{\ln(t) \mid \gamma, \delta_k k_j\} x_{ij} + x_i\beta$ | Royston & Parmar (2002) | 115 |
| **Other Models** | | | |
| Cure Models | $S(t\|X,X^*) = [1-\pi(X^*)]S_u(t\|X) + \pi(X^*)$ | Farewell (1982)<br>Peng et al. (1998)<br>Peng (2003) | 355<br>94<br>32 |
| Frailty Models | $\lambda(t\|X) = \frac{\lambda_0(t)\exp(X\beta)}{1+F(t\|\delta)\exp(X\beta)}$ | Perperoglou et al. (2006b) | 5 |
| Tree-Based Approach | N/A | Xu & Adak (2002) | 28 |
| Bayesian Approaches | Simple Approach<br>Extension to Penalised Poisson Regression<br>Generalised Odds-Rate Hazards (GORH) models<br>Dependent Dirichlet Process model<br>Bayesian Random-Effects Threshold Regression | Gustafson et al. (2003)<br>Lambert & Eilers (2005)<br>Banerjee et al. (2007)<br>De Iorio et al. (2009)<br>Pennell et al. (2010) | 16<br>20<br>5<br>19<br>2 |

* Refer to section 4.9

**Table 4.8** Advantages/Disadvantages associated with Parametric and Other Models

| Model Name | Advantages | Disadvantages |
|---|---|---|
| **Parametric Models**<br><br>Generalised Time-Dependent Logistic Family | • Allows the time-dependence to follow a more flexible, sigmoid pattern<br>• Allows the effects of the covariates to diminish with time in an intuitively appealing way<br>• Relaxes the assumption of PH while preserving the linear influence of the regression parameter on the log-odds scale | • Does not allow modelling of more complex time-dependent effects |
| Flexible Parametric Models | • Ease of which non-proportional effects can be fitted<br>• Parametric model allows easy prediction of hazard and survival rates | • Difficulties associates with choosing the number and location of knots for the baseline cumulative hazards and the non-proportional effects |
| **Other Models**<br>Cure Models | • Able to explain the unexpectedly large survival times<br>• Potential to identify outliers | • In breast cancer studies the patients are often never actually considered cured and so this must be carefully considered |
| Frailty Models | • Does not follow the assumption of proportionality by assuming each person has their own frailty | • Leads to a complex model where covariate effects are hard to interpret |
| Tree-Based Approach | • Time-varying regression effects estimated in simplistic way<br>• Gives clear indication of important variables<br>• Provides exploratory tool for finding multiple change points | • Different split points for different covariates is possible but relies on subjective judgment<br>• Difficulty in interpreting covariate effects<br>• Highly data-dependent |

In 2002, Royston & Parmar developed flexible parametric models based initially on the assumption of proportional hazards or proportional odds scaling of covariate effects (Royston and Parmar, 2002a). This was then extended to include time-varying effects modelled using spline functions. The approach involves modelling on the log *cumulative* hazard scale (rather than the log hazard) which under the PH assumption covariate effects can still be interpreted as hazard ratios and allows easy transformation of the survival and hazard functions. The logarithm of the baseline cumulative hazard function is modelled as a natural (restricted) cubic spline function of log time, with the general function $s(x;\gamma)$ to be approximated by a spline. Natural cubic splines are defined as cubic splines constrained to be linear beyond boundary knots $k_{min}$, $k_{max}$. In addition, *m* distinct internal knots are specified (see section 4.4.4.5 for more detail)

A natural cubic spline function of ln(t) with knots $K_0$ may be written as $s\{\ln(t) \mid \gamma, k_0\}$ which is then used to model the baseline log cumulative hazard in a proportional (cumulative) hazards model:

$$\ln\{H(t|x_i)\} = s\{\ln(t) \mid \gamma, k_0\} + x_i\beta, \qquad 4.23.$$

Non-proportional effects are modelled simply by introducing an interaction between the covariate and the spline variables. If there are D time-dependent effects, this can be written as

$$\ln\{H(t|x_i)\} = s\{\ln(t) \mid \gamma, k_0\} + \sum_{j=1}^{D} s\{\ln(t) \mid \gamma, \delta_k k_j\} x_{ij} + x_i\beta \,. \qquad 4.24.$$

## 4.8 Other Models (Table 4.7)

Several modelling techniques have been proposed so far for the problem of non-proportional hazards. Other approaches are also available including cure and frailty models, tree based

methods and Bayesian approaches with their respective advantages and disadvantages given in Table 4.8.

### 4.8.1 Cure and Frailty Models

One such approach to model long-term survival studies is with the use of mixture models, known as cure models (Perperoglou et al., 2007). In mixture models the overall survival of the patients consists of two parts, a survival function $S_u(t|X)$ that models the survival of not-cured patients, denoted by the subscript u, and a probability a patient has been cured $\pi(X^*)$ which depends on some covariates X* and takes the logistic form $\log[\frac{\pi(X^*)}{\{1-\pi(X^*)\}}]$ .This gives the overall survival function at time *t* for patients with covariates X and X* as

$$S(t|X, X^*) = [1 - \pi(X^*)]S_u(t|X) + \pi(X^*) \ . \qquad 4.25.$$

The survival of uncured patients has been described by parametric forms (Farewell, 1982, Peng et al., 1998) and the proportional hazards model (Peng, 2003).

In some cases it may be possible to use a frailty model to adjust for possible lack of fit and departure from proportionality (Perperoglou et al., 2007). Frailty models are used in survival analysis to describe individual heterogeneity. Since individual heterogeneity is not observable a random effect Z, which is often assumed to follow a gamma distribution (resulting in the Burr model), is entered into the model:

$$\lambda(t|X, Z) = Z\lambda_0(t)\exp(X\beta) \ . \qquad 4.26.$$

The covariate effects are now assumed to disappear over time, i.e. the hazard ratio converges to 1. This model also has the assumption that the frailties are assumed constant. However, with long follow up this assumption may not hold and Perperoglou and colleagues introduced the relaxed Burr model as an approximation to a frailty model with time-varying effects (Perperoglou et al., 2006b):

$$\lambda(t|X) = \frac{\lambda_0(t)\exp(X\beta)}{1 + F(t|\delta)\exp(X\beta)} \quad , \qquad 4.27.$$

with $(t|\delta) = f(t)\delta$ , where $f(t)$ can be a simple time function multiplied by an unknown but estimable coefficient $\delta$.

In terms of cure and frailty models, the interpretation of each model must be distinguished as they are designed to answer different questions. Although they are useful alternatives to the Cox proportional hazards model, the user must determine whether they answer the research question.

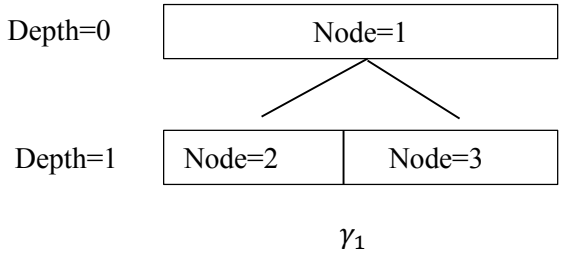
### 4.8.2 Tree-Based Method

Xu & Adak developed a tree-based method to handle survival data for the assessment and estimation of time-dependent regression effects under a Cox-type model (Xu and Adak, 2002). The time-varying regression effects are estimated in a simplistic way, more suitable for clinical applications, using the step-function approach. An algorithm is used that maximises score statistics for recursive segmentation of the time axis. A pruning algorithm is then applied to determine a sparse segmentation with bootstrap resampling to correct for over optimism due to split point optimism. The algorithm for growing the binary tree is as follows:

(1) Start with the root node, i.e. the entire data and fit the model:

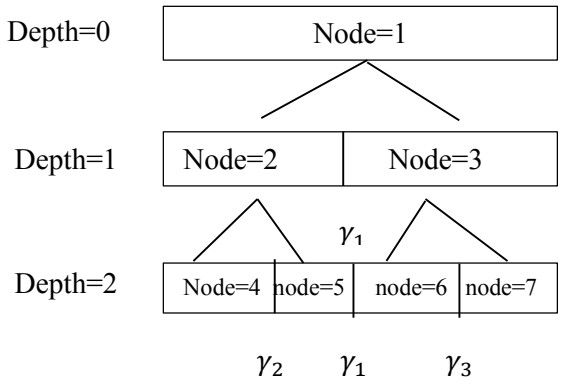
$$M_0: \beta(t) = \beta \; for \; all \; t.$$

Determine the point, $\gamma_1$, at which the score test is maximised. Define the children of the node as the segments resulting from adding the change point to the model and fit model $M_1$.



$$M_1: \beta(t) = \begin{cases} \beta_1, & for\ t < \gamma_1 \\ \beta_2, & for\ t \geq \gamma_1 \end{cases}$$

(2) Determine the optimal change points, $\gamma_2$ , by maximising the score test for node 2, and $\gamma_3$ , for node 3. Define the children of the node as the segmants resulting from adding the change points to the model and fit model $M_2$.




$$\text{Al } M_1: \beta(t) = \begin{cases} \beta_1, & for\ 0 \leq t < \gamma_1 \\ \beta_2, & for\ \gamma_1 \leq t < \gamma_2 \\ \beta_3, & for\ \gamma_2 \leq t < \gamma_3 \\ \beta_4, & for\ \quad t \geq \gamma_3 \end{cases}$$

(3) The recursive segmentation procedure can be continued. In general, determine the optimal change points at depth *d* by fitting model $M_d$. Then obtain model $M_{d+1}$ by determining the optimal change points at depth *d*. The stopping rule for the splitting of the tree can be determined by specifying (1) a maximum allowable depth, (2) a minimum number of events within each node or (3) a threshold for the maximised score statistic for each node.

(4) The growing algorithm results in over-segmentation of the data and a pruning algorithm from classification and regression trees (CART) can be used to merge the intervals that do not provide a sufficient increase in the log partial likelihood.

### 4.8.3 Bayesian Approaches

Gustafson and colleagues proposed a simple approach to fitting Bayesian survival models with the use of the normal approximation to the log-gamma distribution which yields easy and efficient computational methods for estimating the baseline log-hazards and time-dependent effects (Gustafson et al., 2003).

Lambert & Eilers proposed an extension to the penalised Poisson regression approach where the number of events in each narrow time bin is described using a Poisson distribution with the log mean specified using a flexible penalised B-splines model with a large number of equidistant knots (Lambert and Eilers, 2005). They extend this approach to allow the regression coefficients to vary smoothly over time. The regression parameters and penalty weights are estimated efficiently using Bayesian inference tools.

Banerjee and colleagues consider a Bayesian approach to a class of generalised odds-rate hazards (GORH) models (Banerjee et al., 2007). This class of models is governed by a non-proportionality parameter and is general enough to include proportional hazards model and the proportional odds model. The GORH models are given as

$$S(t;x) = \{1 + \theta exp(\varphi_0(t) + x^T\beta)\}^{-\theta^{-1}} \ , \qquad 4.28.$$

thus the hazard rate is given by

$$\lambda_\theta(t;x) = \varphi_0'(t)\exp(\varphi_0(t) + x^T\beta)\{1 + \theta\exp(\varphi_0(t) + x^T\beta)\}^{-1} \ , \qquad 4.29.$$

where $\theta > 0$ is the GORH parameter while $\varphi_0(t)$ controls the form of the baseline hazard function.

De Iorio and colleagues developed a Bayesian nonparametric approach through a dependent Dirichlet process model with inferences easily obtained using MCMC (De Iorio et al., 2009). This approach does not require the resulting survival curve to satisfy the proportional hazards

assumption. The Dirichlet process (DP) is a tool frequently used in nonparametric Bayesian inference and is a probability model for random probability distributions.

An alternative approach, which again does not require the assumption of proportional hazards, is to use a first hitting time model which treats a subject's health status as a latent stochastic process that fails when it reaches a threshold value (Pennell et al., 2010). Pennell and colleagues propose a random-effects threshold regression model and Bayesian methodology for inference (Pennell et al., 2010).

The motivation of this literature search is to determine what other approaches are available when the assumption of proportional hazards in the Cox model is not valid. Due to the time constraints of this project and the extra conceptual leap to Bayesian inference I have decided the various Bayesian approaches will not be considered further. I will focus on promoting the use of alternative approaches to the Cox PH model without the added complexity of introducing Bayesian concepts.

## 4.9 Application of novel methods to data analysis

An interesting part of this literature review was to determine which of the approaches available for investigating non-proportional effects for survival data have been used in practice. This was investigated through forward citation searching in Google Scholar and the number of citations for each of the models can been seen in Table 4.1 through Table 4.7.

The extended Cox model first introduced in Cox's 1972 paper was by far the most cited paper with over 30,000 citations (Cox, 1972). However this paper also introduces the popular Cox proportional hazards model which explains the high number of citations. Popular approaches such as the extended Cox model and the use of piecewise constant effects through splitting of the time scale are often not cited so the focus of this forward citations search was on approaches with potential interest and developed after the year 2000.

Based on the advantages and disadvantages discussed previously, the approaches that were of interest were:

Multiplicative Models

- The MFPT approach proposed by Sauerbrei and colleagues (Sauerbrei et al., 2007)
- The product model introduced by Abrahamowicz and colleagues (Abrahamowicz and MacKenzie, 2007)

Additive-Multiplicative Hazards Models

- The Cox-Aalen model developed by Scheike & Zhang (Scheike and Zhang, 2002)

Parametric Models

- Flexible parametric models introduced by Royston & Parmer (Royston and Parmar, 2002a).

### 4.9.1 Fractional Polynomials

The MFPT approach proposed by Saurerbrei and colleagues involved the use of fractional polynomials in the extended Cox model for selection of non-linear and non-proportional effects and was cited by 23 papers. In their original paper they applied the MFPT approach to 2982 patients from the Rotterdam breast cancer series with follow up time ranging from 1 to 231 months (median 107 months), the outcome was event-free survival and they observed 1,518 events. They identified several variables to have time-dependent effects: tumour size, number of positive lymph nodes, hormonal therapy, chemotherapy and progesterone having the strongest time-varying effect. The forward citations of the MFPT approach by Saurerbrei and colleagues did not highlight any further applications of the approach being used in practice.

### 4.9.2 Regression Splines

The approach by Abrahamowicz & Mackenzie involved the use of regression splines in the extended Cox model. They applied their model to 2,404 breast cancer patients enrolled by the Eastern Cooperative Oncology Group, 1,116 recurrences occurred during a median follow-up time of 4.2 years. They found no violation of the PH or log-linearity assumption for the covariates: tumour size, body mass index and log-transformed number of positive nodes. Their main focus was on the effect of age which was identified to violate the log-linearity assumption but satisfied the PH assumption.

Published in 2007, the product model they proposed has been cited by 35 papers. Application of the product model has been used in practice. Binquet and colleagues used this approach to investigate the effects of prognostic factors on mortality in gastric cancers (Binquet et al., 2009). The flexible modelling revealed the effect of age was most important in the first year after diagnosis and mortality increased for both the youngest and the oldest patients. Gagnon and colleagues used the product model to assess C-reactive protein (CRP) and other biomarkers as a prognostic factor for non-small cell lung cancer (NSCLC) and found violations of both linearity and the PH assumption (Gagnon et al., 2010). The importance of testing the assumptions of the Cox proportional hazards model was highlighted as the Cox PH model failed to identify albumin as a prognostic factor and flexible modelling allowed determining that the prognostic value of CRP and albumin does not extend beyond 6 or 12 months, respectively.

### 4.9.3 Additive and Additive-Multiplicative Models

The Cox-Aalen model introduced by Scheike & Zhang was cited by 44 papers. They applied their model to 4000 patients with myocardial infarction from the Trace study group, 2020 deaths were observed in a 3-year period after entering the study. Ventricular fibrillation was identified to have a very strongly time-varying effect; it is an important predictor for death within the first couple of months but after this time-period its effect vanishes.

In 2006 Baldi and colleagues used the Cox-Aalen model for breast cancer survival on Piedmont cancer registry data (Baldi et al., 2006). Their resulting model had an additive part consisting of a baseline, age at diagnosis and tumour size and the proportional part of the model containing regional lymph nodes involvement and comorbidity score. This model allowed identifying tumour size as having a time-varying effect; during the first 18 months after diagnosis its effect was found not to be significant but later it becomes an important predictor for death.

### 4.9.4 Parametric Models

Royston & Parmer introduced flexible parametric models which allowed simultaneously modelling the baseline hazard and non-proportional effects through the use of splines (Royston and Parmar, 2002a). They analysed recurrence-free survival time of 686 patients with primary node positive breast cancer from the German Breast Cancer Study Group, 299 events were observed with a median follow-up of nearly 5 years. They created three prognostic groups (representing good, medium and poor outcome) based on a linear predictor of standard prognostic factors: age, menopausal status, tumour size, tumour grade, number of positive lymph nodes (constrained to be monotonic with an asymptote), progesterone receptor concentration and estrogen receptor concentration. The use of the flexible parametric model indicated time-varying effects of the prognostic groups. Patients in the poor prognostic group experienced a high hazard of recurrence or death during the first 2 to 3 years after diagnosis. After that time the hazard diminished quite quickly and gradually converged toward that of the other two prognostic groups.

The flexible parametric models is a popular approach due to the parametric nature of the model by modelling the baseline hazard and was cited by 115 papers. It has been used in practice due to the well-established and simple to use software available. For example, the approach has been used to model the absolute difference between the treatment groups over time (Dearnaley et al., 2007) and for more robust prediction of survival probabilities and corresponding

confidence intervals due to modelling of the baseline hazard (Dowsett et al., 2010, Nygard et al., 2004). The approach was used to investigate the effect of tumour characteristics (lymph node status, ER status and tumour size) over time by Colzani and colleagues (Colzani et al., 2011). They found that positive lymph node status conferred an increased hazard of dying as a result of breast cancer, even 10 years after diagnosis, the influence of ER status decreased over time and was not significant 5 years after diagnosis and a pattern of decreasing HR was seen for women with tumours larger than 20 mm compared with women diagnosed with tumours between 1 and 20 mm.

### 4.9.5 Source of Citations

The lack of applications of the approaches highlighted through the forward citation searching can be explained by the majority of the citations being in methodological rather than clinical journals. For example the majority of the citations were papers published in the Statistics in Medicine journal which focuses on influencing practice in medicine and its associated sciences through the publication of new statistical and other quantitative methods. Other journals that also appeared regularly in the forward citations were the Biometrical journal, the BMC Medical Research Methodology journal, the Statistical Modelling journal and the Scandinavian journal of Statistics which all generally focus on statistical methods. Forward citations frequently discussed the methodological aspects of the approach with extension to other areas (e.g. relative survival models).

## 4.10 Reviews of the approaches for modelling non-proportional effects

Four reviews were found in the literature search that compared approaches to deal with non-proportional hazards (Table 4.9) with key points of the reviews given in Table 4.10.

### 4.10.1 Data Sets and Approaches Compared

The four reviews compared several different approaches for analysing non-proportional effects in survival data but all included the extended Cox model.

Lehr & Schemper focused on the extended Cox model and looked at various approaches for the representation of relative risks over time from this model (Lehr and Schemper, 2007). Their data set contained only 90 subjects and 11 events which is too small for any real analyses of time-dependent effects but their main focus was on a simulation study of over-fit and power when adding extra parameters to model the time-dependent effects. Cortese and colleagues focused on the limitations of the Cox proportional hazards model and Perperoglou and colleagues considered approaches in modelling long-term survival (Cortese et al., 2010, Perperoglou et al., 2007). Buchholz & Sauerbrei compared five recently published strategies to assess whether and how the effects of covariates from a multivariable model vary in time (Buchholz and Sauerbrei, 2011).

### 4.10.2 Summary of Findings

Lehr & Schemper and Buchholz & Sauerbrei both determined the use of fractional polynomials with an algorithm for selection of the best fitting FP to be a good approach as it allowed maintaining the power of the standard proportional hazards model if effects were constant and adequately modelled non-proportional effects. The MFPT approach recommended by Buchholz & Sauerbrei also incorporated selection of non-linear functional

**Table 4.9** Summary of Review Papers of Approaches for Dealing with Non-Proportional Hazards

| Author (Year) | Data Set | Approaches Compared | Conclusions |
|---|---|---|---|
| Lehr & Schemper (2007) | Gastric cancer study<br>N=90 (events=11)<br>Max length of follow up=5 years approx. | Extended Cox Model -<br>Interactions with Time<br>Simple monotonic functions<br>Polynomials and FP<br>Piecewise constant effects<br>Splines<br>Penalised likelihood approaches | Fractional polynomials and penalised likelihood approaches are the tools of choice. |
| Perperoglou et al. (2007) | IASO breast cancer study<br>N=2433 (events=602)<br>Max length of follow-up=25 years approx. | Reduced rank models<br>Frailty models<br>Cure rate models | If there are not enough events and long-follow up, it is hard to tell the difference between time-dependent, frailty and cure rate models. |
| Cortese et al. (2010) | Norwegian Register breast cancer data<br>N=9041 (events=6202)<br>Median follow-up time=10 years<br>Max follow-up time=26 years | Cox model<br>Extended Cox model<br>Aalen Additive hazards model<br>Cox-Aalen model<br>Excess Hazards model | Both flexible multiplicative and additive hazards models provide useful instruments for investigating if non-proportional effects are present in the data. |
| Buchholz & Sauerbrei (2011) | GBSG study<br>N=686<br>Rotterdam breast cancer survival<br>N=2982 (events=1518)<br>Median follow-up time=8.9 years<br>Max follow-up time=19.25 years | MFPT approach<br>Dynamic Cox model<br>Reduced Rank model<br>Empirical Bayes model<br>Semi-parametric extended Cox model | When the interest lies on the shape of non-proportional effects the MFPT approach seems to preferable. |

**Table 4.10** Key Points of Review Papers of Approaches for Dealing with Non-Proportional Hazards

| Author (Year) | Key Points |
|---|---|
| Lehr & Schemper (2007) | • Investigation into over-fitting of non-proportional effects<br>• Consequence of over fit<br>  ○ Increase in width of pointwise confidence intervals<br>  ○ Loss of power to confirm non-proportional effect<br>• Increased variability of effect when adding additional parameters<br>  ○ Piecewise constant approach most variable<br>• FP approach<br>  ○ allowed maintaining power if effects were constant<br>  ○ adequately modelled non-proportional effects<br>  ○ chose most parsimonious representation of the non-proportional effect<br>• Width of confidence intervals increased with increasing degrees of freedom |
| Perperoglou et al. (2007) | • Emphasis on plots of hazards and survival rather than the coefficients<br>• Direct comparison was difficult due to the different properties and background assumptions<br>• Reduced Rank Models<br>  ○ Helped identify true nature of time-dependency<br>  ○ Provided means for parsimonious modelling through fitting fewer parameters<br>  ○ Difficult to express coefficients for the time-varying effects<br>  ○ Poor predictive performance with evidence of some over-fitting<br>• Frailty Models<br>  ○ Do not assume proportional hazards<br>  ○ Complex model due to interaction of baseline hazard with main effects<br>  ○ Important to plot survival and hazard curves, not only draw conclusions from coefficients<br>• Cure Models<br>  ○ Can explain unexpectedly large survival times<br>  ○ Potential to identify outliers<br>  ○ However, breast cancer patients are never considered cured |
| Cortese et al. (2010) | • Cox PH model<br>  ○ Clear violation of model assumptions<br>  ○ Did not capture all important aspects of the data<br>• Multiplicative flexible models<br>  ○ Rely on smoothing techniques which makes a description of the uncertainty difficult<br>• Additive models<br>  ○ Easy to fit and require no assumptions on the degree of smoothing<br>  ○ May lead to increasing predicted survival curves |

**Table 4.10 (Continued)** Key Points of Review Papers of Approaches for Dealing with Non-Proportional Hazards

| Author (Year) | Key Points |
|---|---|
| Buchholz & Sauerbrei (2011) | • Demonstrated that<br>  ○ mismodelling of the functional form of a continuous covariate may induce spurious non-proportional effects<br>  ○ omission of covariates can have a severe influence on the assessment of the non-proportional effect<br>• MFPT approach<br>  ○ Easy to use<br>  ○ Addressed variable selection, selection of non-linear functional forms as well as non-proportional effects<br>• Dynamic Cox model<br>  ○ Equally flexible and easy to use<br>  ○ Problems with enlargements of data sets<br>• Reduced Rank model<br>  ○ Theoretically a flexible tool<br>  ○ However algorithm does not include variable selection, selection of non-linear effects, nor selection of non-proportional effects<br>  ▪ May result in serious over-fitting and limits transferability of selected models<br>• Empirical Bayes model<br>  ○ Easily applied<br>  ○ Combinations of non-linear and time-varying effects are not allowed<br>• Semi-parametric extended Cox model<br>  ○ Theoretically flexible tool<br>  ○ Application is more difficult |

forms through the use of fractional polynomials to avoid finding spurious non-proportional effects.

Perperoglou and colleagues and Buchholz & Sauerbrei both determined that the reduced rank model is theoretically a flexible tool for investigating time-dependence but it results in over-fitting due to all effects being fitted as time-varying.

Cortese and colleagues were focused on the limitations of the Cox model and provided a good example of fitting the Cox PH model and various methods for examining how well this model fitted the data and investigating the PH assumption. They briefly mentioned investigating the

functional form of the continuous covariates but results and further implications were not given. Although they investigated several different approaches these were compared to results from the Cox PH model and so a clear comparison of the various approaches was not available. Therefore they did not conclude on a preferred approach but determined all were an improvement on the Cox PH model.

## 4.11 Discussion

The disadvantages of the Cox model are continually being discussed yet it is still being used with the underlying assumptions often not being checked and possibly not valid. The use of the Cox model when the assumptions are not valid can result in misleading conclusions. Bellera and colleagues found variables not satisfying the PH assumption results in a reduction in power of the corresponding tests and so are more likely to make a type II error (accept the null hypothesis when it is false) (Bellera et al., 2010). Testing of the PH assumption should be an integral part of an analysis using the Cox model with both graphical and numerical approaches available to do this (see chapter 3).

Many of those who do check the assumption of proportional hazards and find it not valid use the simple approach of splitting the time scale, known as piecewise constant effects, but the drawbacks of this approach are well established. The categorising of the time-scale results in a reduction of power and the number of intervals chosen can have an impact on the conclusions made. Piecewise models yield 'jumpy', clinically implausible, estimates with difficulties implementing in multivariable modelling (Quantin et al., 1999).

A large number of alternatives to piecewise constant effects were highlighted by this review; from fully parametric to non-parametric models, multiplicative effects of the covariates to additive effects, different methodological assumptions such as cure and frailties, as well as an expanding body of literature on Bayesian approaches.

A key point that was continually made was that the functional form of the covariate on the log-hazard must be correct to avoid spurious evidence of non-proportional effects. Two approaches that specifically mentioned incorporating non-linear effects of continuous covariates on the log-hazard as well as non-proportional effects: the multivariable fractional polynomial time (MFPT) approach by Sauerbrei and colleagues and the product model proposed by Abrahamowicz & Mackenzie (Sauerbrei et al., 2007, Abrahamowicz and MacKenzie, 2007). They both made use of the extended Cox model with fractional polynomials and regression splines respectively. Buchholz & Sauerbrei recommended the use of the MFPT approach which they compared to other multiplicative models (Buchholz and Sauerbrei, 2011). This approach has not been found to be used in practice and it will be of interest to apply to data analyses throughout this project.

Royston & Parmer proposed a flexible parametric approach which involved modelling the baseline hazard and non-proportional effects through the use of regression splines (Royston and Parmar, 2002a). It was a popular approach with a relatively large number of citations and has been used in practice due the parametric nature of the model and the readily available software to implement this approach. They state "The Cox model is convenient with its strength lying in estimating relative hazards for covariate effects. It is semi-parametric in nature and was never intended for the task of comprehensive prediction of survival probabilities and other important quantities"(Royston et al., 2010). They therefore recommend the use of adequately flexible parametric models for estimation and validation of prognostic models. This will be a key approach for modelling the baseline hazard and estimating non-proportional effects.

The citation searching highlighted the lack of application of the available approaches for modelling non-proportional effects to data sets. Binquet and colleagues state the need for more methodological research in this area, more widespread use of flexible modelling in prognostic

studies and encourage others to explore similar flexible analyses on prognostic studies (Binquet et al., 2008, Binquet et al., 2009).

### 4.11.1 Limitations

The limitations of this literature review include the reliance on published research and finding of the studies using the method outlined in the search strategy (section 4.2). The search strategy included finding terms in the title and abstract of studies and the use of focus terms limited the results returned and possibly omitted important studies. However, references of studies that were highlighted as relevant were searched and a relatively large number of papers were included in the final review, giving a broad and comprehensive overview of approaches for analysing non-proportional effects.

Limitations were involved in the citation searching to determine which of the approaches available for investigating non-proportional effects for survival data had been used in practice. Often well-known approaches such as the extended Cox model and piecewise constant effects are not explicitly cited so it was decided to focus on newer approaches that are more likely to be cited. Although the citations revealed a lack of using the approaches in practice this does not explicitly mean it has not been used, just a possibility the approach was not cited.

# Chapter 5: IHC4 and Mammostrat

## 5.1 Introduction

The focus of this project is on two residual risk marker panels, IHC4 and Mammostrat. This chapter will cover the development of these two models and the statistical methods applied in previous evaluations of the models.

This chapter will also cover some of the recent prognostic models for breast cancer which have investigated time-dependent effects.

## 5.2 Model Development and First Validation

### 5.2.1 IHC4

#### *5.2.1.1 Development*

IHC4 score was developed by Cuzick and colleagues to combine four conventional IHC markers, ER, PgR, HER2 and Ki67 to select patients at increased residual risk after adjuvant endocrine therapy (Cuzick et al., 2011).

The risk score was developed on women from the tamoxifen and anastrozole arms of the Arimidex, Tamoxifen, Alone or in Combination (ATAC) trial reported to be ER-positive and/or PgR-positive. Women who received chemotherapy before trial entry and those found to be both ER-negative and PgR-negative on central review of the sections were excluded, leaving 1,125 eligible patients (145 distant recurrences) with measurements for all parameters. The primary end-point was time to distant recurrence (TTDR) and was based on a median follow-up of 8.3 years.

The contribution of each of the four variables was evaluated by the change in likelihood ratio in three ways: by univariate Cox analyses, as an addition to a model containing only the clinical

variables, and as a decrease in likelihood ratio when the variable was removed from the full model containing clinical variables (age, nodal status, tumour size, centrally read grade, and randomised treatment) and all four IHC variables. In the paper it states ‘a model was then created that used all the data to give the best IHC4 score when combined with the classical variables, leading to an overall prognostic score. To allow for the small amount of over-fitting of the IHC4 score, a shrinkage correction was used for the final score’. It is not clear what they mean by ‘best’ IHC4 score and how this was determined. All clinical variables were categorical, which results in a loss of information from continuous age, number of positive nodes and tumour size, which could have inflated the prognostic information added by the IHC markers.

An aim of their paper was to compare the added information from their newly developed IHC4 score to Oncotype Dx. To do this, the IHC4 score was then generated using half the data, and the comparison between IHC4 and Oncotype Dx in the remaining half. It is not clear what the differences were when ICH4 was developed using the full dataset.

It was not mentioned in the paper whether the assumptions of linearity or proportional hazards were assessed when performing the Cox regression. The final score comprised a transformation of one of the IHC markers, suggesting that different functional forms were explored and perhaps this was what was implied by determining the ‘best’ model.

#### *5.2.1.2 Validation*

The first validation was performed as a comparison to the Oncotype Dx test on the remaining half of the data not used in the development of IHC4. The comparison between IHC4 and Oncotype Dx was assessed for all patients and those node-negative by changes in the likelihood ratio statistic and a comparison of Kaplan-Meier survival curves for each score. They found that the ICH4 score provide additional prognostic information that is at least as informative as the Oncotype Dx test.

Further validation of the IHC4 score was performed using a cohort of 786 women treated in Nottingham from 1990 to 1998. All of these patients were ER-positive and received either tamoxifen or no endocrine treatment. There were 174 distant recurrences in this cohort with a median follow-up of 9.5 years. IHC4 was assessed by univariate and multivariate Cox regression analysis, and a comparison between observed and predicted time to distant recurrence. The IHC4 score was significantly prognostic of outcome in a univariate analysis and after adjustment for clinical factors (nodal status, tumour size, grade, age and treatment). There was no mention of testing the modelling assumptions of linearity and proportional hazards in the Cox model.

### 5.2.2 Mammostrat Score

#### *5.2.2.1 Development*

Mammostrat score was developed by Ring and colleagues, to explore whether insights from gene expression studies could be translated into robust IHC-based assays by generating hundreds of novel antibody reagents targeted to predicted protein products of genes selected on the basis of their gene expression patterns (Ring et al., 2006).

The development cohort was 466 patients (122 recurrences) with primary invasive breast cancer seen at the Comprehensive Cancer Institute of Huntsville (CCIH). Close to 700 gene targets for generation were chosen on the basis of interesting gene expression patterns in published data sets. Prognostic models were derived using 195 ER-positive, node-negative patients in the CCIH training cohort and the subset of 20 antibodies that had a significant association (based on a log-rank test) with outcome in this patient set. The number of recurrences in this subgroup of patients and the median length of follow-up were not provided in the paper. An explicit statement of the primary outcome was not given, however it is assumed that disease-free survival was used, due to discussion of disease-free survival rates in the paper.

Cox proportional hazards regression model building used initially all of the antibodies nominated by the univariate tests. A backward selection method was used to remove antibodies that did not add significantly to the fit of the model, until removal of antibodies caused a significant reduction in the strength of the model. This resulted in a final model consisting of five antibodies. The prognostic score was classified into three prognostic categories, but it was not stated how they determined the cut-points for these categories. There was no mention of testing the modelling assumptions of linearity and proportional hazards in the Cox model.

#### *5.2.2.2 Validation*

In their initial paper, Ring and colleagues presented the results of the validation of Mammostrat score in two independent cohorts: 299 consecutive primary invasive breast cancer patients (61 recurrences) seen at the Cleveland Clinic Foundation (CCF); 344 consecutive primary invasive breast cancer patients (56 deaths – recurrences not available) seen at Vancouver General Hospital then linked to an anonymous British Columbia Cancer Agency (BCCA) database containing follow-up information. For the CCF data set, disease-free survival was assessed. For the BCCA data set, overall survival was assessed (disease-free survival was unavailable), and death from other causes was censored. Follow-up time in both validation cohorts was limited to 5 years.

The prognostic ability of Mammostrat score was assessed using Kaplan-Meier plots, univariate and multivariate Cox regression (adjusted for stage, grade and lymph node status), sensitivity, specificity and positive and negative predicted values. Overall, Mammostrat score was found to provide good stratification of patients into prognosis groups and remained an independent predictor of outcome in multivariable cox regression.

## 5.3 Independent Evaluations of IHC4 Score

Since the publication of IHC4 score in 2011, the score has been evaluated in several publications.

### 5.3.1 Contribution of the IHC4+C score to decision making in clinical practice

Barton and colleagues compared the prognostic information gained from IHC4+C with that from Adjuvant! Online (AoL) and NPI, in a group of postmenopausal women with early breast cancer receiving treatment (Barton et al., 2012). Their aim was to determine whether assessment of IHC4+C assists in the selection of patients who may be safely treated with adjuvant hormone therapy alone and spared the side-effects of chemotherapy.

The primary endpoint was the proportion of patients re-allocated from AoL-defined intermediate risk of distant recurrence at 10 years, to either high or low risk, by application of the IHC4+C score. The first secondary endpoint was the proportion of patients reallocated from NPI-defined moderate risk to either high or low risk, by application of the IHC4+C score.

The analysis was performed on 101 hormone receptor-positive, HER2-negative early breast cancer patients with ≤3 axillary lymph nodes containing metastases.

The IHC4+C score downgraded more than half of the patients in the AoL-defined intermediate-risk group to a low-risk group. Nearly two-thirds in the NPI-defined intermediate-risk group were reallocated into either a low- or high-risk group, with risk stratification most often lowered. They concluded the use of IHC4+C may substantially improve decision-making on adjuvant therapy.

### 5.3.2 Comparison of PAM50 Risk of Recurrence Score with Oncotype DX and IHC4

Dowsett and colleagues compared the ability of the risk of recurrence (ROR) score generated from the 50-gene PAM50 test with that of the Oncotype DX 21-gene recurrence score (RS) and IHC4 score in predicting risk of distant recurrence (DR) after endocrine therapy (Dowsett et al., 2013*)*.

The analysis was performed on the same dataset from the development of IHC4, women from the tamoxifen and anastrozole arms of the ATAC trial reported to be ER positive and/or PgR

positive. Therefore the comparison between IHC4 and PAM50 was performed using a sample-splitting approach. . 940 patients had data on all three scores with 154 distant recurrences. Subgroup analysis was performed on node-negative patients and HER2-negative/node-negative patients.

To assess the prognostic value of each score (ROR, RS, IHC4), two approaches were used: the changes in the likelihood ratio values to measure quantitatively the relative amount of information of one score compared with another; and the concordance index (c index) measure of concordance for time-to-event data.

IHC4 added as much information to clinical factors (nodal status, tumour size, histopathologic grade, age, and treatment) as ROR but in the HER2-negative/node-negative subgroup, ROR added significantly more information than IHC4.

Sestak and colleagues then continued the analysis of ROR, RS and IHC4 for early (0-5 year) and late (5-10 year) distant recurrence (Sestak et al., 2013). After adjustment for the clinical treatment score (tumour size and grade, lymph node status, age, and treatment), IHC4, RoR and RS were prognostic of early recurrence with IHC4 providing more prognostic information than RoR and RS. The scores also remained prognostic in the 5-10 year interval, with RoR providing more prognostic information than IHC4 and RS.

### 5.3.3 Comparison of the breast-cancer index (BCI) assay, Oncotype Dx, and IHC4

Sgroi and colleagues compared the prognostic ability of the breast-cancer index (BCI) assay, 21-gene recurrence score (Oncotype DX), and IHC4 for both early and late recurrence in patients with ER-positive, node-negative disease who took part in the Arimidex, Tamoxifen, Alone or in Combination (ATAC) clinical trial (Sgroi et al., 2013). Again the analysis was performed on the same dataset used in the development of IHC4 and sample-splitting was used.

Distant recurrence was the primary endpoint and the primary analysis population was patients with ER-positive, node-negative breast cancer, with the secondary analysis populations including patients with ER-positive, node-negative, HER2-negative breast cancer and those with ER-positive, node-positive breast cancer. The analysis included 665 patients with 72 distant recurrences.

The primary objective was to determine the prognostic ability of the BCI model for distant recurrence. Secondary objectives were to assess the BCI model and it's components for overall (0-10 year), early (0-5 year) and late (5-10 year) distant recurrence and compare with that of RS and IHC4.

Likelihood ratio tests based on Cox proportional hazards regression models were used to test for a significant difference between a reduced proportional hazards model based on clinical treatment score and a full proportional hazards model, including BCI, RS, or IHC4. The improvement in prediction was quantified by the change in the likelihood ratio chi-squared value, which measures the amount of information added to the proportional hazards model by the gene signatures compared with clinical treatment score.

BCI was significantly associated with both early and late distant recurrence with adjustment for clinical treatment score (tumour size and grade, lymph node status, age, and treatment). For early recurrence, in all node-negative patients, IHC4 was more prognostic than RS and BCI. However, neither IHC4 nor RS were prognostic of late distant recurrence.

### 5.3.4 Clinical Impact of using the IHC4 Score

Yeo and colleagues presented their results of 'the clinical impact of using the IHC4 score: our MDT experience in a prospective series of postmenopausal women with ER-positive early breast cancer' at the IMPAKT Breast Cancer Conference (Yeo et al., 2014). Their study aim was to prospectively evaluate the change in adjuvant treatment decision making based on the availability of the IHC4 score. The primary endpoint was the percentage change in the MDT

recommendation for systemic therapy before and after the availability of the IHC4 score. They concluded the IHC4 score has the potential to change a significant proportion of patients' adjuvant treatment recommendations given through the MDT in this small prospective study.

## 5.4 Independent Evaluations of Mammostrat Score

### 5.4.1 Chemo-sensitivity and Stratification of Mammostrat in the NSABP B14 and B20 Trials

Ross and colleagues aimed to determine the association between the Mammostrat test and clinical outcomes in ER-positive, node-negative tamoxifen-treated breast cancer patients and to determine whether the test identified patients who would have selectively benefited from adjuvant chemotherapy treatment (Ross et al., 2008).

The NSABP B14 and B20 studies were randomized prospective studies of the role of adjuvant tamoxifen and adjuvant chemotherapy in the treatment of node-negative, ER-positive breast cancer, respectively. The analysis included 711 patients with number of events and median length of follow-up not stated.

The primary aim was to test on tamoxifen treated patients from both the B14 and B20 trials that the Mammostrat risk groups were significantly associated with recurrence-free interval (RFI) in a univariate Cox model and was tested using the likelihood ratio test.

Prediction of chemotherapy benefit was carried out by comparing patients treated with tamoxifen only to those treated with tamoxifen plus chemotherapy using patients from the B20 trial. The primary aim was to test that the distribution of recurrence events in comparison between tamoxifen-treated and tamoxifen plus chemotherapy-treated patients during the 10-year study period was significantly distinct and was tested with an interaction between Mammostrat risk groups and chemotherapy.

The Mammostrat score was shown to be associated with RFI in a univariate analysis and after adjustment for age and tumour size. It was shown patients in both the low and high-risk groups

benefited from chemotherapy with an improvement in a univariate analysis, but there was no significant interaction between the risk groups and chemotherapy.

### 5.4.2 Mammostrat as a tool to stratify breast cancer patients at risk of recurrence during endocrine therapy

Bartlett and colleagues explored the relationship of Mammostrat risk stratification on patients from the Edinburgh Breast Conservation Series (BCS), where patients were those considered suitable for breast-conserving therapy with a tumour size of < 3 cm, node-negative or positive, and no metastases on conventional TNM (tumour size, nodes, metastases) staging (Bartlett et al., 2010). The analysis included 1,540 patients treated with adjuvant tamoxifen, other hormonal therapy and with both hormonal therapy and chemotherapy, excluding patients treated with chemotherapy only. The number of events and median length of follow-up was not stated in the paper.

The primary analysis was to assess the association with Mammostrat risk groups on distant-recurrence free survival (DRFS) in ER-positive, node-negative cases treated with tamoxifen only and in ER-positive, tamoxifen-treated cases irrespective of nodal status. Univariate and multivariate Cox regression analysis was used to estimate hazard ratios and 95% confidence intervals and corresponding p-value.

Mammostrat risk score as a predictor of DRFS was of borderline significance for ER-positive, node-negative cases treated with tamoxifen only after adjustment for nodal status, grade, size, multifocality, menopausal status, age, HER2, PgR and ER. Mammostrat score remained an independent predictor when ER-positive, tamoxifen-treated cases irrespective of nodal status were considered.

It was concluded the data added support for the potential role of Mammostrat in management of early-stage breast cancer.

### 5.4.3 Mammostrat as a predictor of Early Relapse Risk in the TEAM Study

Bartlett and colleagues then explored the effect of Mammostrat on residual risk of relapse after treatment with either 5 years of exemestane (steroidal aromatase inhibitor) or 5 years of endocrine therapy by using a switch strategy with tamoxifen (2 to 3 years) followed by exemestane (3 to 2 years) within the large clinical trial cohort, the Tamoxifen versus Exemestane Adjuvant Multicenter (TEAM) trial of postmenopausal hormone-receptor positive patients (Bartlett et al., 2012). The TEAM pathology substudy consisted of 4,598 cases suitable for TMA construction and Mammostrat staining, with 954 disease relapses with a median follow-up of 5.1 years.

The primary analyses were to test that Mammostrat was an independent marker of residual risk of DRFS after 5 years of adjuvant endocrine therapy (either exemestane or tamoxifen followed by exemestane): in ER-positive, node-negative patients treated with endocrine therapy (but not with chemotherapy), in ER-positive patients irrespective of nodal status treated with endocrine therapy (but not chemotherapy), and in all patients with hormone receptor-positive breast cancer treated with endocrine therapy. Univariate and multivariate Cox regression analysis was used to estimate hazard ratios and 95% confidence intervals and corresponding p-value.

Mammostrat was an independent predictor of DRFS after adjustment for PgR, HER2, age, grade, size, nodes and chemotherapy in all patient subgroups.

It was concluded for postmenopausal women with ER-positive early breast cancers, irrespective of nodal status, the Mammostrat panel provided independent information on residual risk of distant recurrence or death as a result of breast cancer after treatment with exemestane or a switch regimen (tamoxifen followed by exemestane) in a retrospective analysis of the TEAM pathology study cohort.

## 5.5 Time-dependent analysis of other prognostic scores

It was of interest to examine any previous analyses of prognostic models in breast cancer in relation to evidence of non-proportional effects.

### 5.5.1 Validation and Clinical Utility of MammaPrint

Buyse and colleagues aimed to validate the 70-gene signature, MammaPrint (Buyse et al., 2006).

Participants in this study included 326 patients from five European centres who were younger than 61 years old at diagnosis, diagnosed before 1999 with node-negative, T1 – T2 (≤ 5 cm) breast cancer, and had not received adjuvant systemic therapy.

Three main endpoints were analysed: time from surgery to distant metastases; overall survival; and disease-free survival. There was a median follow-up of 13.6 years and events included 68 recurrences, 31 second primary cancers, 77 distant metastases, and 82 deaths.

Validation was based on the estimation of hazard ratios and 95% confidence intervals and sensitivity and specificity were estimated for distant metastases within 5 years and for death within 10 years. The impact of follow-up duration was evaluated by looking at different time-points with arbitrary censoring of all observations at increasing time points.

The gene signature was shown to be a strong prognostic factor for time to distant metastases and overall survival in untreated patients with node-negative breast cancer after adjustment for various risk classifications that take into account clinicopathologic factors (Adjuvant! Online). The gene signature was shown to have the greatest prognostic value at identifying patients at high risk of an event within 5 years.

### 5.5.2 Strong Time Dependence of the 76-Gene Prognostic Signature

Desmedt and colleagues aimed to independently validate the 76-gene signature, identified by Erasmus Medical Center and Veridex that could be used to predict the development of distant

metastases within 5 years in node-negative primary breast cancer patients (irrespective of age and tumor size) who did not receive systemic treatment (Desmedt et al., 2007).

198 patients from the TRANSBIG study were included in the analysis who were younger than the age of 61 years and had node-negative, T1-T2 (≤5 cm) tumours. Patients in this series had been diagnosed between 1980 and 1998 (median follow-up, 13.6 years) and had been seen at six centres.

The study end-points were time from diagnosis to distant metastases (TDM) and overall survival. The median follow-up for the 198 patients included was 14.0 years, with 51 distant metastases and 35 with progression within 5 years.

Hazard ratios were calculated from Cox proportional hazards models, stratified by clinical risk as defined by Adjuvant! Online to estimate the prognostic effect of the signature above that provided by clinicopathological variables. Sensitivity and specificity was calculated for distant metastasis within 5 or 10 years. The effect of the duration of follow-up on HRs was analysed by censoring all observations at increasing time points.

The signature was found to be prognostic of outcome after adjustment for clinical risk when considering 10-year follow-up, but a strong time-dependence of the signature was identified providing the most prognostic information in the first 5 years of follow-up.

It was concluded the prognostic ability of the 76-gene signature was confirmed. However, again there was no mention of assessing the proportional hazards assumption and all hazard ratio estimates had very wide confidence intervals.

### 5.5.3 Comparison of prognostic gene expression signatures for breast cancer

Haibe-Kains and colleagues aimed to statistically compare three gene signatures in terms of predicting clinical outcome for the individual patient. The signatures were the 70-gene

signature, MammaPrint, the 76-gene signature and the Gene expression Grade Index (GGI) score based on 97 genes (Haibe-Kains et al., 2008).

198 patients from the TRANSBIG study were included in the analysis. The patients were younger than the age of 61, had node-negative, T1–T2 (≤ 5 cm) tumours and did not receive any adjuvant treatment. The endpoints considered were distant metastases free survival, time to distant metastases and overall survival. The number of events and length of follow-up were not stated in the paper.

Survival analysis was performed considering 10-year and full follow-up (25 years) using Cox regression models to estimate hazard ratios and the C-index to quantify concordance. Sensitivity and specificity was estimated at several lengths of follow-up (3, 5, 10, 15 years and full follow-up).

The three signatures added significant information to the traditional parameters (age, tumour size, ER status, and grade) when follow-up was censored at 10-years. The 76-gene signature was not prognostic of outcome when full follow-up was considered. A decrease in model performance was observed, according to the sensitivity and specificity, with increasing follow-up duration (10 years and more) for all gene signatures.

It was concluded that the gene signatures showed similar prognostic performance with a strong time dependence of the classification performance. However, there was no mention of assessing the proportional hazards assumption when the Cox regression analysis was performed over the 10 or 25-year follow-up.

### 5.5.4 EndoPredict prognostic of late distant metastases

Dubsky and colleagues assessed whether the prognostic EndoPredict (EP) score, a multigene score that combines the expression levels of proliferative and ESR1 signalling/differentiation-

associated genes, identifies late relapse events in ER+/HER2- breast cancer patients (Dubsky et al., 2013).

1701 patients were included in this study who participated in the ABCSG6 (tamoxifen-only arm) or ABCSG8 trial and received either tamoxifen for 5 years or tamoxifen for 2 years followed by anastrozole for 3 years. None of the patients received adjuvant chemotherapy. The primary end-point used in the analysis was distant metastases with number of events not stated in the paper.

Kaplan-Meier survival estimates were estimated in two different time periods: 0-5 years and >5 years. Multivariable Cox regression analysis was also performed on the two time intervals. The c-index was used to assess the prognostic information provided by the EP signature and clinicopathological variables.

In the multivariate analysis EP was an independent predictor of distant metastases after adjustment for age, grade, lymph node status, tumour size and Ki67 in the first and second time interval. The prognostic performance as assessed by the C-index was significantly improved by adding the molecular information of the EP test to the clinicopathological variables.

It was concluded that the EPclin (EP score combined with tumour size and nodal status) reliably identified a subgroup of patients who have an excellent long-term prognosis after 5 years of endocrine therapy.

### 5.5.5 Breast Cancer Index Identifies Breast Cancer Patients at Risk for Early- and Late-Distant Recurrence

Zhang and colleagues aimed to develop an optimized BCI risk model and validate its prognostic performance with pre-specified risk groups for predicting early (0–5 years) and late (>5 years) distant recurrences (Zhang et al., 2013).

This BCI model was validated in a retrospective analyses of two cohorts: first, 808 tamoxifen-treated ER+, LN- patients enrolled in the Stockholm trial and second, a multi-institutional cohort consisting of ER+, LN- tamoxifen-treated patients from 2 academic medical centres which included 265 and 93 patients respectively.

The study endpoint was distant recurrence-free survival. The Stockholm cohort had a median follow-up time for all patients of 17 years and the numbers of distant recurrences were 56 and 33 for untreated and tamoxifen-treated patients, respectively, with 52% and 61% of the recurrences occurring after 5 years. The multi-institutional cohort had an overall median follow up time for all patients of 10 years and 57 distant recurrences with 40% of the recurrences occurring after 5 years.

Multivariate Cox proportional hazard models were used to evaluate whether BCI provided prognostic information independent of traditional clinical and pathologic factors (age, tumour size, tumour grade, HER2 status, PR status, and chemotherapy if appropriate) using likelihood ratio tests.

BCI was a significant predictor of overall (0-10 years), early (0-5 years) and late distant recurrence risk (5-10 years) after adjustment for clinicopathological variables in both cohorts.

It was concluded that the prognostic ability of BCI to assess early- and late-distant recurrence risk at diagnosis has clinical use for decisions of chemotherapy at diagnosis and for decisions for extended adjuvant endocrine therapy beyond five years.

### 5.5.6 Long-term impact of MammaPrint on breast cancer outcome

Drukker and colleagues aimed to evaluate the effect of the 70-gene signature, MammaPrint after longer follow-up (Drukker et al., 2014).

The cohort included 295 patients, who were females diagnosed at the Netherlands Cancer Institute between 1984 and 1995, younger than 53 years with histologically proven, operable,

invasive breast cancer (tumour size <3cm, node-negative or positive, no metastases). All patients were primarily treated with breast conserving surgery or mastectomy. Adjuvant treatment consisted of radiotherapy, chemotherapy and/or endocrine therapy as indicated by guidelines used at the time of treatment.

End-points included overall survival (OS) and distant metastasis-free survival (DMFS). There was a median follow-up of 18.5 years, with 121 distant metastasis events, 127 deaths of which 114 were attributed to death due to breast cancer.

Hazard ratios and 95% confidence intervals and p-values were calculated from the Cox regression analysis for the full follow-up (max 25 years) and per 5 year intervals.

The 70-gene signature was shown to have the largest prognostic value for DMFS and OS in the first 5 years of follow-up. A prognostic effect for distant metastases was claimed over the 25 year follow-up, despite a significant hazard ratio confined to first five years of follow-up and no mention of assessing the proportional hazards assumption when full follow-up (25 years) was considered. The analysis was also not adjusted for any known prognostic clinical factors.

## 5.6 Discussion

IHC4 has been shown to provide significant prognostic information which is at least as informative as the Oncotype Dx test, the PAM50 test and BCI when considering 10 year follow-up. When considering early (0-5 years) and late (5-10 years) distant recurrence, IHC4 provided more prognostic information on early distant recurrence compared to Oncotype Dx and PAM50 but PAM50 was the stronger predictor for late distant recurrence. Similarly IHC4 was the stronger predictor of early distant recurrence compared with BCI but only BCI remained prognostic in the 5-10 year period.

However, the majority of evaluations of the IHC4 residual risk model have been performed on the same dataset as it was developed, performing the analysis on half of the dataset not used in the development of IHC4 score. It is therefore of benefit to evaluate this model on completely independent datasets. The IHC4 score has also not been evaluated after longer follow-up.

Mammostrat has been shown to provide prognostic information on residual risk of distant recurrence or death as a result of breast cancer after treatment in several cohorts, but it has not been compared with other prognostic models. The impact of follow-up time has also not been explored for this score.

Overall, there is no mention of assessing the proportional hazards assumption despite all of the papers using the Cox proportional hazards model. The only method used to evaluate the time-dependence of the scores was with splitting of the time-scale into 5 year intervals and arbitrarily increasing the censoring time.

# Chapter 6: The Data

In this work data analyses will be performed on two breast cancer patient series.

## 6.1 Materials

### 6.1.1 The Edinburgh Breast Conservation Series

The Edinburgh Breast Conservation Series (BCS) represents a fully documented, consecutive cohort of 1,812 patients treated by breast conservation surgery, axillary node sampling or clearance, and whole breast radiotherapy at the Edinburgh Cancer Centre between 1981 and 1998 (Bartlett et al., 2010, Thomas et al., 2009). Following ethical approval (Lothian Local Research Ethics 04) tissue blocks were retrieved from all cases and sufficient material was available from 1,686 cases for assembly into tissue microarrays (TMAs)(see section 6.2.2.1 for more detail). For all cases with available tissue, tumours were regraded on whole sections by a single pathologist (Thomas et al., 2009). Data were not available on patients treated with chemotherapy only (n=146) and patients treated with non-tamoxifen based endocrine therapy (n=91) resulting in 1,449 patients available for analyses

### 6.1.2 The Tamoxifen Exemestane Adjuvant Multinational Trial

The TEAM (Tamoxifen Exemestane Adjuvant Multinational) trial is a multinational randomised, open-label, phase III trial in postmenopausal women with hormone receptor-positive early breast cancer testing the efficacy of 5 years of exemestane (25 mg once per day) versus tamoxifen (20 mg once per day for 2.5-3 years) followed by exemestane for a total of 5 years (van de Velde et al., 2011). Patients were recruited in the years 2001 to 2006. Five of nine participating countries provided paraffin-embedded tumour samples for pathology sub-studies (Bartlett et al., 2011a) resulting in 6,120 patients available for analyses

## 6.2 Variables

### 6.2.1 Clinical Variables

Prognostic clinical variables that will be used in the analyses areage (continuous), tumour size (continuous), number of positive nodes (a count), histological grade (grade I-III), treatment (exemestane/tamoxifen in TEAM) and chemotherapy (yes/no).

### 6.2.2 Biomarkers

The IHC4 model (Cuzick et al., 2011) utilized a linear combination of multiple markers: ER, PgR, HER2 and Ki67. Continuous marker scores were normalised prior to inclusion in the IHC4 model. ER histoscores were divided by 30 and PgR as a percentage of cells staining positive were divided by 10 to obtain continuous values between 0 and 10. Ki67 scores were represented as percentage positive cells and HER2 was treated as a dichotomous variable. The IHC4 risk score was generated according to the previously specified algorithm(Cuzick et al., 2011):

$$IHC4\ Score = 94.7 * \Big(-0.1 * ER_{10} - 0.079 * PgR_{10} + 0.586 * HER2_{0/1} + 0.24 * Ln\big(1 + 10 * Ki67_{\%pos}\big)\Big).$$

The IHC4 score is analysed as a continuous risk score, except for Kaplan-Meier analyses where the IHC4 score is categorised into three groups using two cutoff points that correspond to a 10 year distant recurrence rate of 10% and 20% from the original study, however these cutoffs have not been prospectively validated (Cuzick et al., 2011).

The Mammostrat model (Ring et al., 2006) used five IHC markers: SLC7A5, CEACAM5, NDRG1, HTF9C and p53. The Mammostrat risk score was generated by combining binary staining results for all markers as either positive or negative according to the previously

specified algorithm (Bartlett et al., 2010, Ring et al., 2006, Ross et al., 2008, Bartlett et al., 2012):

$$Mammsotrat\ Score = (1.54 * SLC7A5) + (1.12 * p53) + (1.06 * NDRG1) + (0.72 * HTF9C) + (0.50 * CEACAM5) - 0.86\ .$$

The Mammostrat score was categorised into low (≤ 0), medium (>0 and <0.7) and high (≥ 0.7) risk categories as previously specified (Ring et al., 2006, Bartlett et al., 2010, Bartlett et al., 2012, Ross et al., 2008). The Mammostrat score is not analysed as a continuous score due to the skewed distribution (Figure 6.1). Model fit of continuous versus categorical variable was assessed by the AIC (Akaike information criterion(Akaike, 1974)). It is an index used to aid in choosing between competing models and is defined as $-2L_m + 2m$ where $L_m$ is the maximised log-likelihood and *m* is the number of parameters in the model. The AIC will be used in further analyses as well as the BIC (Bayes information criterion(Schwarz, 1978)) which is defined as $-2L_m + m \times \ln(n)$ where *n* is the number of observations. Analysing Mammostrat score as a categorical variable improves the model fit as assessed by the AIC from a univariate Cox regression model whereas categorising IHC4 results in a loss of information (Table 6.1).

**Table 6.1** Comparison of scores as categorical or continuous predictors as assessed by AIC from a univariate Cox regression model.

| | Mammostrat | | | IHC4 | | |
|---|---|---|---|---|---|---|
| AIC | Cont. | Cat. | Diff. | Cont. | Cat. | Diff. |
| Edinburgh BCS | 3347 | 3340 | 7 | 2575 | 2593 | -18 |
| TEAM | 7666 | 7644 | 22 | 9242 | 9281 | -39 |

NOTE. Difference (diff.) of continuous (cont.) minus categorical (cat.).
Lower values of AIC indicate the preferred model.

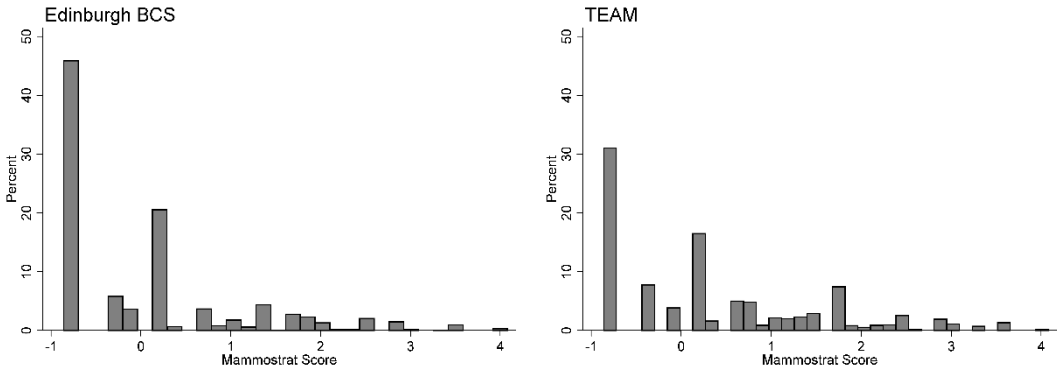



**Figure 6.1** Distribution of Mammostrat Score in the Edinburgh BCS and Team Cohorts.

### *6.2.2.1 Staining Methodology*

Tissue microarray (TMA) is a recent innovation in the field of pathology that overcomes problems with standard histopathological techniques that they are time consuming and labour intensive and therefore costly. It is a robust method of tissue analysis, where a large number of patient samples can be examined in a short time using a minimum number of slides. The method was designed as a high-throughput molecular biology technique for researchers that allows for assessment of expression of interesting candidate disease related genes or gene products simultaneously on hundreds of tissue samples (Jawhar, 2009). It also allows parallel molecular profiling of clinical samples at the DNA, RNA, and protein level. In a TMA, cylinders of tissue are cored out of formalin-fixed, paraffin-embedded tissue blocks and slotted in a regular grid pattern into a blank recipient paraffin wax block. The TMA block is then cut using a standard laboratory microtome (Chandler et al., 2011). The number and size of cores in a TMA block can be varied from approximately forty 2 mm cores to hundreds of 0.6 mm cores. This technique enables pathologists to perform large-scale analyses using immunohistochemistry (IHC), fluorescence in situ hybridization (FISH), or RNA in situ hybridization (ISH) at substantially faster and at markedly lower costs compared with conventional approaches (Skacel et al., 2002, Richani et al., 2006).

IHC refers to the process of detecting antigens (e.g., proteins) in cells of a tissue section by exploiting the principle of antibodies binding specifically to antigens in biological tissues (Ramos-Vara and Miller, 2014). It is widely used for diagnosis of cancers because specific tumour antigens are expressed *de novo* or up-regulated in certain cancers (Duraiyan et al., 2012). FISH is a cytogenetic technique that can be used to visualize specific genes or portions of genes.

IHC staining was performed for a panel of biomarkers including ER, PgR, HER2, Ki67, HTF9C, CEACAM5, NDRG1, p53 and SLC7A5 and FISH for HER2 was performed using either sextuplet (ER, PgR) or triplicate (all other markers) 0.6 $mm^2$ TMA cores. Results were derived from dual scoring by expert observers (as described by Kirkegaard snd colleagues (Kirkegaard et al., 2006)) for the Edinburgh BCS cohort for all markers. For TEAM patients, ER, PgR and Ki67 scores were derived by quantitative image analysis using the Ariol system using algorithms validated against both whole sections and manual assessment (Bartlett et al., 2011a, Faratian et al., 2009). Data for ER were recorded as a histoscore *(*Kirkegaard et al., 2006) and for Ki67 and PgR as percentage positive cells (ATAC and Ki67 guidelines (Dowsett et al., 2011)). Results for HER2 were scored according to the UK guidelines (Bartlett et al., 2011b, Walker et al., 2008) with cases regarded as HER2 amplified if any core showed amplification/overexpression. Positivity for p53, HTF9C (recently re-named TRIMT2A), CEACAM5, NDRG1, and SLC7A5 was recorded as previously described (Ring et al., 2006, Bartlett et al., 2010, Ross et al., 2008, Bartlett et al., 2012).

### 6.2.3 Survival Outcome

The primary end point selected for this study was time to distant recurrence (TTDR) as this is the event which drives subsequent death from breast cancer. TTDR was defined as the interval between operation (Edinburgh BCS) or randomisation (TEAM) and distant metastasis (van de Velde et al., 2011) or death with evidence of recurrent breast cancer with patients censored at the time of last follow-up.

Of the 1,449 (ER-negative and ER-positive) patients in the Edinburgh BCS cohort, there were 273 distant recurrences with a median follow-up of 12.9 years (max 25 years). Of the 6,120 (ER-positive) patients in the TEAM cohort, with a recently updated clinical database, there were 770 distant recurrences with a median follow-up of 6.2 years (max 10 years).

Similar TTDR was observed between the Edinburgh BCS and TEAM cohorts (Figure 6.2A) when including all patients with observed 10 year freedom from distant recurrence estimates of 83.3% (95% CI 81.2-85.2) and 80.5% (95% CI 77.5-83.2) respectively. Better survival was observed in the Edinburgh BCS cohort compared to the TEAM cohort when only ER-positive patients were considered (Figure 6.2B) with 10 year freedom from distant-recurrence estimates of 85.0% (95% CI 82.7-87.0) and 78.8% (95% CI 74.8-82.3) respectively. Longer follow-up was observed for the Edinburgh BCS cohort (Figure 6.2C) with observed 20 year freedom from distant recurrence estimates of 77.1% (95% CI 74.0-79.8) and 78.5% (95% CI 75.1-81.5) for all patients and ER-positive patients respectively.

## 6.3 Patient Demographics

Baseline characteristics of patients in both the Edinburgh BCS and TEAM cohorts are given in Table 6.2. The TEAM cohort is a higher risk population compared with the Edinburgh BCS cohort with higher mean tumour size (24mm vs 16mm respectively), a smaller proportion of low grade tumours (grade 1: 10% vs 27% respectively), higher proportion of node-positive pateints ( 57% vs 23% respectively) and higher proportion of patients given chemotherapy (37% vs. 10% respectively).

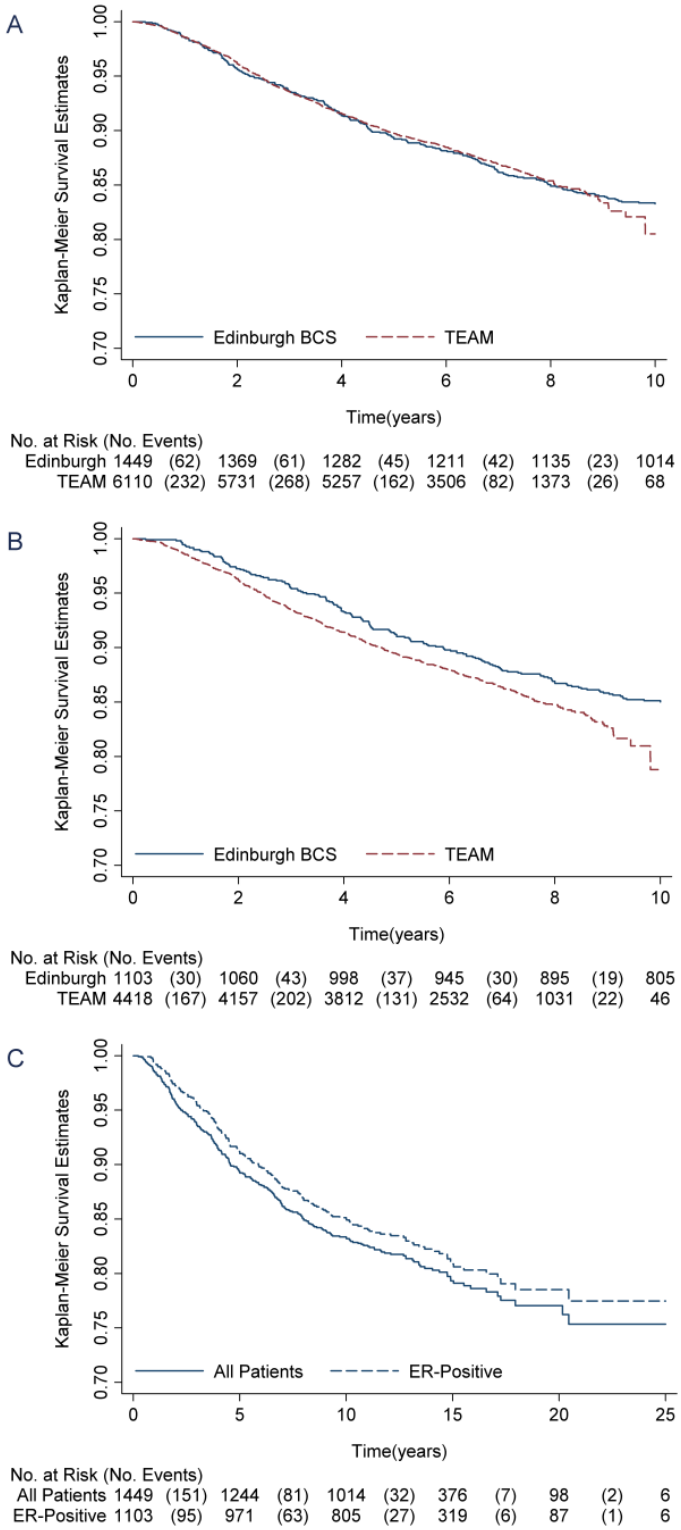



**Figure 6.2** Kaplan-Meier curves for time to distant recurrence. Plots represent (A) All patients in the Edinburgh BCS and TEAM cohorts up to 10 years follow-up, (B) ER-positive patients in the Edinburgh BCS and TEAM cohorts up to 10 years follow-up and (C) All patients and ER-positive patients in the Edinburgh BCS cohort with full follow-up. Note. ER-positive patients defined as having an Allred score above zero and not missing.

**Table 6.2** Baseline Characteristics of Patients

| | Edinburgh BCS (n=1449) | | TEAM (n=6120) | |
|---|---|---|---|---|
| | n (%) | mean (std. dev) or median (Q1-Q3) | n (%) | mean (std. dev) or median (Q1-Q3) |
| Age (Years) | 1449 (100.0) | 56.4 (10.1) | 6120 (100%) | 64.5 (8.9) |
| Tumour size (cm) | 1376 (95.0) | 1.6 (0.7) | 5760 (94.1%) | 2.4 (1.3) |
| Grade | | | | |
| I | 396 (27.3) | - | 616 (10.1) | - |
| II | 664 (45.8) | - | 3168 (51.8) | - |
| III | 352 (24.3) | - | 1835 (30.0) | - |
| Nodal Status | | | | |
| Negative | 1116 (77.0) | - | 2624 (42.9) | - |
| Positive | 333 (23.0) | - | 3496 (57.1) | - |
| | | | | |
| HER2 Status | | | | |
| Negative | 1190 (87.2) | - | 3825 (62.5) | - |
| Positive | 175 (12.8) | - | 560 (9.2) | - |
| ER Status | | | | |
| Negative | 256 (17.7) | - | 41 (0.9) | - |
| Positive | 1103 (76.1) | - | 4422 (99.1) | - |
| ER Histoscore | 1360 (93.9) | 87.5 (4.0-157.1) | 4464 (72.9) | 190.6 (159.6-212.5) |
| PgR Histoscore | 1355 (90.4) | 116.3 (3.5-190.4) | 4401 (71.9) | 138.2 (26.1-197.0) |
| PgR Percent Positive | 984 (67.9) | 63.9 (6.9-97.3) | 4526 (74.0) | 64.0 (13.0-86.6) |
| Mammostrat Score | | | | |
| Low | 673 (46.5) | - | 1530 (25.0) | - |
| Medium | 289 (19.9) | - | 896 (14.6) | - |
| High | 254 (17.5) | - | 1160 (19.0) | - |
| IHC4 Score | | | | |
| Low | 41 (4.4) | - | 281 (6.6) | - |
| Medium | 310 (33.3) | - | 2136 (49.9) | - |
| High | 580 (62.3) | - | 1863 (43.5) | - |
| Treatment | | | | |
| Exe | NA | - | 3075 (50.3) | - |
| Tam & Exe | NA | - | 3045 (49.8) | - |
| Chemotherapy | | | | |
| No | 1300 (89.7) | - | 3863 (63.1) | - |
| Yes | 149 (10.3) | - | 2253 (36.8) | - |

Note. Nodal status defined as negative when number of positive nodes equal zero and positive otherwise. ER Status defined as negative with an Allred score of zero and positive otherwise. Abbreviations: n, number; std. dev, standard deviation; Q, quartile; Exe, exemestane; Tam, tamoxifen; NA, not applicable.

The distributions of the IHC4 and Mammostrat scores were different between the two cohorts, with the median IHC4 scores being 54 and 27 in the Edinburgh BCS and TEAM cohorts respectively (Figure 6.3). A larger proportion of patients (47%) were allocated to the low Mammostrat risk group in the Edinburgh BCS cohort compared with 25% in the TEAM cohort.

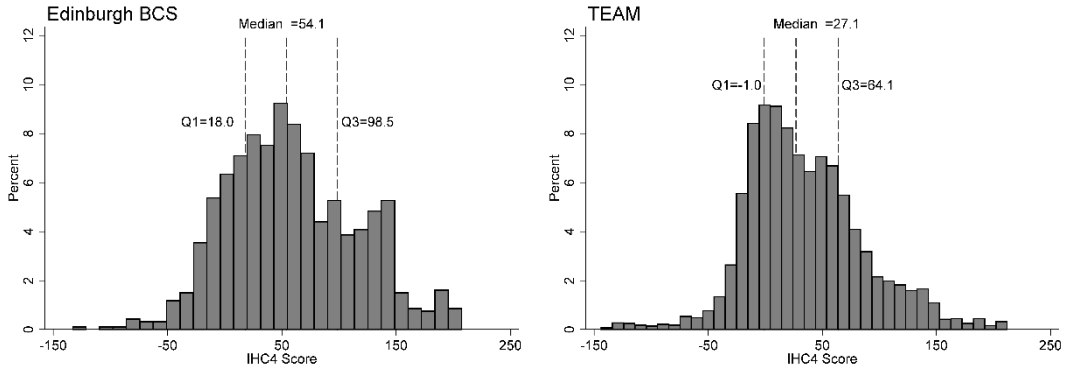



**Figure 6.3** Distribution of IHC4 Score in the Edinburgh BCS and Team Cohorts, displaying quartiles (Q) 1 and 3 and the median.

## 6.4 Missing Data

### 6.4.1 Edinburgh BCS Cohort

The technical limitations of TMAs, such as loss of tissue cores during processing and the unreliability of IHC staining (Avninder et al., 2008), inevitably result in missing data. A large amount of data was missing (32.1%) for the PgR variable measured as percentage of positive cells. Due to correlations between the missingness of TMA data and tumour characteristics such as size, complete case analysis would result in bias (Hoppin et al., 2002). Therefore, multiple imputation was performed.

#### *6.4.1.1 Multiple Imputation*

We used the *mi impute chained* command in Stata to perform multiple imputation using chained equations (MICE) (Biagini et al., 1991) (Biagini et al., 1991) (Biagini et al., 1991) (Biagini et al., 1991) (Biagini et al., 1991) to generate 42 imputed datasets. Multiple imputation was performed using all predictors plus the outcome variable to avoid bias towards the null (White and Royston, 2009). This involves creating multiple copies of the data and imputing the missing values for each dataset with sensible values randomly selected from their predicted distribution. The number of imputed datasets was based on the rule of thumb suggested by White and colleagues that the number of imputations should be at least equal to the percentage of incomplete cases (see Table 6.3)(White et al., 2011). Data were available on the PgR variable measured as histoscores and due to the high correlation between the histoscores and percentage of positive cells, this variable was also included in the imputation model to improve the imputations and hence reduce the standard errors of the estimates in the analysis. The event indicator and the Nelson-Aalen estimator (see section 3.4) of the cumulative baseline hazard were included in the imputation model as recommended by White & Royston (White and Royston, 2009). The results from analyses on each of the imputed datasets were combined using Rubin's rules (Rubin, 1987) which involves incorporating both the within and between imputation variability to produce estimates and confidence intervals that incorporate the uncertainty of imputed values. Suppose that $\hat{Q}_j$ is an estimate of a scalar quantity of interest (e.g. a regression coefficient) obtained from data set j (j=1,2,..,m) and $U_j$ is the standard error associated with $\hat{Q}_j$. The overall estimate is the average of the individual estimates,

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m} \hat{Q}_j \quad . \qquad 6.1.$$

The total variance of $\bar{Q}$ is formed from the within-imputation variance $\overline{U} = \frac{1}{m}\sum_{j=1}^{m} U_j$ and the between-imputation variance $B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{Q}_j - \bar{Q}\right)^2$:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B \quad . \qquad 6.2.$$

**Table 6.3** Information on Completeness of Data

| No. Risk Factors* not recorded (per patient) | No. (%) of All Patients (n=1,449) |
|---|---|
| 0 (complete data) | 844 (58.2) |
| 1 | 363 (25.1) |
| 2 | 83 (5.7) |
| 3 | 23 (1.6) |
| 4 | 24 (1.7) |
| 5 | 19 (1.3) |
| 6 | 19 (1.3) |
| 7 | 17 (1.2) |
| 8 | 18 (1.2) |
| 9 | 33 (2.3) |
| 10 | 6 (0.4) |

* Risk factors are ER, PgR, HER2, Ki67, SLC7A5, p53, NDRG1, HTF9C, CEACAM5, age, grade, size and nodal status.

*Statistics that cannot be combined using Rubin's rules*

Regression coefficients can be combined easily across imputations using Rubin's rules, however some statistics cannot and others require a transformation to ensure that they are approximately normally distributed. Provided by White and colleagues, Table 6.4 summarizes some common statistics which can and cannot be combined using Rubin's rules (White et al., 2011).

**Table 6.4** Statistics that can and connot be combined using Rubin's rules

| | |
|---|---|
| Statistics that can be combined without any transformation | Mean, proportion, regression coefficient, linear predictor, C-index, area under the ROC curve |
| Statistics that may require sensible transformation before combination | Odds ratio, hazard ratio, baseline hazard, survival probability, standard deviation, correlation, proportion of variance explained, skewness, kurtosis |
| Statistics that cannot be combined | P-value, likelihood ratio test statistic, model chi-squared statistic, goodness-of-fit test statistic |

### *6.4.1.2 Distribution of Continuous Variables*

The first step is to assess the distribution of the continuous variables as an approximate normal distribution is required for imputation models. ER, PgR and Ki67 showed departures from normality (Figure 6.4A), however even after a shifted log transformation (Figure 6.4B) ER and PgR still did not satisfy the normality assumption due to the large number of zero values (and values of 100 for PgR measured as % positive cells).

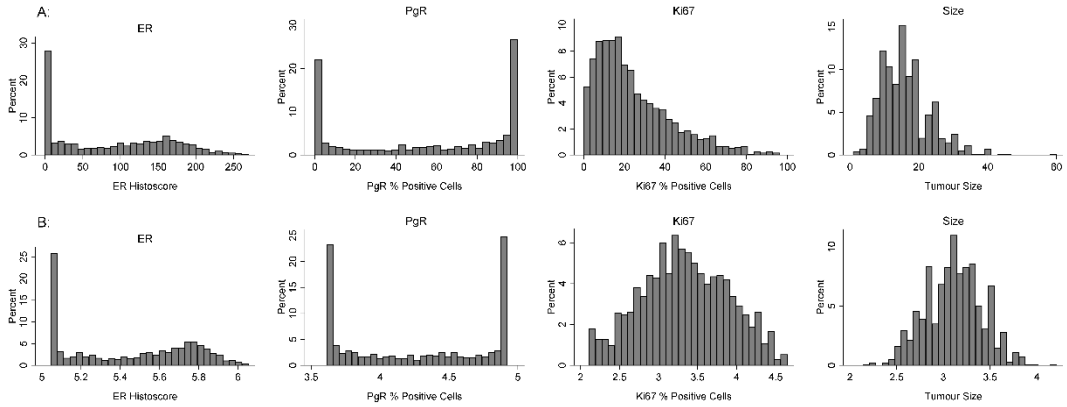



**Figure 6.4** Distribution of the continuous covariates. Plots represent (A) Distribution of the continuous covariates with missing data and (B) Distribution of the transformed continuous covariates.

The essence of the MICE algorithm is regression of an incomplete variable on other variables (Royston and White, 2011). A summary of the variables and their associated regression command is given in Table 6.5. The default method is logistic regression when there are two distinct values of the variable to be imputed, multinomial or ordinal logistic regression when there are 3-5 values and linear regression otherwise. To avoid values being predicted outside the possible range of values, truncated regression was used. Due to the problems with non-

normal distributions of ER and PgR even after transformation, predictive mean matching (PMM) was also considered as the imputation method. PMM imputes missing values of a continuous variable z such that imputed values are sampled only from the observed values of z by matching predicted values as closely as possible. The resulting distribution of imputed z often closely matches that of observed z.

**Table 6.5** Summary of variables in the imputation model

| Variable | Type | Levels/Range (min-max) | % missing | Imputation Method |
|---|---|---|---|---|
| ER Histoscore | Continuous | 0-300 | 6.1 | truncated regression/pmm |
| PgR Histoscore | Continuous | 0-300 | 6.5 | truncated regression/pmm |
| PgR % Pos. Cells | Continuous | 0-100 | 32.1 | pmm |
| HER2 | Binary | 2 | 5.8 | logistic |
| Ki67 % Pos Cells | Continuous | 0-100 | 7.7 | truncated regression |
| SLC7A5 | Binary | 2 | 8.8 | logistic |
| CEACAM5 | Binary | 2 | 9.2 | logistic |
| NDRG1 | Binary | 2 | 8.8 | logistic |
| HTF9C | Binary | 2 | 8.7 | logistic |
| p53 | Binary | 2 | 10.3 | logistic |
| Age | Continuous | 24-91 | 0 | NA |
| Tumour Size (mm) | Continuous | 1-60 | 5 | linear regression |
| Tumour Grade | Ordinal | 3 | 0.8 | ordinal |
| No. Positive Nodes | Continuous | 0-28 | 0 | NA |
| Survival Outcomes included in imputation model | | | | |
| Event Indicator | Event/Censored | | 0 | NA |
| Nelson-Aalen estimator of the cumulative baseline hazard | | | 0 | NA |

### *6.4.1.3 Non-Linear Terms*

Another issue to consider is non-linear covariate effects as they cannot be correctly assessed if they have not been included in the imputation model.

A starting point to determine the functional form of the continuous variables involves the use of Martingale residuals defined in section 3.5. Plotting the martingale residuals from the null model against the continuous covariates can give an indication of the functional form of the

covariate, a linear effect when the fitted smooth curve is flat. The plots (Figure 6.5) indicate a strong non-linear functional form for tumour size and number of positive nodes as well as a slight non-linear functional form for age.

Fractional polynomials were then used to determine the transformation which best modelled the non-linear effect in a complete case analysis, based on the model with the lowest deviance (defined as minus twice the log likelihood). The addition of non-linear effects significantly improved the fit of the model for tumour size and number of positive nodes, with the effects on the log-hazard displayed in Figure 6.6.

The transformed tumour size and number of positive nodes were included in the imputation model. If any convincing non-linear terms were found in the analysis of the imputed data, the imputations would be re-created including the non-linear terms. However, no further non-linear terms were identified.

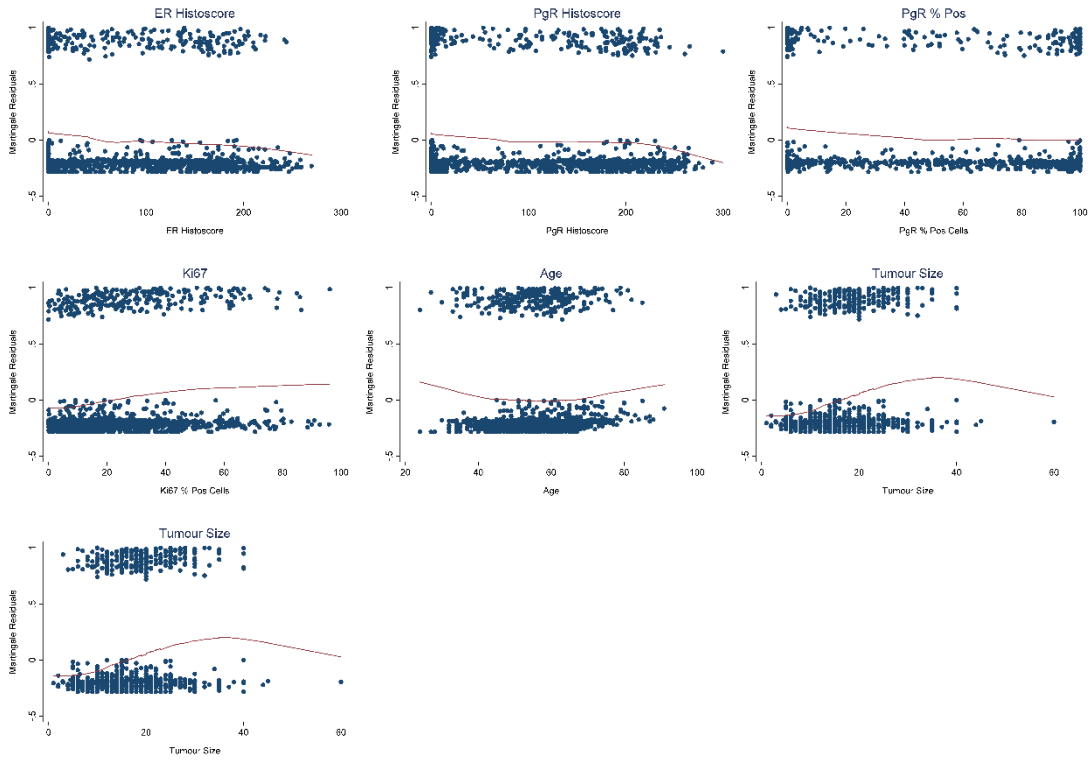



**Figure 6.5** Plots of martingale residuals versus continuous covariates. The red line is a smoothed curve produced by locally weighted regression.

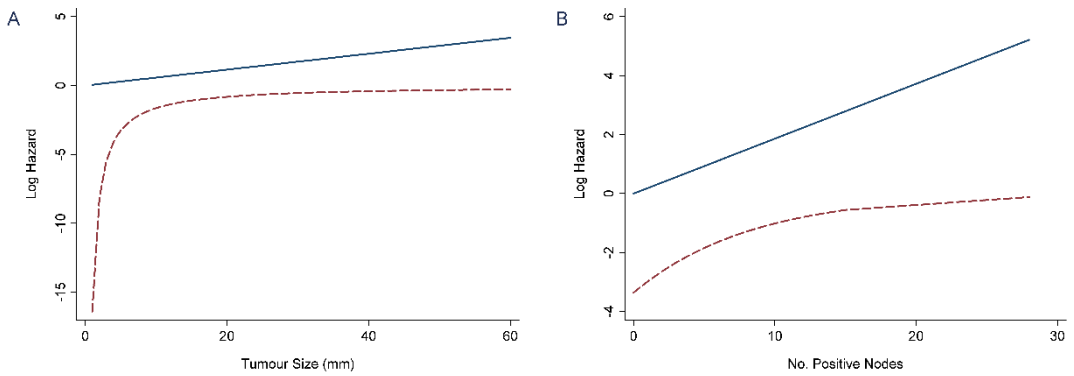



**Figure 6.6** Comparison of linear and non-linear effects. Plots represent linear (solid line) and non-linear (dashed line) effects of (A) tumour size and (B) number of positive nodes on the log-hazard.

#### *6.4.1.4 Comparison between observed and imputed values*

We compared the observed and imputed data for continuous variables using box plots over the imputations (Figure 6.7). For variables satisfying the normality assumption (transformed Ki67 and tumour size), similar distributions were observed between observed and imputed values. ER and PgR histoscores have skewed distributions, with a large number of zero values and using truncated regression has resulted in differences in the distribution between the observed and imputed values. The imputation of ER and PgR histoscores was also considered using PMM (Figure 6.8). This resulted in more similar distributions between the observed and imputed values but still some variation in the median values, especially for PgR histoscores. Some differences are to be expected however, as the data are assumed to be missing at random (MAR) and therefore any systematic difference between the missing values and the observed values can be explained by differences in the observed data (Sterne et al., 2009).

For categorical variables, we compared the observed and imputed data using frequency tables (Figure 6.9). In the majority of imputations a smaller number of patients were predicted to be biomarker positive than in the observed data, however quite large variation was observed.

When performing a univariate Cox regression analysis of IHC4 score on TTDR, similar estimated coefficients were observed between a complete case analysis and analysis on imputed data (Table 6.6). The imputed analysis had a reduced standard error (0.22 versus 0.19 respectively) as expected due in part to the larger sample size. A very small difference was observed in the estimated coefficient when imputing IHC4 using regression for ER histoscores compared to PMM (0.52 versus 0.53), with PMM giving a closer estimate to that of a complete case analysis. A larger difference in the estimated coefficient for Mammostrat score from a complete case analysis compared to an analysis on imputed data was observed (0.90 versus 0.97). However the standard error was not reduced, possibly due to the large variation between imputations seen for the categorical variables used in the calculation of Mammostrat score.

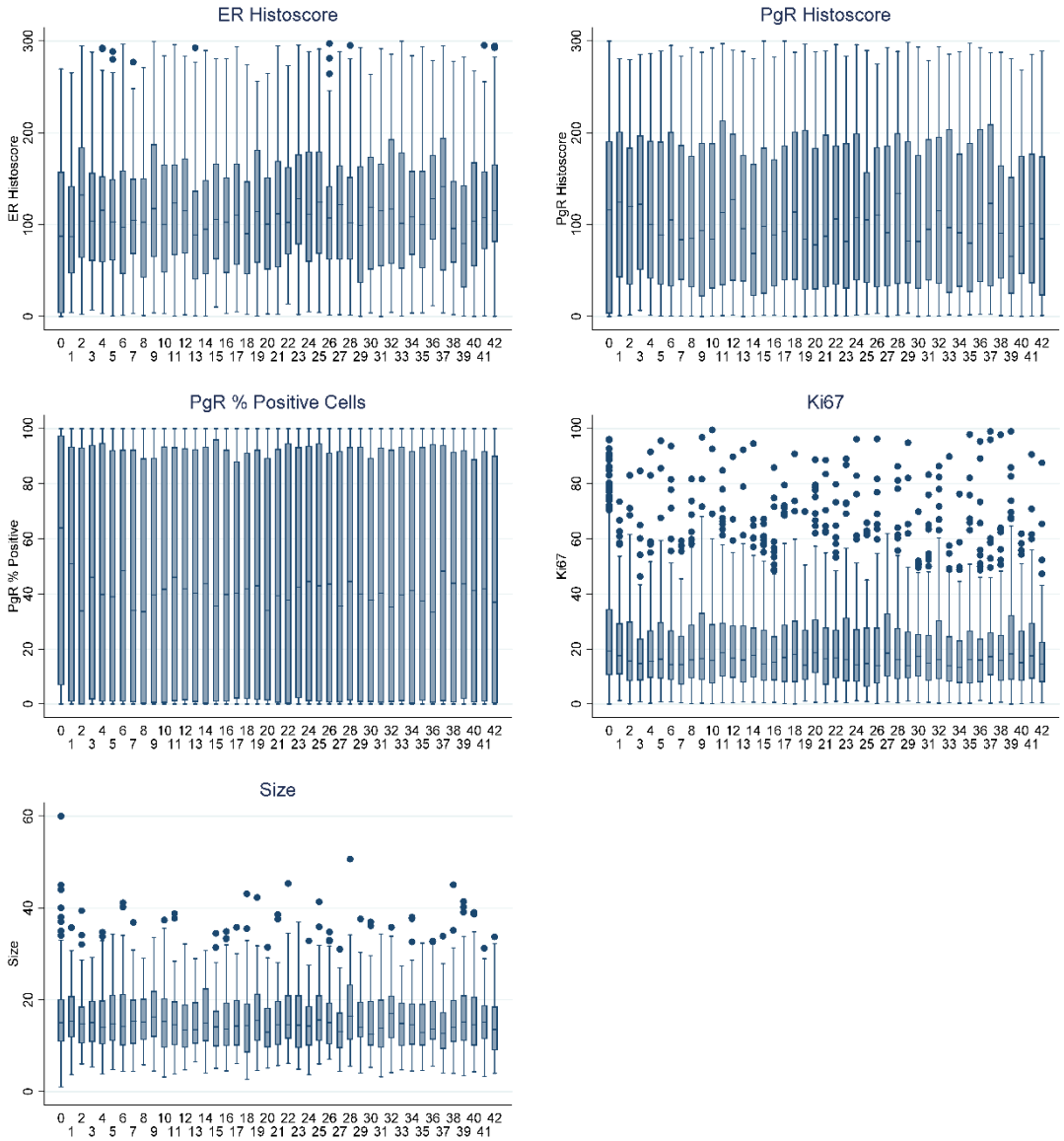



**Figure 6.7** Comparison of Observed (0) and Imputed (1-42) data for continuous variables. Imputation methods: regression for ER and PgR histoscores, Ki67 and tumour size. Predictive mean matching for PgR % positive cells.

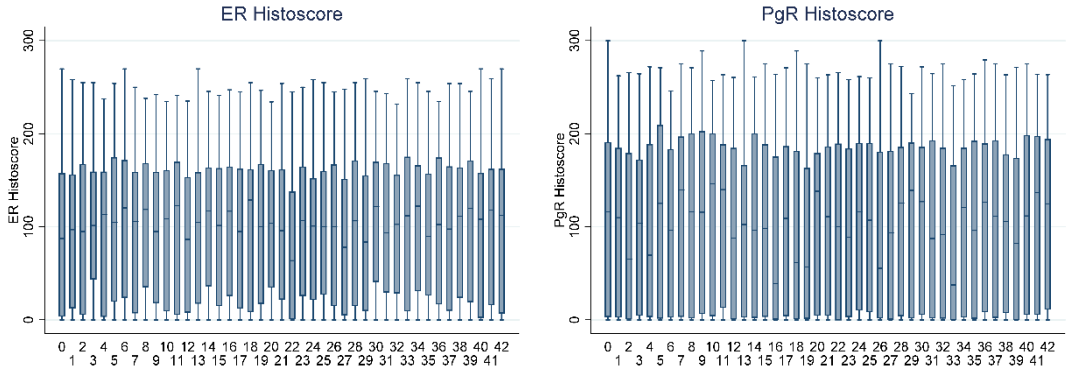



**Figure 6.8** Comparison of observed (0) versus imputed (1-42) data using predictive mean matching (PMM) for imputation of ER and PgR histoscores.

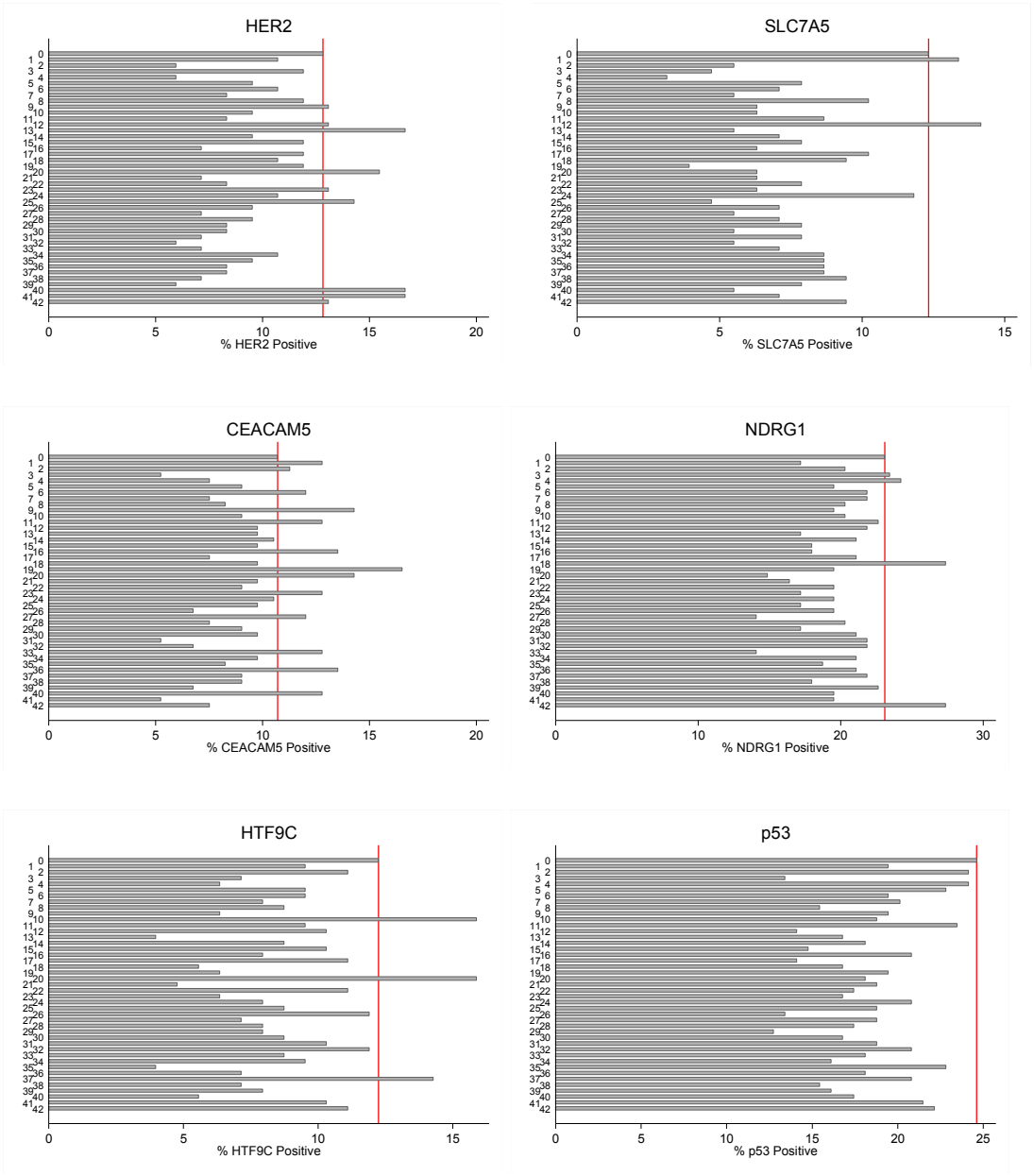



**Figure 6.9** Comparison of observed (0) versus imputed (1-42) data for categorical variables. Vertical (red) line indicates the observed percentage of positive values.

**Table 6.6** Estimated coefficients from a univariate Cox regression of IHC4 and Mammostrat score from a complete case analysis and different imputation procedures.

| | IHC4 | | Mammostrat | |
|---|---|---|---|---|
| Imputation Procedure | Estimate | Standard Error | Estimate | Standard Error |
| Complete Case | 0.54 | 0.22 | 0.90 | 0.15 |
| Regression | 0.52 | 0.19 | 0.97 | 0.15 |
| PMM | 0.53 | 0.19 | 0.96 | 0.15 |

Note. Estimates represent coefficients calculated between the IQR of the continuous IHC4 score and high risk compared to low risk as categorised by the Mammostrat score. Abbreviations: IQR, interquartile range; PMM, predictive mean matching.

### 6.4.2 TEAM Cohort

Five of nine countries that participated in the TEAM trial (6,120 randomised patients) provided tumour samples for pathology sub-studies. Tissue blocks were received at a central laboratory and 4,598 were suitable for TMA construction. Patient tumour characteristics were similar among samples from the analysed pathology subset and in all patients from countries participating in the pathology substudy (Bartlett et al., 2011a). Multiple imputation was not performed on the TEAM cohort due to the large number of patients still available for analysis and the difficulties with some patients having no biomarker data at all.

## 6.5 Software

All analyses were performed using STATA IC version 12 (StataCorp, 2011).

# Chapter 7: Analysis of IHC4 and Mammostrat

## 7.1 Introduction

In this chapter we compare the two prognostic IHC biomarker panels, IHC4 and Mammostrat, to determine which provides more information on the risk of recurrence in the context of additional clinical factors or whether combining both approaches would increase the information available to patients and clinicians.

This chapter will focus on fitting the standard Cox proportional hazards model and assessing the validity of the model assumptions.

## 7.2 Methods

### 7.2.1 Materials

The Edinburgh BCS and TEAM cohorts will be used as described in chapter 6.

#### *7.2.1.1 Clinical subgroups of interest*

The primary analysis was based on ER-positive patients treated with adjuvant endocrine therapy only (without chemotherapy) (Figure 7.1). Secondary analysis was performed on all ER-positive patients irrespective of treatment. Exploratory analyses were performed on node-negative and node-positive ER-positive patients treated with adjuvant endocrine therapy (without chemotherapy) and node-positive ER-positive patients irrespective of treatment. Kaplan-Meier curves for the clinical subgroups are displayed in Figure 7.2.

Baseline characteristics for all ER-positive patients are given in Table 7.1.

.

A

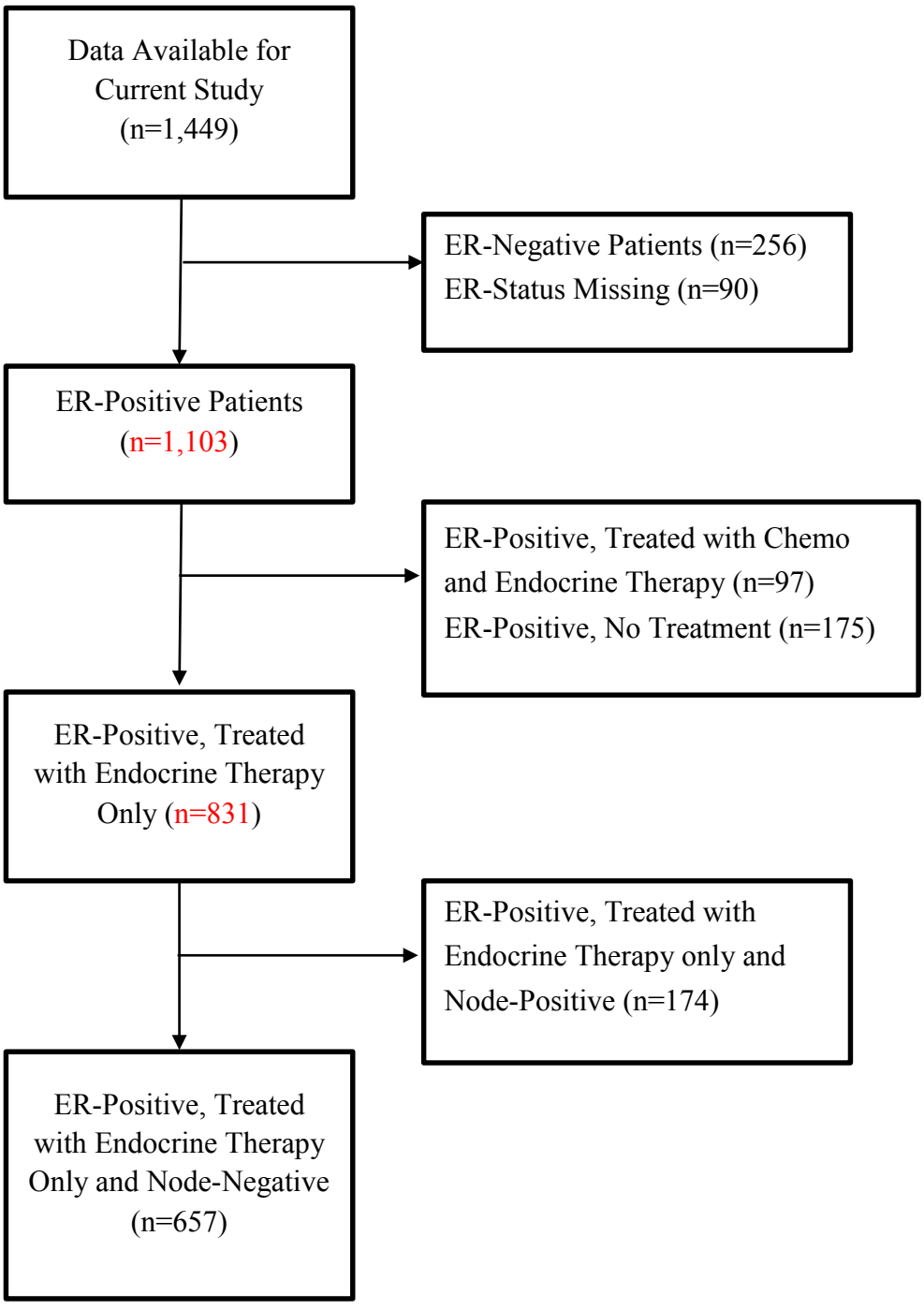



**Figure 7.1** Clinical subgroups of interest for (A) Edinburgh BCS cohort.

**B**

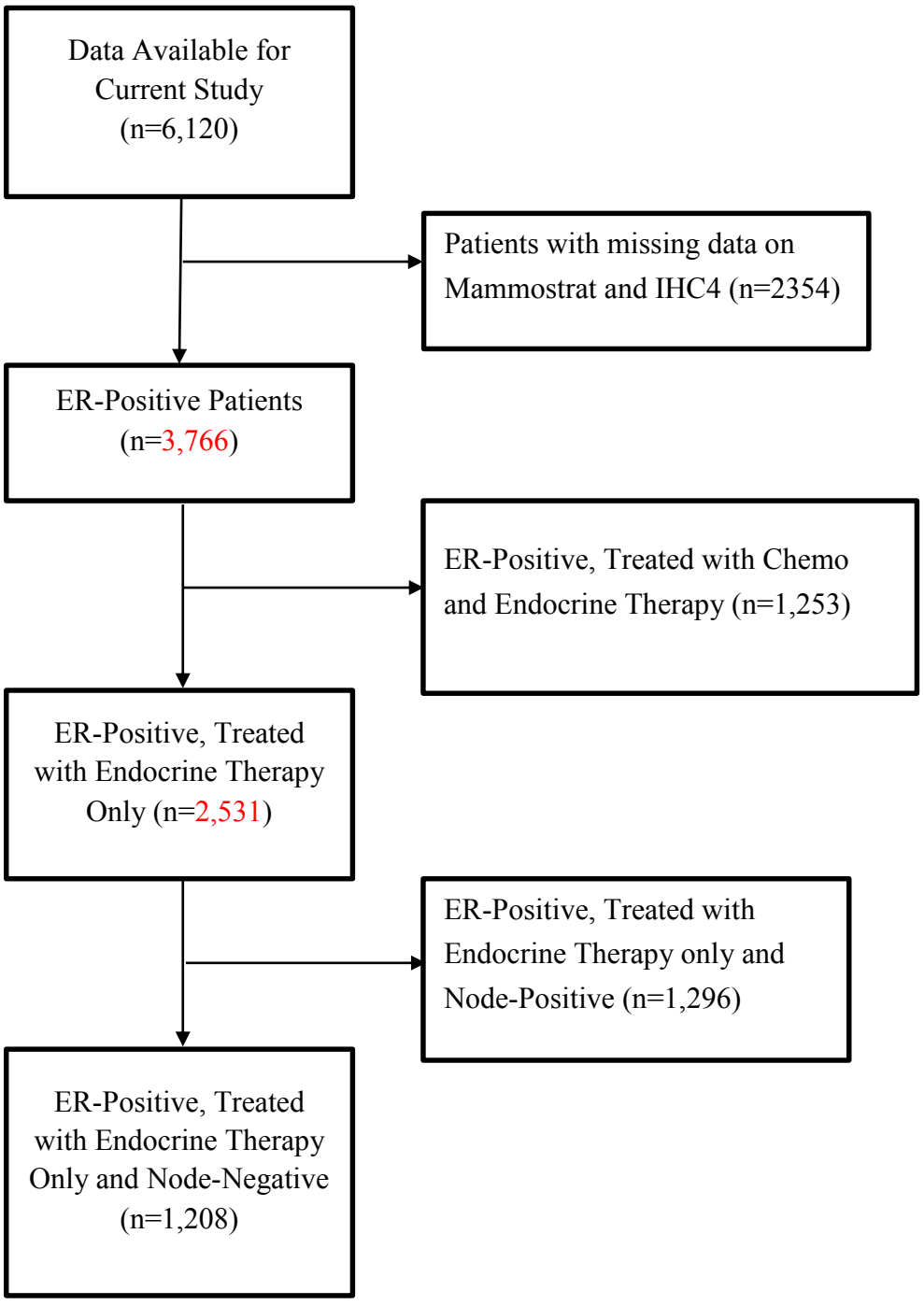
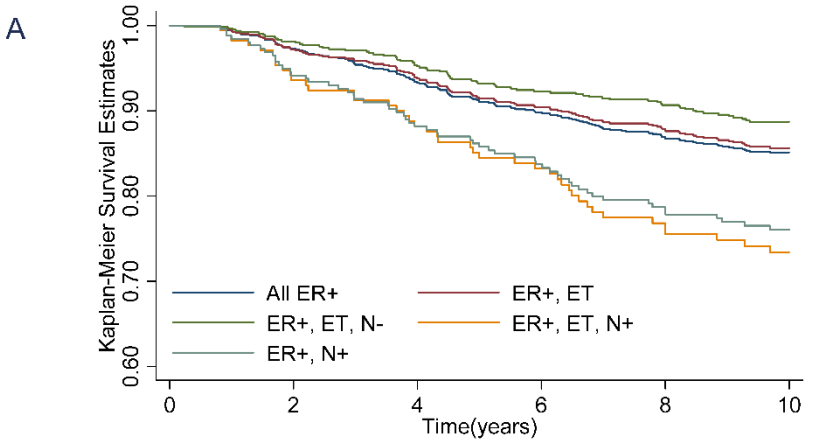
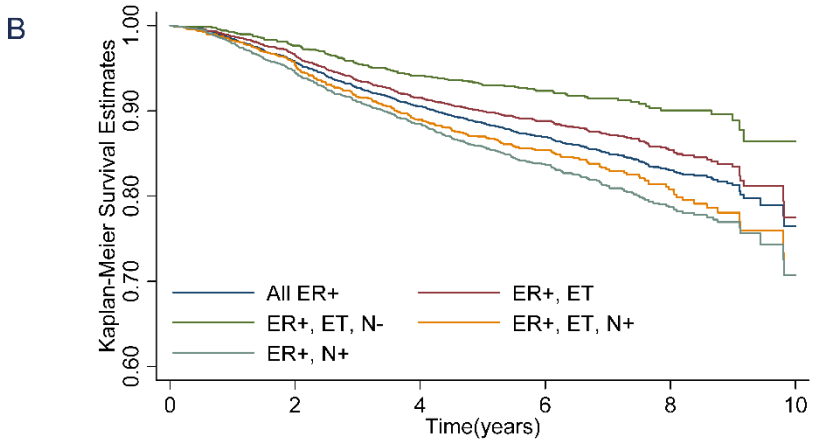



**Figure 7.1** Clinical subgroups of interest for (B) the TEAM cohort.




No. at Risk (No. Events)

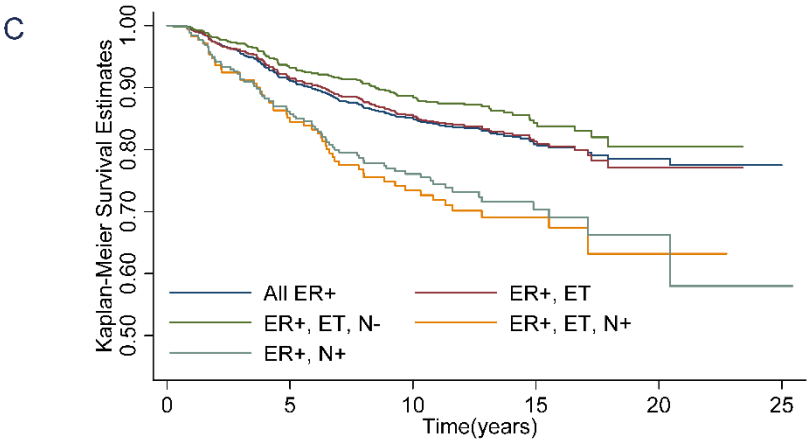
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| All ER+ | 1103 | (30) | 1060 | (43) | 998 | (37) | 943 | (31) | 893 | (17) | 803 |
| ER+, ET | 831 | (23) | 801 | (28) | 758 | (27) | 718 | (21) | 678 | (16) | 614 |
| ER+, ET, N- | 657 | (12) | 641 | (19) | 614 | (19) | 586 | (10) | 561 | (12) | 514 |
| ER+, ET, N+ | 174 | (11) | 160 | (9) | 144 | (8) | 132 | (11) | 117 | (4) | 100 |
| ER+, N+ | 260 | (15) | 241 | (15) | 219 | (11) | 202 | (13) | 185 | (5) | 153 |




No. at Risk (No. Events)

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| All ER+ | 3766 | (159) | 3530 | (188) | 3224 | (117) | 2139 | (65) | 890 | (19) | 39 |
| ER+, ET | 2513 | (87) | 2367 | (120) | 2148 | (57) | 1438 | (37) | 593 | (15) | 25 |
| ER+, ET, N- | 1208 | (28) | 1158 | (41) | 1078 | (19) | 760 | (12) | 313 | (4) | 7 |
| ER+, ET, N+ | 1296 | (59) | 1201 | (79) | 1062 | (38) | 672 | (25) | 278 | (11) | 18 |
| ER+, N+ | 2229 | (121) | 2059 | (130) | 1855 | (90) | 1186 | (49) | 492 | (13) | 26 |

**Figure 7.2** Kaplan-Meier curves for time to distant recurrence for clinical subgroups. Plots represent (A) Edinburgh BCS with 10 year follow-up, (B) TEAM cohort with 10 year follow-up and (C) Edinburgh BCS cohort with full follow-up. Groups represent: All ER-positive patients (ER+); ER-positive patients treated with endocrine therapy only (ER+, ET); ER-positive patients treated with endocrine therapy only and either node-negative (ER+, ET, N-) and node-positive (ER+, ET, N+); and ER-positive, node-positive patients irrespective of treatment (ER+, N+).




No. at Risk (No. Events)

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| All ER+ | 1103 | (97) | 969 | (61) | 803 | (27) | 319 | (6) | 86 | (1) | 6 |
| ER+, ET | 831 | (70) | 735 | (45) | 614 | (21) | 228 | (6) | 22 | (0) | 0 |
| ER+, ET, N- | 657 | (44) | 599 | (28) | 514 | (16) | 179 | (4) | 18 | (0) | 0 |
| ER+, ET, N+ | 174 | (26) | 136 | (17) | 100 | (5) | 49 | (2) | 4 | (0) | 0 |
| ER+, N+ | 260 | (36) | 211 | (23) | 153 | (8) | 60 | (2) | 9 | (1) | 1 |

**Figure 7.2 (Continued)** Kaplan-Meier curves for time to distant recurrence for clinical subgroups. Plots represent (A) Edinburgh BCS with 10 year follow-up, (B) TEAM cohort with 10 year follow-up and (C) Edinburgh BCS cohort with full follow-up. Groups represent: All ER-positive patients (ER+); ER-positive patients treated with endocrine therapy only (ER+, ET); ER-positive patients treated with endocrine therapy only and either node-negative (ER+, ET, N-) and node-positive (ER+, ET, N+); and ER-positive, node-positive patients irrespective of treatment (ER+, N+).

**Table 7.1** Baseline characteristics of patients in the analysis set

| | ER-Positive Patients | | | |
|---|---|---|---|---|
| | Edinburgh BCS (n=1103) | | TEAM (n=3766) | |
| | n (%) | mean (std. dev) or median (Q1-Q3) | n (%) | mean (std. dev) or median (Q1-Q3) |
| Age (Years) | 1103 (100.0) | 56.5 (9.9) | 3766 (100.0) | 65.1 (9.0) |
| Tumour size (cm) | 1103 (100) | 1.6 (0.7) | 3657 (97.1) | 2.4 (1.3) |
| Grade | | | | |
| I | 354 (32.1) | - | 391 (10.4) | - |
| II | 542 (49.1) | - | 1901 (50.5) | - |
| III | 207 (18.8) | - | 1249 (33.2) | - |
| Nodal Status | | | | |
| Negative | 843 (76.4) | - | 1528 (40.6) | - |
| Positive | 260 (23.6) | - | 2229 (59.2) | - |
| | | | | |
| HER2 Status | | | | |
| Negative | 933 (84.6) | - | 3259 (86.5) | - |
| Positive | 170 (15.4) | - | 507 (13.5) | - |
| ER Status | | | | |
| Negative | 0 (0.0) | - | 38 (1.0) | - |
| Positive | 1103 (100.0) | - | 3728 (99.0) | - |
| ER Histoscore | 1103 (100.0) | 117.1 (37.9-164.6) | 3766 (100.0) | 191.2 (161.4-213.1) |
| PgR Histoscore | 1103 (100.0) | 140.0 (23.2-195.7) | 3766 (100.0) | 138.6 (22.9-197.3) |
| PgR Percent Positive | 1103 (100.0) | 56.1 (4.2-96.4) | 3766 (100.0) | 65.7 (12.1-87.5) |
| Mammostrat Score | | | | |
| Low | 704 (63.8) | - | 1606 (42.6) | - |
| Medium | 246 (22.4) | - | 942 (25.0) | - |
| High | 152 (13.8) | - | 1218 (32.3) | - |
| IHC4 Score | 1103 (100.0) | 44 .4 (12.1-78.1) | 3766 (100.0) | 26.9 (-0.7-64.7) |
| Low | 63 (5.7) | | 239 (6.4) | |
| Medium | 424 (38.4) | | 1889 (50.2) | |
| High | 616 (55.9) | | 1638 (43.5) | |
| Treatment | | | | |
| Exe | NA | - | 1889 (50.2) | - |
| Tam & Exe | NA | - | 1877 (49.8) | - |
| Chemotherapy | | | | |
| No | 1006 (91.2) | - | 2512 (66.7) | - |
| Yes | 97 (8.8) | - | 1253 (33.3) | - |

Note. Nodal status defined as negative when number of positive nodes equal zero and positive otherwise. ER Status defined as negative with an Allred score of zero and positive otherwise. TEAM cohort refers to the subset with IHC4 and Mammostrat data. Abbreviations: n, number; std. dev, standard deviation; Q, quartile; Exe, exemestane; Tam, tamoxifen; NA, not applicable.

### 7.2.2 Statistical Analysis

The primary end point was time to distant recurrence (TTDR) as this is the event which drives subsequent death from breast cancer. TTDR was defined as the time to distant metastasis (van de Velde et al., 2011) or death with evidence of recurrent breast cancer with patients censored at the time of last follow-up.

Firstly the survival curves were computed for the scores using the Kaplan-Meier product limit estimator.

Cox proportional hazards regression was performed and the added value of the addition of either IHC4 or Mammostrat scores or both to clinical factors was assessed using Wald tests.

The functional form of the continuous covariates on the log-hazard was investigated using the multivariable fractional polynomial (MFP) algorithm (Royston and Altman, 1994), step one in the MFPT approach described in section 4.4.4, using the Stata *mfp* command, which selects the fractional polynomial (FP) model that best predicts the outcome variable. Assessment of the proportional hazards model was assessed using Schoenfeld residuals and the Grambsch-Therneau test.

The predictive performance of the IHC4 and Mammostrat risk scores along with conventional clinical risk factors was assessed using measures of calibration and discrimination, two fundamental measures of evaluating model performance as stated by Royston and Altman (Royston and Altman, 2013).

Model calibration refers to how closely the estimates of survival from the model agree with the survival from the observed data (Altman et al., 2009, Moons et al., 2009). This was assessed for each decile of predicted risk, ensuring 10 equally sized groups, by producing a calibration plot (observed versus predicted probabilities of 5 year distant recurrence) and

calculating the calibration slope. An estimate of the baseline hazard function from the original study is not available and therefore a strict assessment of calibration is not possible.

Discrimination is the ability of a risk score to differentiate between patients who do and do not experience an event during the study period (McGeechan et al., 2008, Altman et al., 2009). However the degree of separation is hard to assess, with several measures available but not one seems to be universally accepted. Some recent descriptions and comparisons of measures are given by Choodari-Oskooei and Hielscher (Hielscher et al., 2010, Choodari-Oskooei et al., 2012). We are going to focus on Royston & Sauerbrei's $R^2$ statistic based on their index of discrimination (D)(Royston and Sauerbrei, 2004). The D statistic measures prognostic separation of survival curves, and is closely related to the standard deviation of the prognostic index (or risk score). It is computed by ordering the prognostic index (PI) across patients, calculating the rankits (expected standard normal order statistics) corresponding to these values, dividing the latter by a factor $k = \sqrt{8/\pi} \approx 1.596$ and performing Cox regression on the scaled rankits. The resulting regression coefficient is D. The conversion to $R_D^2$ is given by

$$R_D^2 = \frac{D^2/k^2}{\sigma^2 + D^2/k^2} \quad , \qquad 7.1.$$

where $\sigma^2 = \pi^2/6 \approx 1.645$. A difference in D of at least 0.1 may be needed to see any important difference in separation between the relevant survival curves.

## 7.3 Results

Data were available on 1,103 ER-positive patients with 192 distant recurrences from the Edinburgh BCS cohort (median follow-up of 12.2 years) and on 3,766 ER-positive patients with 548 distant recurrences from the TEAM cohort (median follow-up of 6.2 years). See Table 7.2 for other subgroups of interest.

**Table 7.2** Information on number of patients, number of events and median survival time for patient subgroups

| | Edinburgh BCS | | | TEAM | | |
|---|---|---|---|---|---|---|
| Patient Subgroup | N | No. Events | Median Survival Time (years) | N | No. Events | Median Survival Time (years) |
| All ER-Positive | 1,103 | 192 | 12.2 | 3,766 | 548 | 6.2 |
| ER-Positive treated with Endocrine Therapy Only | 831 | 142 | 12.2 | 2,513 | 316 | 6.2 |
| ER-Positive, Endocrine Only and Node-Negative | 657 | 92 | 12.6 | 1,208 | 104 | 6.4 |
| ER-Positive, Endocrine Only and Node-Positive | 174 | 50 | 11.0 | 1,296 | 212 | 6.0 |
| ER-Positive and Node-Positive | 260 | 70 | 11.0 | 2,229 | 402 | 6.1 |

### 7.3.1 Kaplan-Meier Estimates of TTDR

Lower scores for IHC4 and Mammostrat are associated with better TTDR survival than higher scores in both the Edinburgh BCS and TEAM cohorts for ER-positive patients treated with endocrine therapy only and all ER-positive patients irrespective of treatment (Figure 7.3) as well as for the exploratory subgroups (Figure 7.4). Less separation of the survival curves were seen for those categorised as medium and high risk according to the Mammostrat score in the Edinburgh BCS cohort. The cut-points from the original study of IHC4 did not validate well, allocating a small proportion of patients to low risk in both cohorts (5.7% and 6.3% of all ER-positive patients in the Edinburgh BCS and TEAM cohorts respectively). The 5-year TTDR estimates were similar between studies and 15 year TTDR estimates were only available for the Edinburgh BCS cohort (Table 7.3).

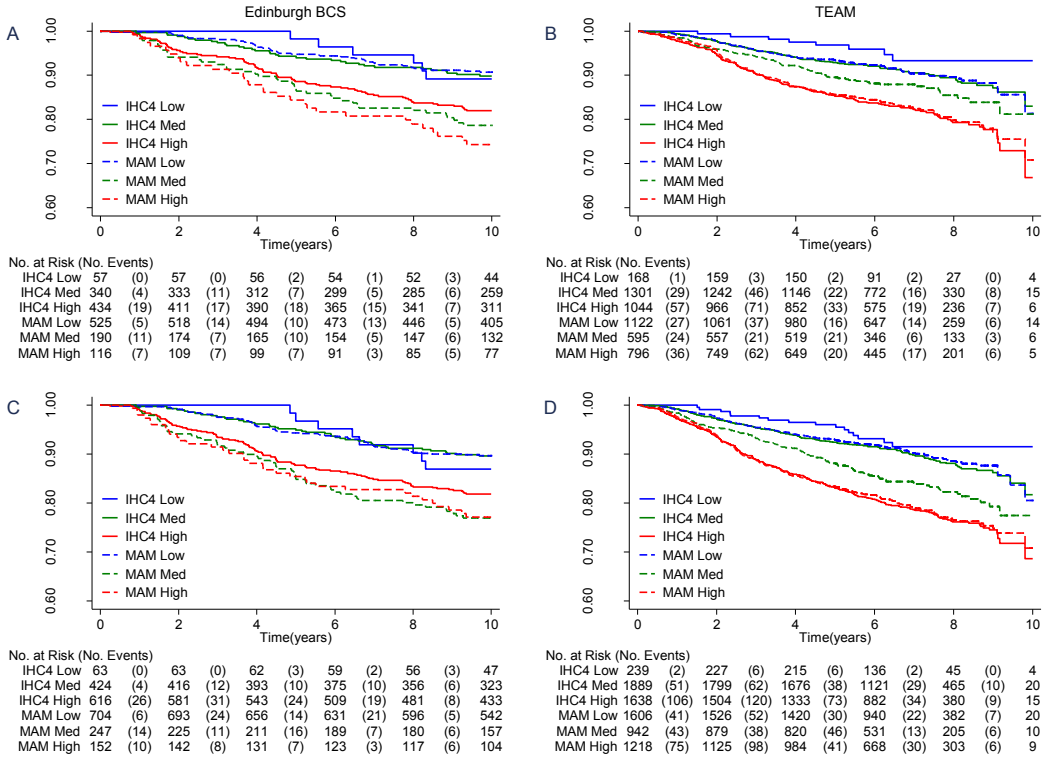



**Figure 7.3** Kaplan-Meier curves for time-to distant recurrence up to 10 years for the IHC4 score (solid lines) and Mammostrat score (dashed lines). Plots represent ER-positive patients treated with endocrine therapy only (A and B) and all ER-positive patients (C and D) in the Edinburgh BCS (left column) and TEAM (right column) cohorts.

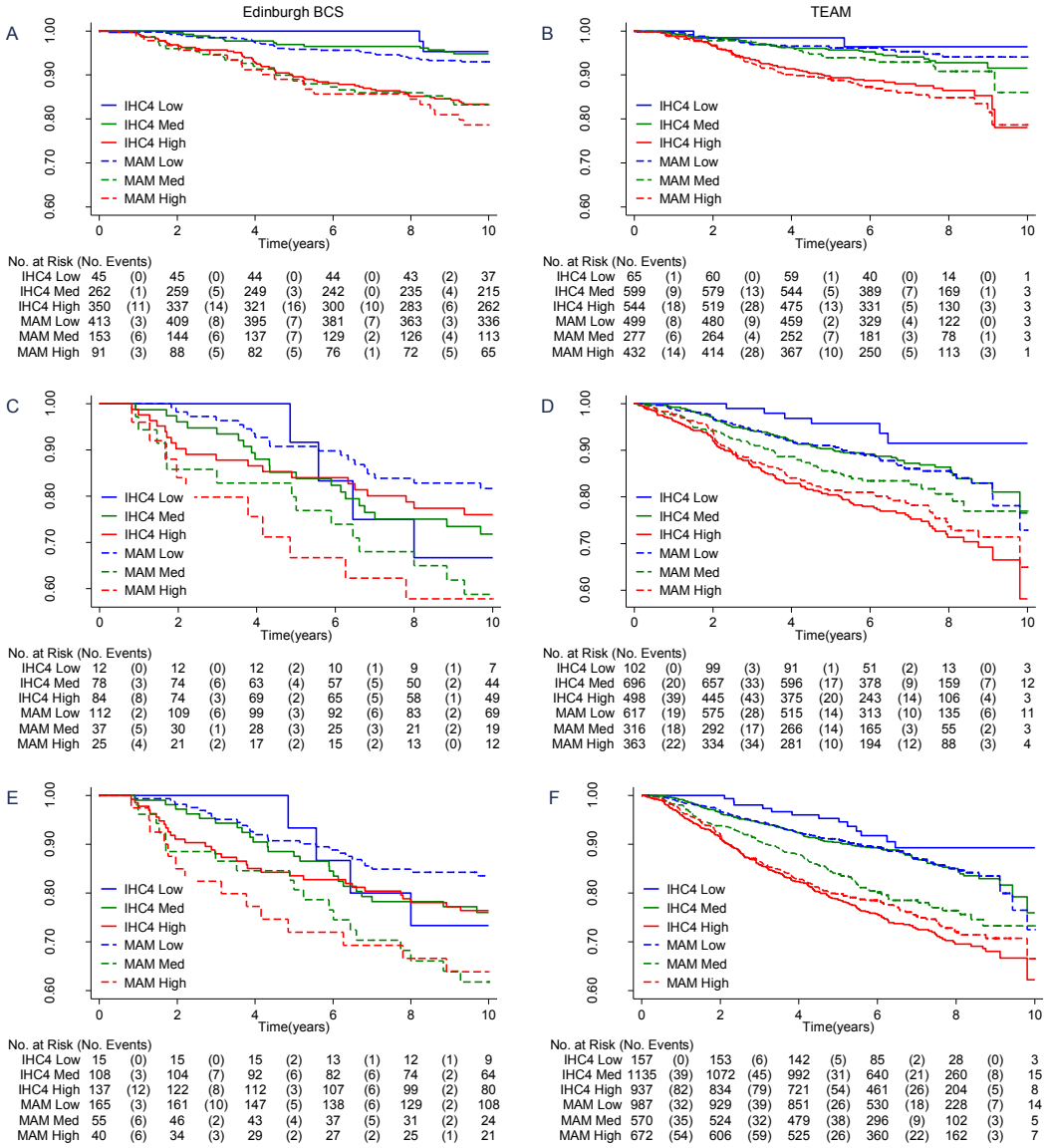



**Figure 7.4** Kaplan-Meier curves for time-to distant recurrence (exploratory subgroups) up to 10 years for the IHC4 score (solid lines) and Mammostrat score (dashed lines). Plots represent ER-positive patients treated with endocrine therapy only and either node-negative (A and B) or node-positive (C and D) and ER-positive, node-positive patients irrespective of treatment (E and F) in the Edinburgh BCS (left column) and TEAM (right column) cohorts.

**Table 7.3** Five and Fifteen-year freedom from distant recurrence estimates stratified by IHC4 and Mammostrat scores

| | All ER-Positive | | ER-Positive Endocrine Only[a] | |
|---|---|---|---|---|
| | Edinburgh BCS (n=1,103) | TEAM (n=3,766) | Edinburgh BCS (n=831) | TEAM (n=2,513) |
| **5-Year** | | | | |
| IHC4 L | 96.7 (87.7-99.2) | 96.0 (92.5-97.9) | 98.2 (88.0-99.8) | 96.8 (92.5-98.7) |
| IHC4 M | 94.9 (92.3-96.7) | 92.4 (91.1-93.6) | 94.0 (90.8-96.1) | 92.9 (91.3-94.2) |
| IHC4 H | 87.7 (84.8-90.1) | 83.1 (81.2-84.9) | 88.6 (85.1-91.2) | 85.3 (83.0-87.4) |
| Mam L | 94.4 (92.4-95.8) | 92.9 (91.5-94.1) | 94.8 (92.5-96.4) | 93.4 (91.7-94.7) |
| Mam M | 84.9 (79.6-88.8) | 88.0 (85.7-90.0) | 86.4 (80.6-90.6) | 89.5 (86.7-91.8) |
| Mam H | 85.4 (78.7-90.2) | 83.4 (81.1-85.4) | 84.3 (76.3-89.8) | 85.6 (82.9-87.9) |
| **15-Year** | | | | |
| IHC4 L | 79.3 (65.1-88.3) | NA | 81.1 (66.1-89.9) | NA |
| IHC4 M | 84.5 (79.8-88.2) | NA | 85.8 (80.7-89.6) | NA |
| IHC4 H | 78.8 (75.1-82.0) | NA | 78.3 (73.8-82.1) | NA |
| Mam L | 85.0 (81.7-87.8) | NA | 85.9 (82.0-89.0) | NA |
| Mam M | 72.5 (65.7-78.2) | NA | 73.7 (65.9-80.0) | NA |
| Mam H | 75.0 (66.8-81.4) | NA | 73.3 (64.1-80.5) | NA |

NOTE. Freedom from DR estimates with 95% CIs for patients with low (L), medium (M) or high (H) risk as stratified by IHC4 or Mammostrat scores. Abbreviations: DR, distant-recurrence; CI, confidence interval; IHC4, IHC4 score; Mam, Mammostrat score; NA, not applicable. [a]Endocrine only, patients treated with adjuvant endocrine therapy without adjuvant chemotherapy

### 7.3.2 Classification of patients according to IHC4 and Mammostrat score

Table 7.4 illustrates the classification of all ER-positive patients according to the prognostic scores. Agreement in classification was only seen in 251 of 1,103 patients (23%) in the Edinburgh BCS cohort and 1,291 of 3,766 patients (34%) in the TEAM cohort. The largest proportion of patients were classified as low Mammostrat and either medium or high IHC4 score (60% and 40% in the Edinburgh BCS and TEAM cohorts respectively) with 11 and 18% classified as high Mammostrat and high IHC4 in the Edinburgh BCS and TEAM cohorts respectively. Figure 7.5 illustrates the Kaplan-Meier estimates of TTDR for different risk categories assigned by the two scores by collapsing the low and intermediate ICH4 score to

low risk and intermediate and high Mammostrat score to high risk. Good separation between the curves is observed, with having an overall low score (IHC4 and Mammostrat

**Table 7.4** Frequencies of patients according to their classification of IHC4 and Mammostrat scores

| | Edinburgh BCS (n=1,103) | | | TEAM (n=3,766) | | |
|---|---|---|---|---|---|---|
| | Mammostrat | | | Mammostrat | | |
| IHC4 | L | M | H | L | M | H |
| L | 48 (4.4) | 11 (1.0) | 4 (0.4) | 134 (3.6) | 51 (1.4) | 54 (1.4) |
| M | 305 (27.7) | 87 (7.9) | 32 (2.9) | 944 (25.1) | 469 (12.5) | 476 (12.6) |
| H | 351 (31.8) | 149 (13.5) | 116 (10.5) | 528 (14.0) | 422 (11.2) | 688 (18.3) |

Note: Values represent number (percentage of total) for each combination of IHC4 and

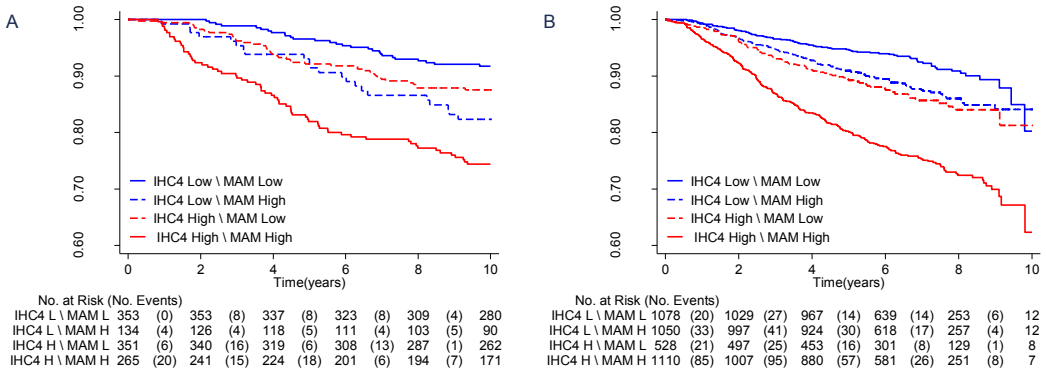



**Figure 7.5** Kaplan-Meier curves for TTDR for IHC4 score vs Mammostrat score. Plots represent all ER-positive patients in the (A) Edinburgh BCS and (B) TEAM cohorts. Categorisations refer to: IHC4 Low (low/intermediate risk); IHC4 High (high risk); Mam Low (low risk); Mam High(intermediate/high risk).

low risk) associated with better TTDR than an overall high score (ICH4 and Mammostrat high risk). Discordant scores (low risk IHC4 and high risk Mammostrat and vice versa) have more similar outcomes. This potentially identifies three risk groups from a combined IHC4 and Mammostrat risk score.

### 7.3.3 Assumption of Linearity

The functional form of the continuous covariates was investigated using step 1 of the MFPT approach described in section 4.4.4. The MFP algorithm identified tumour size and number of

positive nodes to have non-linear effects on the log-hazard in both the Edinburgh BCS and TEAM cohorts (Figure 7.6). The plots show the unrealistic estimated hazard ratio for increasing tumour size and number of positive nodes if a linear effect on the log-hazard is

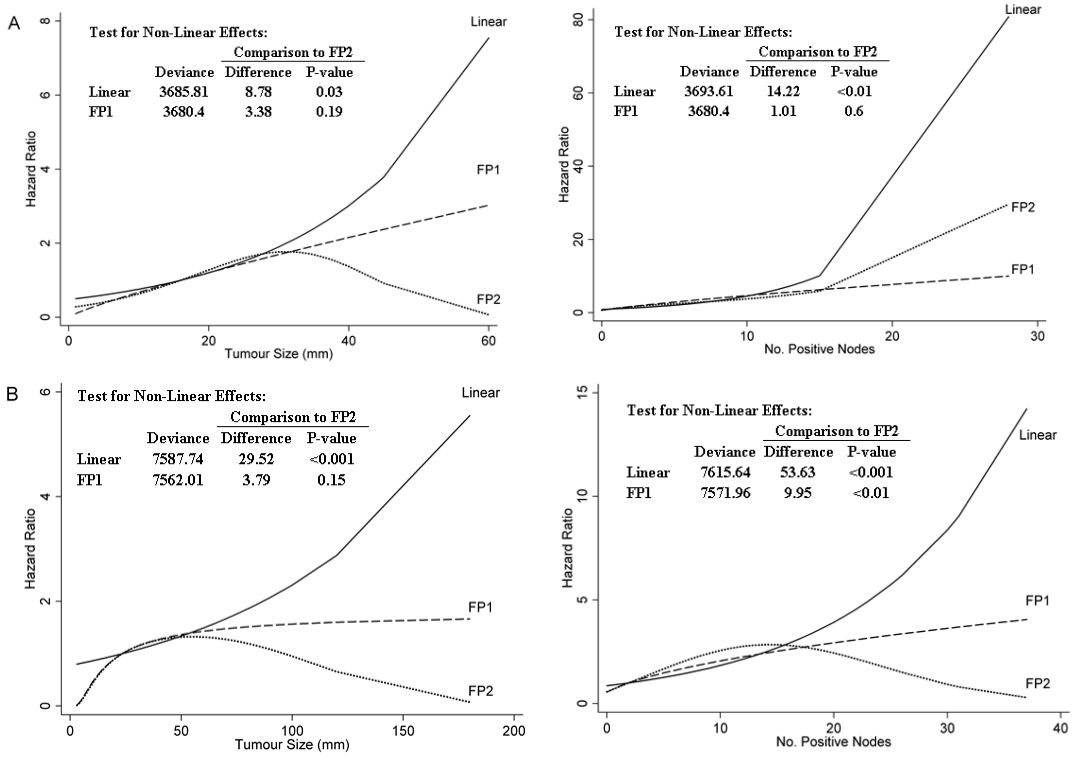



**Figure 7.6** Results of the MFP algorithm along with plots of the linear and non-linear effects of tumour size and number of positive nodes in the (A) Edinburgh BCS and (B) TEAM cohorts.

assumed. The FP1 model was not statistically significantly different from the FP2 model and therefore the simpler FP1 transformations were chosen to model the effects of tumour size and number of positive nodes. This was true except for the number of positive nodes in the TEAM cohort where the FP2 model was chosen, however the FP2 model results in a clinically unrealistic decrease in hazard for higher values of number of positive nodes and suggests an over-fitting of the data. This is due to the skewed distribution of the number of positive nodes and highlights the need for an initial exploration of the functional form with the use of

martingale residuals (Figure 7.7). FP1 transformations of tumour size and number of positive nodes will be used in all further analyses.

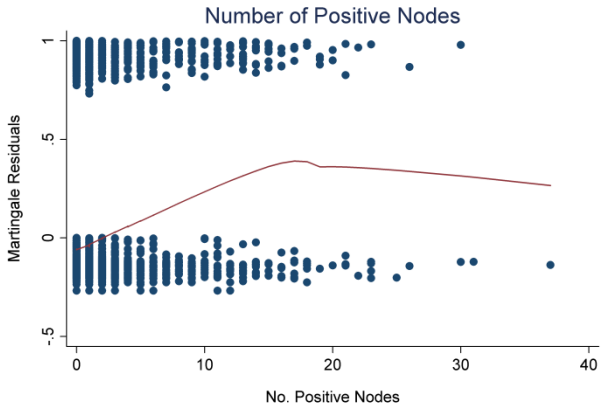



**Figure 7.7** Plot of martingale residuals versus number of positive nodes for all ER-positive patients in the TEAM cohort. The red line is a smoothed curve produced by locally weighted regression.

## 7.3.4 Comparison of IHC4 and Mammostrat

### *7.3.4.1 Separate analysis of scores*

Both scores were significant independent predictors of outcome when analysed in separate multivariate regression (Table 7.5 and Table 7.6) with other recognised prognostic factors (histologic grade, tumour size, age, number of positive nodes, treatment and chemotherapy) for all ER-positive patients and ER-positive patients treated with endocrine therapy only in the TEAM cohort. In this cohort, the IHC4 score provided more prognostic information beyond that of clinical factors compared to the Mammostrat score (increase in $R^2$: 3.3% vs 1.4% and D-statistic: 0.11 vs 0.05 for ER-positive patients treated with endocrine therapy only and increase in $R^2$: 3.5% vs 2.2% and D-statistic: 0.11 vs 0.07 for all ER-positive patients).

In the Edinburgh BCS cohort, the Mammostrat score was the stronger predictor of outcome with the prognostic information added by the IHC4 score being not statistically significant (HR (95% CI): 1.02 (0.99-1.05), p-value 0.1) and resulted in a negligible increase in $R^2$ and D-statistic of 0.6% and 0.02 respectively for all ER-positive patients treated with endocrine therapy only.

The Mammostrat score did not provide independent information regarding the risk of TTDR for any of the exploratory subgroups in the Edinburgh BCS cohort, although the small sample sizes should be noted: node-negative (n=657) and node-positive (n=174) ER-positive patients treated with endocrine therapy only and node-positive ER-positive patients irrespective of treatment (n=260) (Table 7.7 and Table 7.8). However, the IHC4 score was a predictor of outcome for ER-positive node-negative patients treated with endocrine therapy only (n=657) (p-value: <0.01; increase in $R^2$: 2.7% and D-statistic: 0.09).

Despite not adding statistically significant information based on the Wald test, possibly due to the small sample size of ER-positive, node-positive patients treated with endocrine therapy only (n=174), the addition of IHC4 to clinical factors statistically improved the ability to discriminate between events and non-events (increase in $R^2$: 3.7% and D-statistic: 0.12). In the TEAM cohort, both scores remained significant predictors of outcome in multivariate analysis for all exploratory subgroups except for the Mammostrat score in ER-positive, node-positive patients treated with endocrine therapy only (n=1,296) which was not statistically significant (HR (95% CI) MvL : 1.16 (0.81-1.68); HvL : 1.18 (0.83-1.67); p-value 0.2), and provided only a small increase in model performance (increase in $R^2$: 0.5% and D-statistic: 0.02).

Good agreement between observed and predicted risk of TTDR was seen in the TEAM cohort with no clear trend in improvement when adding either score to clinical factors (Figure 7.8). In the Edinburgh BCS cohort, the models under-estimated then over-estimated the risk of distant recurrence at 5 years when the observed risk was less than or greater than 10% respectively. This suggests the models tend to under-predict risk in the low-risk groups and over-predict risk in the high-risk groups. However, the addition of Mammostrat score to clinical factors improved the calibration compared with the IHC4 score.

**Table 7.5** Multivariate Cox regression of IHC4 score and Mammostrat score

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Edinburgh BCS (n=1,103) | TEAM (n=3,766) | Edinburgh BCS (n=831) | TEAM (n=2,513) |
| Main Effect | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| Separate* | | | | |
| IHC4 | 1.02 (0.99-1.05) | 1.06 (1.05-1.08) | 1.03 (0.99-1.06) | 1.06 (1.04-1.09) |
| MvL | 1.59 (1.13-2.24) | 1.42 (1.12-1.81) | 1.64 (1.10-2.44) | 1.28 (0.94-1.75) |
| HvL | 1.31 (0.86-2.00) | 1.89 (1.52-2.36) | 1.44 (0.90-2.31) | 1.57 (1.18-2.07) |
| Combined* | | | | |
| IHC4 | 1.02 (0.99-1.05) | 1.06 (1.04-1.07) | 1.02 (0.99-1.06) | 1.06 (1.03-1.08) |
| MvL | 1.56 (1.11-2.20) | 1.34 (1.05-1.71) | 1.63 (1.09-2.43) | 1.19 (0.87-1.63) |
| HvL | 1.21 (0.78-1.88) | 1.67 (1.34-2.10) | 1.33 (0.81-2.17) | 1.35 (1.01-1.81) |

NOTE. Multivariate analysis for Mammostrat score and IHC4 entered separately or simultaneously (combined analysis) into a Cox regression model. Values represent hazard ratios and 95% CIs for risk of TTDR calculated as a ten-unit increase in IHC4 score and either high-risk or moderate-risk Mammostrat score compared with low Mammostrat scores at baseline for all ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts. Abbreviations: TTDR, time-to distant recurrence; CI, confidence interval; HvL, high risk v low risk Mammostrat score; MvL, moderate risk v low risk Mammostrat score. *Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

**Table 7.6** Wald tests and performance data for assessing the amount of information added by the IHC4 score or the Mammostrat score or both to clinical factors

| Model | Edinburgh BCS | | | TEAM | | |
|---|---|---|---|---|---|---|
| | Wald test (p-value) | Increase in $R^2$ (%) | Increase in D Statistic | Wald test (p-value) | Increase in $R^2$ (%) | Increase in D Statistic |
| **All ER-Positive** | (n=1,103) | | | (n=3,766) | | |
| C vs. Null (7 df) | 13.1 (<0.001) | 24.0 | 1.15 | 326.1 (<0.001) | 27.5 | 1.26 |
| C + IHC4 vs. C (1 df) | 2.6 (0.1) | 0.5 | 0.01 | 56.1 (<0.001) | 3.5 | **0.11** |
| C + Mam vs. C (2 df) | 3.6 (0.03) | 1.4 | 0.04 | 32.5 (<0.001) | 2.2 | 0.07 |
| C + IHC4 + Mam vs. C (3 df) | 3.0 (0.03) | 1.8 | 0.06 | 76.4 (<0.001) | 4.8 | **0.15** |
| C + IHC4 + Mam vs. C + IHC4 (2 df) | 3.3 (0.04) | 1.3 | 0.04 | 20.3 (<0.001) | 1.3 | 0.04 |
| C + IHC4 + Mam vs. C + Mam (1 df) | 1.9 (0.2) | 0.4 | 0.01 | 42.3 (<0.001) | 2.6 | 0.08 |
| **ER-Positive Endocrine Only** | (n=831) | | | (n=2,513) | | |
| C vs. Null (6 df) | 16.1 (<0.001) | 25.7 | 1.20 | 217.3 (<0.001) | 29.5 | 1.33 |
| C + IHC4 vs. C (1 df) | 2.5 (0.11) | 0.6 | 0.02 | 30.1 (<0.001) | 3.3 | **0.11** |
| C + Mam vs. C (2 df) | 3.1 (0.04) | 1.4 | 0.05 | 9.9 (0.007) | 1.4 | 0.05 |
| C + IHC4 + Mam vs. C (3 df) | 2.7 (0.04) | 2.2 | 0.07 | 34.4 (<0.001) | 3.8 | **0.12** |
| C + IHC4 + Mam vs. C + IHC4 (2 df) | 2.9 (0.06) | 1.6 | 0.05 | 4.2 (0.12) | 0.5 | 0.02 |
| C + IHC4 + Mam vs. C + Mam (1 df) | 1.9 (0.17) | 0.8 | 0.02 | 23.6 (<0.001) | 2.4 | 0.08 |

NOTE. Values represent Wald tests (significance level) and the increase in $R^2$ and D statistic for the addition of IHC4 or Mammostrat score or both with a difference in D of at least 0.1 indicating improved prognostic separation. Results are given for ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts. Abbreviations: C, clinical model with age, grade, tumour size, nodes positive, treatment and chemotherapy; Null, null model with no covariates; IHC4, IHC4 score; Mam, Mammostrat score; df, degrees of freedom.

**Table 7.7** Multivariate Cox regression of IHC4 score and Mammostrat score for Exploratory Subgroups

| | ER-Positive Endocrine Only and Node-Negative | | ER-Positive Endocrine Only and Node-Positive | | ER-Positive Any Treatment and Node-Positive | |
|---|---|---|---|---|---|---|
| | Edinburgh BCS (n=657) | TEAM (n=1208) | Edinburgh BCS (n=174) | TEAM (n=1,296) | Edinburgh BCS (n=260) | TEAM (n=2,229) |
| Main Effect | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| Seperate* | | | | | | |
| IHC4 | 1.06 (1.02-1.11) | 1.10 (1.06-1.14) | 0.96 (0.90-1.02) | 1.05 (1.02-1.08) | 0.99 (0.94-1.04) | 1.05 (1.03-1.07) |
| MvL | 1.71 (1.04-2.81) | 1.44 (0.78-2.65) | 1.54 (0.77-3.08) | 1.25 (0.87-1.79) | 1.48 (0.82-2.67) | 1.42 (1.09-1.86) |
| HvL | 1.58 (0.88-2.85) | 2.37 (1.42-3.97) | 1.36 (0.59-3.12) | 1.33 (0.94-1.86) | 1.39 (0.70-2.77) | 1.62 (1.25-2.09) |
| Combined* | | | | | | |
| IHC4 | 1.06 (1.02-1.11) | 1.08 (1.04-1.13) | 0.95 (0.89-1.01) | 1.05 (1.02-1.08) | 0.99 (0.93-1.04) | 1.05 (1.03-1.07) |
| MvL | 1.69 (1.02-2.78) | 1.32 (0.72-2.44) | 1.51 (0.75-3.03) | 1.16 (0.81-1.68) | 1.48 (0.82-2.68) | 1.35 (1.03-1.77) |
| HvL | 1.38 (0.76-2.51) | 1.95 (1.16-3.30) | 1.74 (0.74-4.08) | 1.18 (0.83-1.67) | 1.48 (0.72-3.07) | 1.48 (1.14-1.92) |

NOTE. Multivariate analysis for Mammostrat score and IHC4 entered seperately or simultaneously (combined analysis) into a Cox regression model. Hazard ratios for risk of TTDR for patients with a ten-unit increase in IHC4 score and with either high-risk or moderate-risk Mammostrat score compared with low Mammostrat scores at baseline. Abbreviations: TTDR, time-to distant-recurrence; CI, confidence interal; HvL, high risk v low risk Mammostrat score; MvL, medium risk v low risk Mammostrat score. * Adjusted for age, grade, tumour size,nodal status, treatment and chemotherapy.

**Table 7.8** Wald tests and performance data for assessing the amount of information added by IHC4 score or Mammostrat score or both to clinical factors for exploratory subgroups.

| | Edinburgh BCS | | |
|---|---|---|---|
| Model | Wald test (p-value) | Increase in $R^2$ | Increase in D Statistic |
| **ER-Positive Endocrine Only and Node-Negative** | | | |
| C vs. Null (7 df) | 7.0 (<0.001) | 17.8 | 0.95 |
| C + IHC4 vs. C (1 df) | 8.7 (0.003) | 2.7 | 0.09 |
| C + Mam vs. C (2 df) | 2.5 (0.08) | 1.4 | 0.05 |
| C + IHC4 + Mam vs. C (3 df) | 4.2 (0.006) | 4.3 | **0.14** |
| C + IHC4 + Mam vs. C + IHC4 (2 df) | 2.1 (0.12) | 1.6 | 0.05 |
| C + IHC4 + Mam vs. C + Mam (1 df) | 7.8 (0.005) | 2.9 | 0.09 |
| **ER-Positive Endocrine Only and Node-Positive** | | | |
| C vs. Null (7 df) | 5.1 (<0.001) | 23.7 | 1.14 |
| C + IHC4 vs. C (1 df) | 2.0 (0.15) | 3.7 | **0.12** |
| C + Mam vs. C (2 df) | 0.8 (0.47) | 0.9 | 0.03 |
| C + IHC4 + Mam vs. C (3 df) | 1.4 (0.24) | 4.5 | **0.14** |
| C + IHC4 + Mam vs. C + IHC4 (2 df) | 1.1 (0.34) | 0.8 | 0.03 |
| C + IHC4 + Mam vs. C + Mam (1 df) | 2.7 (0.1) | 4.6 | **0.10** |
| **ER-Positive Any Treatment and Node-Positive** | | | |
| C vs. Null (7 df) | 5.0 (<0.001) | 21.5 | 1.07 |
| C + IHC4 vs. C (1 df) | 0.1 (0.8) | -0.2 | 0.00 |
| C + Mam vs. C (2 df) | 1.0 (0.4) | 1.3 | 0.04 |
| C + IHC4 + Mam vs. C (3 df) | 0.7 (0.5) | 1.1 | 0.03 |
| C + IHC4 + Mam vs. C + IHC4 (2 df) | 1.1 (0.3) | 1.3 | 0.04 |
| C + IHC4 + Mam vs. C + Mam (1 df) | 0.3 (0.6) | -0.2 | 0.00 |

**Table 7.8 Continued** Wald tests and performance data for assessing the amount of information added by IHC4 score or Mammostrat score or both to clinical factors for exploratory subgroups.

| | TEAM | | |
|---|---|---|---|
| Model | Wald test (p-value) | Increase in $R^2$ | Increase in D Statistic |
| **ER-Positive Endocrine Only and Node-Negative** | | | |
| C vs. Null (7 df) | 35.0 (<0.001) | 19.9 | 1.02 |
| C + IHC4 vs. C (1 df) | 23.3 (<0.001) | 9.9 | **0.32** |
| C + Mam vs. C (2 df) | 11.8 (0.003) | 5.3 | **0.17** |
| C + IHC4 + Mam vs. C (3 df) | 29.5 (<0.001) | 11.4 | **0.36** |
| C + IHC4 + Mam vs. C + IHC4 (2 df) | 6.7 (0.03) | 1.5 | 0.04 |
| C + IHC4 + Mam vs. C + Mam (1 df) | 17.4 (<0.001) | 6.1 | **0.19** |
| **ER-Positive Endocrine Only and Node-Positive** | | | |
| C vs. Null (7 df) | 125.1 (<0.001) | 28.1 | 1.28 |
| C + IHC4 vs. C (1 df) | 12.0 (<0.001) | 2.1 | 0.07 |
| C + Mam vs. C (2 df) | 2.9 (0.2) | 0.5 | 0.02 |
| C + IHC4 + Mam vs. C (3 df) | 13.0 (0.005) | 2.1 | 0.07 |
| C + IHC4 + Mam vs. C + IHC4 (2 df) | 1.0 (0.6) | 0.02 | 0.00 |
| C + IHC4 + Mam vs. C + Mam (1 df) | 9.9 (0.002) | 1.6 | 0.05 |
| **ER-Positive Any Treatment and Node-Positive** | | | |
| C vs. Null (7 df) | 200.4 (<0.001) | 25.5 | 1.2 |
| C + IHC4 vs. C (1 df) | 26.4 (<0.001) | 2.4 | 0.08 |
| C + Mam vs. C (2 df) | 13.8 (0.001) | 1.3 | 0.04 |
| C + IHC4 + Mam vs. C (3 df) | 35.3 (<0.001) | 3.2 | 0.10 |
| C + IHC4 + Mam vs. C + IHC4 (2 df) | 9.0 (0.01) | 0.8 | 0.03 |
| C + IHC4 + Mam vs. C + Mam (1 df) | 20.9 (<0.001) | 1.9 | 0.06 |

NOTE. Values represent Wald tests (significance level) and the increase in $R^2$ and D statistic for the addition of IHC4 or Mammostrat score or both with a difference in D of at least 0.1 indicating improved prognostic separation. Results are given for exploratory subgroups in the Edinburgh BCS and TEAM cohorts. Abbreviations: C, clinical model with age, grade, tumour size, nodes positive, treatment and chemotherapy; Null, null model with no covariates; IHC4, IHC4 score; Mam, Mammostrat score; df, degrees of freedom.

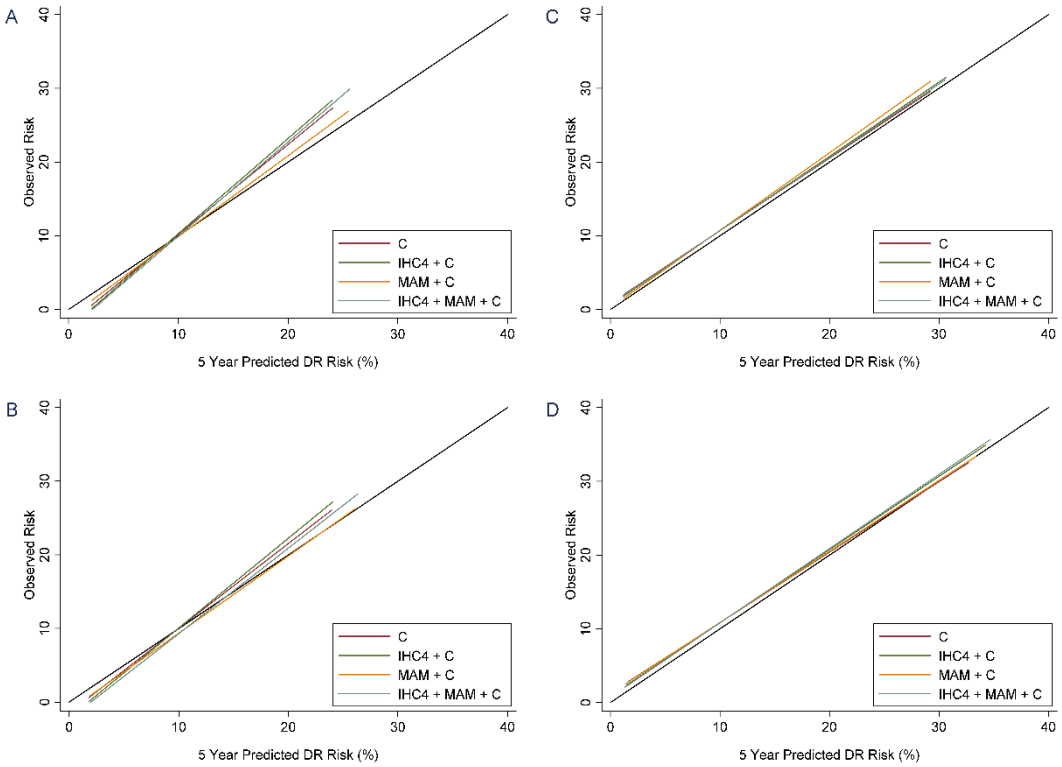



**Figure 7.8** Calibration slope estimates of observed versus predicted 5 year risk of distant recurrence (DR). Plots represent ER-positive patients treated with endocrine therapy only (A and C), all ER-positive patients (B and D) in the Edinburgh BCS (A and B) and TEAM (C and D) cohorts. Black line shows perfect agreement between observed and predicted risk. Abbreviations: C, clinical model with age, grade, tumour size, number of positive nodes, treatment and chemotherapy; IHC4 + C, IHC4 score plus clinical model; MAM+C, Mammostrat score plus clinical model; IHC4+MAM+C, IHC4 and Mammostrat score plus clinical model.

### *7.3.4.2 The addition of both scores to clinical factors*

The scores were entered simultaneously into a multivariate Cox regression model (Table 7.5 and Table 7.6) and the addition of both scores to clinical factors provided statistically significant information ($p<0.05$) for both subsets of patients across both cohorts with increases of $R^2$ between 2 and 5% and increases in D-statistic of between 0.06 and 0.15. Similarly in the exploratory subgroup analysis (Table 7.7 and Table 7.8), the addition of both scores provided significant additional prognostic information to clinical factors only in the TEAM cohort but this was not seen in those with node-positive disease in the Edinburgh BCS cohort.

The Mammostrat score remained the stronger predictor of TTDR in the Edinburgh BCS cohort with the IHC4 score not providing any statistically significant information beyond that of

Mammostrat and clinical factors combined for both patient subsets (increase in $R^2$: 0.8 and 0.4% and D-statistic: 0.02 and 0.01 for ER-positive patients treated with endocrine therapy only and all ER-positive patients respectively).

In the TEAM cohort, the IHC4 provided more prognostic information than the Mammostrat score in both subgroups (increase in $R^2$: 2.4 vs 0.5% and D-statistic: 0.08 vs 0.02 for ER-positive patients treated with endocrine therapy only (n=2,513) and increase in $R^2$: 2.6 vs 1.3% and D-statistic: 0.08 vs 0.04 for all ER-positive patients (n=3,766)). The Mammostrat score offered prognostic information beyond that of IHC4 score and clinical factors combined in all ER-positive patients ($p<0.001$), although only a small improvement in model discrimination was observed (increase in $R^2$: 1.3% and D-statistic: 0.04). The prognostic information added by Mammostrat score was not significant in ER-positive patients treated with endocrine therapy only ($p=0.1$, increase in $R^2$: 0.5% and D-statistic: 0.02).

Again varied results were seen for the exploratory analysis depending on the subgroup and study: the IHC4 score was the stronger predictor of TTDR in the TEAM cohort but additional information was gained from including Mammostrat score for ER-positive, node-negative patients treated with endocrine therapy ($p=0.03$, increase in $R^2$: 1.5% and D-statistic: 0.04) and ER-positive, node-positive patients irrespective of treatment ($p=0.01$, increase in $R^2$: 0.8% and D-statistic: 0.03).

Model calibration was also investigated with no clear improvements observed from the combined use of both scores in addition to clinical factors (Figure 7.8).

### 7.3.5 Assumption of Proportional Hazards

A graphical procedure can be used which plots Schoenfelds residuals against time (or the logarithm of time), with patterns in these plots indicating non-proportional effects. The plots of the residuals are given in Figure 7.9. In the Edinburgh BCS cohort several explanatory variables showed evidence of a non-proportional effect (IHC4 score, Mammostrat score and

grade). IHC4 score and Mammostrat high risk also showed evidence of a non-proportional effect in the TEAM cohort, however the effect was more prominent in the longer follow-up of the Edinburgh BCS cohort.

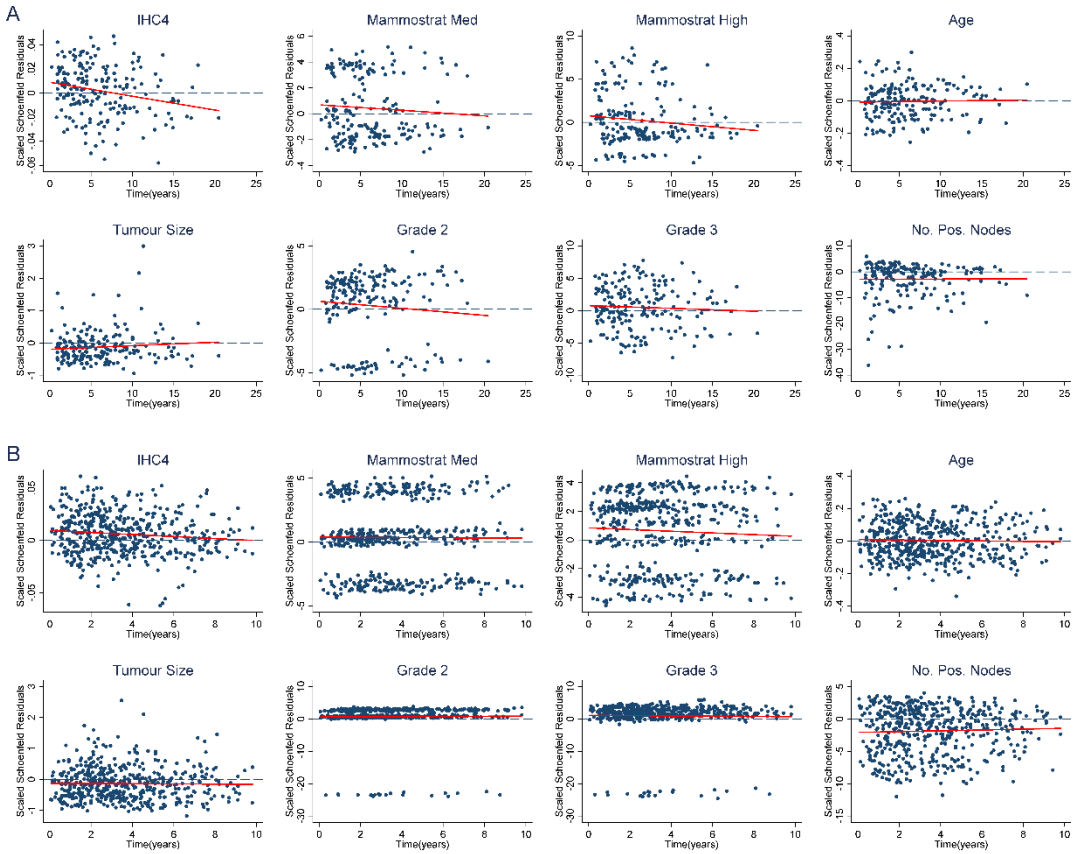



**Figure 7.9** Plots of Schoenfeld residuals for each explanatory variable versus survival time in the (A) Edinburgh BCS and (B) TEAM cohorts with red (solid) line indicating the fitted linear regression line and the blue (dashed) line a reference line at zero.

A formal test can also be performed which tests, for individual covariates and globally, the null hypothesis of zero slope, which is equivalent to testing that the log-hazard function is constant over time. The formal tests identify IHC4 to have a non-proportional effect in both cohorts (Table 7.9). It is however important to look at plots as well as the significance of the

tests as significant results may not be clinically important and tests will not be good at detecting curved relationships.

**Table 7.9** Grambsch-Therneau test of non-proportional hazards

| | P-value of Grambsch-Therneau Test | | | |
|---|---|---|---|---|
| | Edinburgh BCS | | TEAM | |
| Variable | Univariable | Multivariable | Univariable | Multivariable |
| IHC4 | <0.001 | <0.001 | <0.001 | 0.02 |
| MAM Medium | 0.10 | 0.31 | 0.79 | 0.89 |
| MAM High | 0.01 | 0.10 | 0.07 | 0.49 |
| Age | 0.44 | 0.71 | 0.82 | 0.54 |
| Size | 0.03 | 0.13 | 0.31 | 0.80 |
| Grade 2 | 0.04 | 0.23 | 0.90 | 0.74 |
| Grade 3 | 0.01 | 0.44 | 0.42 | 0.80 |
| No. Pos. Nodes | 0.71 | 0.94 | 0.19 | 0.35 |

It is also important to note the differences when variables are analysed in a univariable model, with Mammostrat high risk and tumour size violating the PH assumption in a univariate but not multivariate analysis. This relates to the issue of spurious non-proportional effects when other important variables are omitted from the model (Buchholz and Sauerbrei, 2011).

## 7.4 Discussion

To our knowledge, these prognostic algorithms have not been evaluated in a head-to-head comparison before. Assessment was complicated as IHC4 was a continuous risk score and Mammostrat was considered as a risk category (high, moderate or low). The results showed that the scores have different capabilities in predicting TTDR depending on the cohort and the subgroup of patients. As a single panel IHC4 provided more information than Mammostrat in the TEAM cohort whereas in the Edinburgh BCS cohort the Mammostrat score was the stronger predictor of TTDR. However, particularly in the larger TEAM cohort and when all ER-positive patients were considered, statistically significant benefit in estimation of residual recurrence risk after treatment was observed from a combined use of both marker panels.

We have shown that two distinct IHC marker panels provide independent information on risk of distant recurrence following treatment with tamoxifen or aromatase inhibitors within the Edinburgh BCS and TEAM cohorts. In multivariate analysis, the "IHC4 algorithm" derived from a mathematical transformation of quantitative scores for ER, PgR, HER2 and Ki67 and Mammostrat risk score derived from a panel of five antibodies p53, NDRG1,CEACAM5, SLC7A5 and HTF9C were significantly associated with risk of distant recurrence. These results would appear to provide an independently generated validation of both the "IHC4 algorithm" (Cuzick et al., 2011) and Mammostrat score (Ring et al., 2006).

The IHC4 is analysed as a continuous score, but for Kaplan-Meier analysis cut-points are required. To avoid the biases that occur from choosing our own cut-points, we used those from the original study (Cuzick et al., 2011). However, these did not validate well in our cohorts, allocating only a small number of patients to the low risk group (<10%).

From this initial analysis, the IHC4 and Mammostrat risk scores were significantly associated with risk of time to distant recurrence and added prognostic information beyond that provided by clinical factors. The combined use of IHC4 and Mammostrat in choosing patients to be

treated with chemotherapy warrants further study and could be a much more cost-effective option than the considerably more expensive mRNA based assays currently available.

Depending on the study and the subgroup of patients analysed, the scores demonstrated different capabilities in predicting TTDR. One potential reason for the difference between study cohorts could be the differences in the lengths of follow-up, with the studies having median follow-ups of 6.2 years and 12.9 years for the TEAM and Edinburgh BCS cohorts respectively. The Cox analysis performed in this study has the assumption of proportional hazards (PH) which corresponds to assuming that the hazard ratio comparing one level of a covariate to another is constant over time. The misleading conclusions that can be inferred if this assumption has been violated has been well documented such as poor parameter estimates and we are less likely to conclude for a significant effect when there actually is one (Bellera et al., 2010). Violations of the PH assumption were found and a detailed analysis looking at the impact of follow-up duration and non-proportional effects follows in the next chapter.

# Chapter 8: Analysis of IHC4 and Mammostrat Incorporating Non-Proportional Effects

## 8.1 Introduction

We have previously shown in chapter 7, that when assessed as a single panel IHC4 provided consistently more residual risk information than Mammostrat. However, particularly in the wider TEAM cohort including all ER-positive patients (irrespective of treatment), significant benefit in estimation of residual recurrence risk after treatment was obtained by a combined use of both maker panels. However, differences were observed in the performance of the scores across the two studies and it was thought this could be due to differences in length of follow-up and the impact of non-proportional (time-dependent) effects.

In this chapter we investigate the impact of follow-up duration on the IHC4 and Mammostrat scores to determine whether these two prognostic panels provide information on the risk of early or late recurrence.

## 8.2 Piece-wise constant effects

The most simple and often used approach when dealing with non-proportional effects is the use of piece-wise constant effects (see section 4.4.4), where time is partitioned into intervals resulting in a step function for $\beta(t)$ in the extended Cox model.

A cut-point of 5 years was pre-specified due to the clinical relevance as a decision point for continued endocrine therapy at this time. Smaller intervals of two years were also explored.

## 8.2.1 Results

### *8.2.1.1 Model Performance*

We assessed the performance of the scores, using measures of discrimination and calibration as described in section 7.2.2, in addition to clinical factors at differing lengths of follow-up. Measures of discrimination are given in Table 8.1 for full follow-up (using all available data rather than censoring at a specific time point) and follow-up censored at 5 years. In the Edinburgh BCS cohort the models performed statistically better with shorter follow-up (5 years) compared to full follow-up (25 years) with differences in D statistic between 0.4 and 0.5 and $R^2$ between 7 and 13%. There was a small improvement in model performance with shorter follow-up in the TEAM cohort, with increases in $R^2$ between 1.5 and 3% and differences in D statistic between 0.05 and 1. Plots of observed versus predict risk of distant recurrence are displayed in Figure 8.1 and Figure 8.2 for all ER-positive patients and ER-positive patients treated with endocrine therapy respectively. The calibration of the combined model (IHC4 and Mammostrat score in addition to clinical factors) was improved censoring follow-up at 5 years in the Edinburgh BCS cohort only. The combined model calibration slope estimate was 1.0 (95% CI, 0.8-1.1) for follow-up censored at 5 years versus 1.2 (95% CI, 0.8-1.5) for full follow-up for all ER-positive patients.

**Table 8.1** Performance data on IHC4 and Mammostrat score in addition to clinical factors

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| Model | $R^2$ (95 % CI) | D Statistic (95% CI) | $R^2$ (95 % CI) | D Statistic (95% CI) |
| **Edinburgh BCS** | | | | |
| Full Follow-up | | | | |
| Clinical | 24.0 (16.2-31.7) | 1.15 (0.91-1.39) | 25.7 (16.7-34.6) | 1.20 (0.92-1.48) |
| IHC4 | 24.4 (16.7-32.1) | 1.16 (0.92-1.41) | 26.3 (17.4-35.1) | 1.22 (0.94-1.50) |
| MAM | 25.4 (17.6-33.2) | 1.19 (0.95-1.44) | 27.1 (18.1-36.1) | 1.25 (0.97-1.53) |
| Comb | 25.7 (18.0-33.5) | 1.21 (0.96-1.45) | 27.9 (19.0-36.8) | 1.27 (0.99-1.56) |
| 5 Years | | | | |
| Clinical | 31.1 (20.8-41.4) | 1.51 (1.17-1.85) | 35.3 (23.3-47.4) | 1.51 (1.12-1.91) |
| IHC4 | 33.7 (23.6-43.8) | 1.63 (1.28-1.98) | 39.0 (27.2-50.7) | 1.63 (1.23-2.04) |
| MAM | 33.6 (23.4-43.8) | 1.58 (1.23-1.92) | 38.5 (26.8-50.3) | 1.62 (1.22-2.02) |
| Comb | 36.5 (26.4-46.5) | 1.69 (1.34-2.04) | 41.3 (29.9-52.7) | 1.72 (1.31-2.12) |
| **TEAM** | | | | |
| Full-Follow-up | | | | |
| Clinical | 27.5 (23.0-32.0) | 1.26 (1.12-1.40) | 29.5 (23.6-35.3) | 1.33 (1.14-1.51) |
| IHC4 | 31.0 (26.5-35.4) | 1.37 (1.23-1.52) | 32.8 (27.0-38.4) | 1.43 (1.24-1.62) |
| MAM | 29.8 (25.3-34.2) | 1.33 (1.19-1.48) | 31.0 (25.1-36.7) | 1.37 (1.18-1.56) |
| Comb | 32.3 (27.8-36.7) | 1.41 (1.27-1.56) | 33.3 (27.5-38.9) | 1.45 (1.26-1.63) |
| 5 Years | | | | |
| Clinical | 29.0 (23.8-34.1) | 1.31 (1.14-1.47) | 30.5 (23.7-37.0) | 1.36 (1.14-1.57) |
| IHC4 | 34.0 (28.9-38.9) | 1.47 (1.31-1.63) | 34.9 (28.3-41.2) | 1.50 (0.29-1.71) |
| MAM | 31.5 (26.4-36.5) | 1.39 (1.23-1.55) | 32.2 (25.5-38.6) | 1.41 (1.20-1.62) |
| Comb | 35.4 (30.3-40.2) | 1.51 (1.35-1.68) | 35.5 (28.9-41.7) | 1.52 (1.31-1.73) |

NOTE. Measures of discrimination for patients in the Edinburgh BCS and TEAM cohorts. The four models being assessed are: Clinical, age, grade, tumour size, nodal status, treatment and chemotherapy; IHC4, IHC4 score in addition to clinical factors; Mammostrat (MAM), Mammostrat score in addition to clinical factors; Combined (Comb), IHC4 score and Mammostrat score in addition to clinical factors.

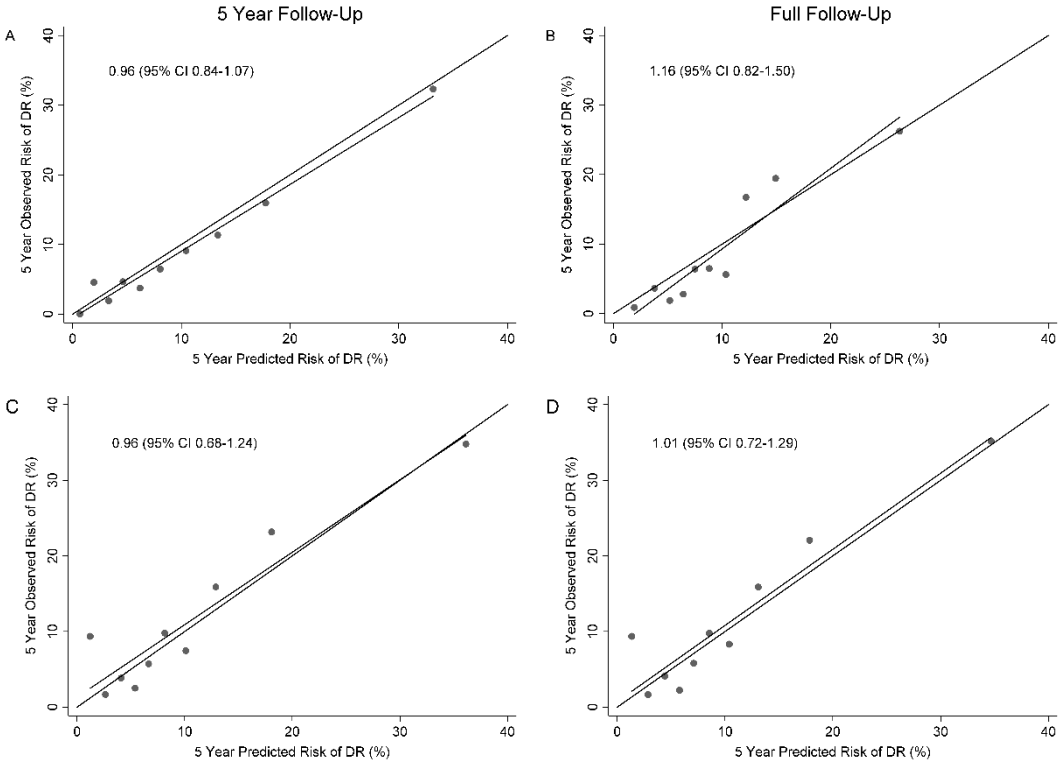



**Figure 8.1** All ER-positive patients. Observed versus predicted 5 year risk of distant recurrence (DR) and the calibration slope estimate for the combined model (IHC4, Mammostrat and clinical factors, age, grade, nodes positive, grade, size, treatment and chemotherapy) with full follow-up (right column) and follow-up censored at 5 years (right column) for all ER-positive patients in the Edinburgh BCS (A and B) and TEAM (C and D) cohorts. Dashed line represents perfect agreement between observed and predicted risk and the solid line is the fitted regression line.

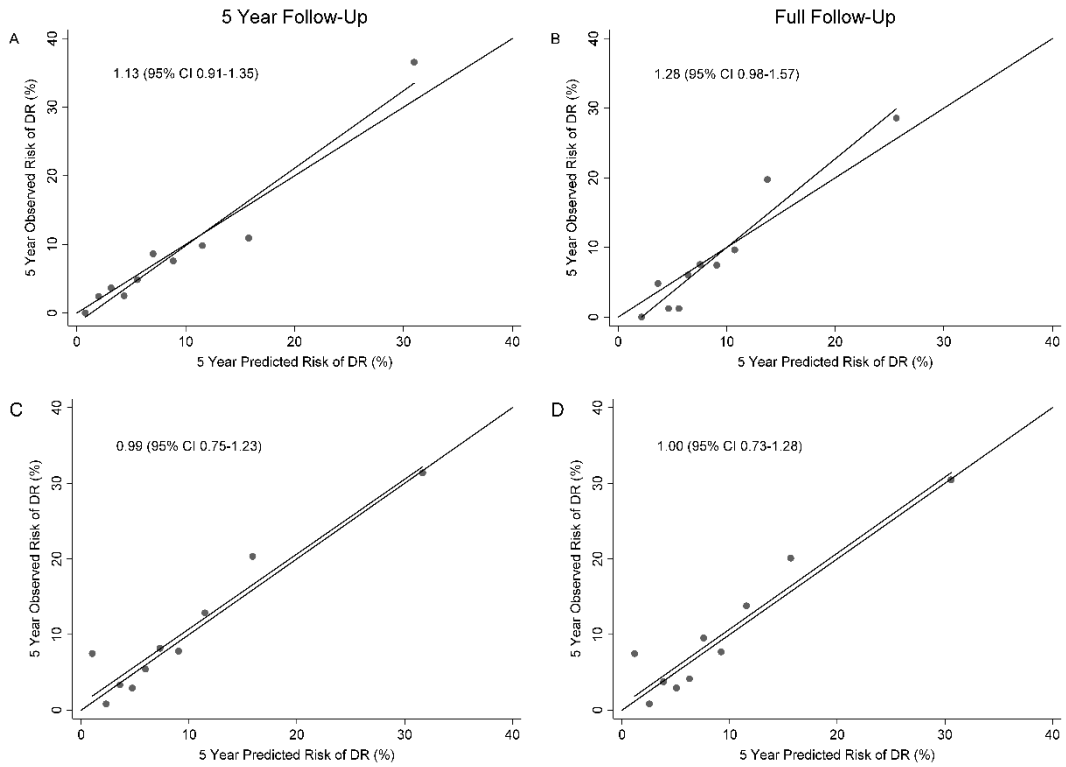



**Figure 8.2** ER-positive patients treated with endocrine therapy only. Observed versus predicted 5 year risk of distant recurrence (DR) and the calibration slope estimate for the combined model (IHC4, Mammostrat and clinical factors, age, grade, nodes positive, grade, size, treatment and chemotherapy) with full follow-up (right column) and follow-up censored at 5 years (right column) for ER-positive patients treated with endocrine therapy only in the Edinburgh BCS (A and B) and TEAM (C and D) cohorts. Dashed line represents perfect agreement between observed and predicted risk and the solid line is the fitted regression line.

#### *8.2.1.2 Prognostic Value of Scores within the first 5 years and beyond 5 years after diagnosis*

Cox regression was performed with follow-up time divided into the intervals 0-5 years and 5-10 years for both the Edinburgh BCS and TEAM cohorts.

Period specific K-M curves are displayed for all ER-positive patients in Figure 8.3 and ER-positive patients treated with endocrine therapy in Figure 8.4. Clear separation of the survival curves can be seen in the 0-5 year interval, however this separation does not remain in the 5-10 year interval.

Both scores were significant independent predictors of outcome restricted to the first five years of follow-up, after which there was no evidence the scores were associated with TTDR (Table 8.2). For example, the HR for a ten-unit increase in IHC4 score was 1.07 (95% CI, 1.03-1.12) in the first five years following diagnosis compared with 0.97 (95% CI, 0.92-1.02) after 5 years for all ER-positive patients in the Edinburgh BCS cohort. Despite categorising the time interval there are a sufficient number of events, especially in the TEAM cohort, in both time intervals suggesting that the finding of no evidence of a prognostic effect of the scores in the 5-10 year time interval does not just reflect a lack of endpoints.

There was evidence of a prognostic effect after 5 years for Mammostrat high risk versus low risk for all ER-positive patients (Table 8.2) and ER-positive node-negative patients treated with endocrine therapy (Table 8.3) in the TEAM cohort with HRs of 1.6 (95% CI, 1.0-2.4) and 3.3 (95% CI, 1.1-10.5) respectively. This effect of Mammostrat high versus low risk was also seen in the Edinburgh BCS cohort for ER-positive node-negative patients treated with endocrine therapy only (Table 8.3) for the 5-10 year time period after diagnosis with a HR of 2.8 (95% CI, 1.0-7.8).

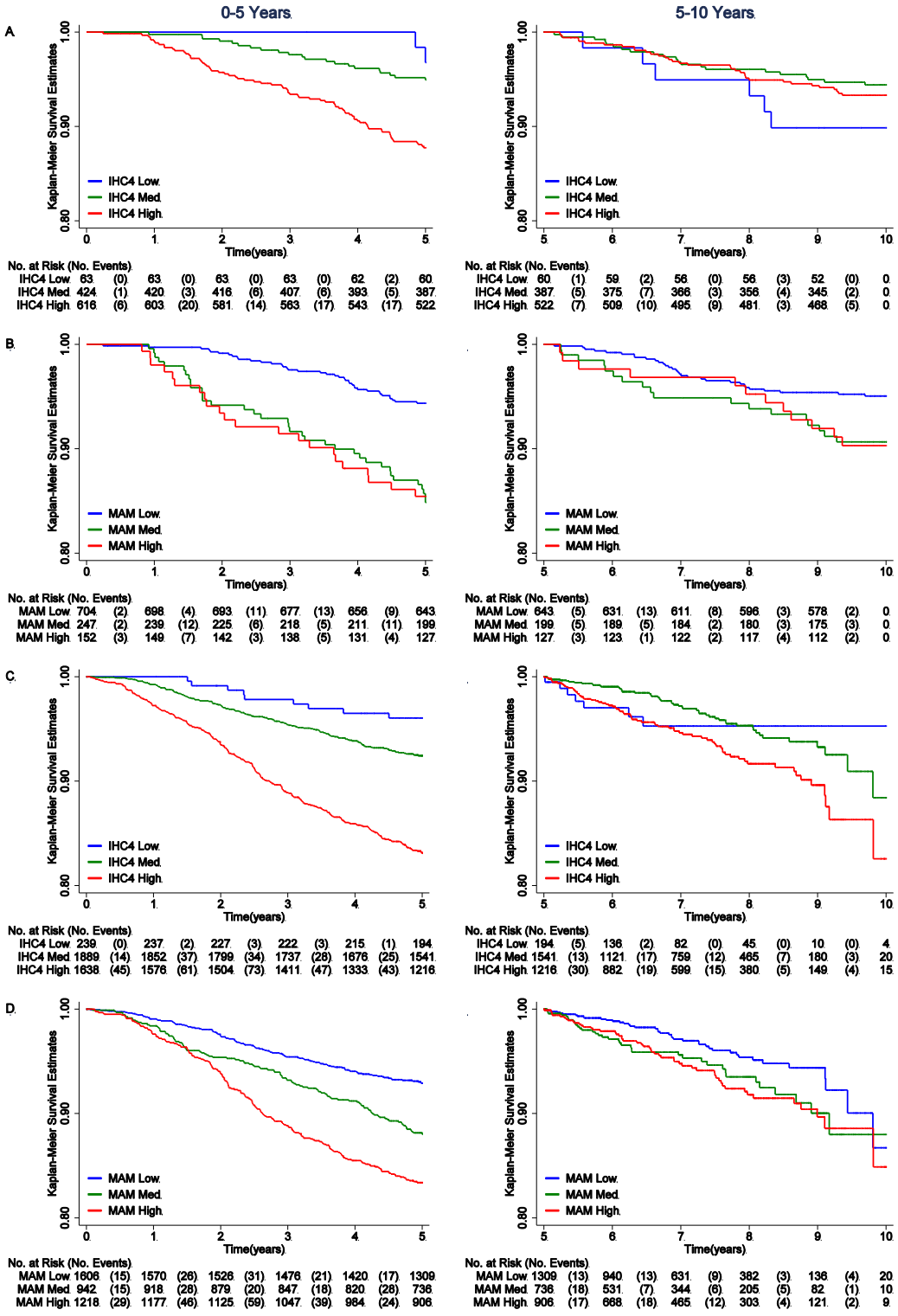



**Figure 8.3** All ER-positive patients. Period-specific Kaplan-Meier curves for time to distant recurrence for IHC4 score and Mammostrat s**core**. Plots represent (A and B) Edinburgh BCS dataset and (C and D) TEAM dataset. Curves are shown for the period 0-5 years, censoring follow-up of all patients at 5 years after diagnosis. The 5-10 year interval is assessed from the subset of patients who remained distant-recurrence free for at least 5 years and censoring follow-up of patients at 10 years.

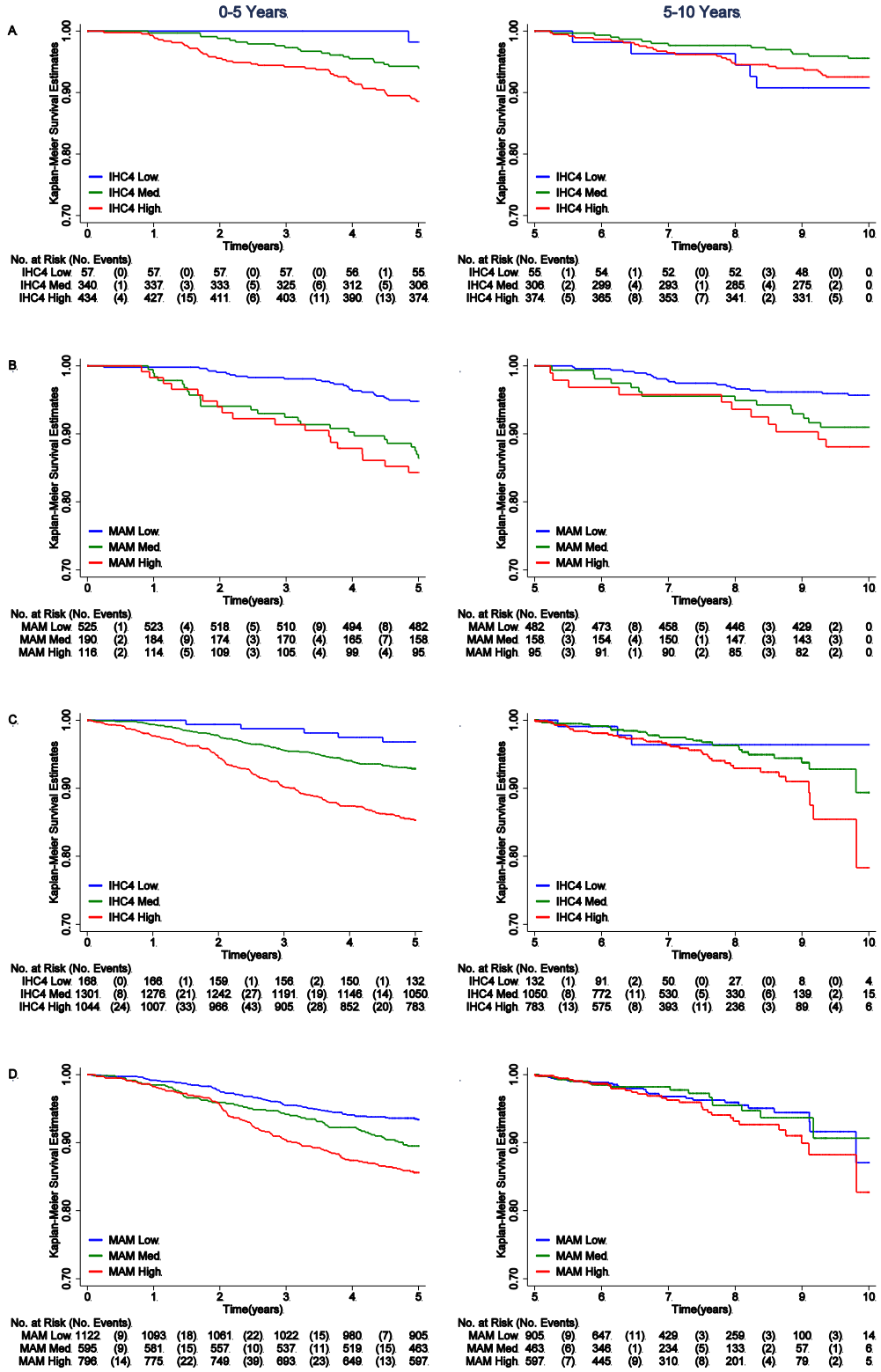



**Figure 8.4** ER-positive patients treated with endocrine therapy only. Period-specific Kaplan-Meier curves for time to distant recurrence for IHC4 score and Mammostrat score. Plots represent (A and B) Edinburgh BCS dataset and (C and D) TEAM dataset. Curves are shown for the period 0-5 years, censoring follow-up of all patients at 5 years after diagnosis. The 5-10 year interval is assessed from the subset of patients who remained distant-recurrence free for at least 5 years and censoring follow-up of patients at 10 years.

**Table 8.2** Period-Specific Multivariate Cox Regression of IHC4 and Mammostrat score

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | 0-5 Years | 5-10 Years | 0-5 Years | 5-10 years |
| Main Effect* | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| Edinburgh BCS | (97 events) | (61 events) | (70 events) | (45 events) |
| IHC4 | 1.07 (1.03-1.12) | 0.97 (0.92-1.02) | 1.06 (1.01-1.11) | 0.99 (0.93-1.05) |
| MvL | 1.84 (1.14-2.97) | 1.50 (0.80-2.80) | 2.00 (1.13-3.54) | 1.50 (0.72-3.13) |
| HvL | 1.65 (0.94-2.92) | 1.36 (0.65-2.85) | 1.87 (0.98-3.57) | 1.68 (0.75-3.77) |
| TEAM | (416 events) | (132 events) | (242 events) | (74 events) |
| IHC4 | 1.08 (1.06-1.10) | 1.02 (0.99-1.06) | 1.08 (1.05-1.10) | 1.01 (0.97-1.06) |
| MvL | 1.44 (1.09-1.90) | 1.38 (0.86-2.21) | 1.41 (0.98-2.01) | 0.96 (0.51-1.81) |
| HvL | 2.01 (1.56-2.60) | 1.57 (1.02-2.44) | 1.72 (1.24-2.38) | 1.18 (0.69-2.04) |

NOTE. Separate multivariate analysis of IHC4 and Mammostrat scores with conventional histopathologic variables for TTDR before 5 years and after 5 years for patients in different subgroups. Values represent estimated hazard ratios and 95% CIs calculated as a ten-unit increase in IHC4 score and high and medium risk compared to low risk as categorised by Mammostrat score for all patients with ER-positive breast cancer and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts. *Adjusted for age, grade, tumour size, nodes positive, treatment and chemo. Abbreviations: IHC4, IHC4 score; MvL, medium v low risk Mammostrat score; HvL, high v low risk Mammostrat score; CI, confidence interval.

**Table 8.3** Period-Specific Multivariate Cox Regression of IHC4 and Mammostrat score for Exploratory Subgroup Analysis

| | ER-Positive, Endocrine Only and Node-negative | | ER-Positive, Endocrine Only and Node-Positive | | ER-Positive, Any Treatment and Node-Positive | |
|---|---|---|---|---|---|---|
| | 0-5 Years | 5-10 Years | 0-5 Years | 5-10 Years | 0-5 Years | 5-10 Years |
| Main Effect* | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| Edinburgh BCS | (44 events) | (28 events) | (26 events) | (17 events) | (36 events) | (23 events) |
| IHC4 | 1.10 (1.04-1.17) | 1.04 (0.97-1.12) | 0.99 (0.91-1.07) | 0.89 (0.79-1.00) | 1.04 (0.98-1.12) | 0.92 (0.84-1.01) |
| MvL | 2.16 (1.06-4.40) | 1.79 (0.67-4.75) | 1.61 (0.60-4.35) | 1.35 (0.43-4.29) | 1.21 (0.52-2.83) | 2.09 (0.79-5.52) |
| HvL | 1.81 (0.78-4.19) | 2.82 (1.02-7.76) | 1.85 (0.63-5.42) | 0.77 (0.15-4.08) | 1.83 (0.74-4.48) | 0.97 (0.25-3.83) |
| TEAM | (81 events) | (23 events) | (161 events) | (51 events) | (306 events) | (97 events) |
| IHC4 | 1.11 (1.07-1.16) | 1.04 (0.95-1.14) | 1.06 (1.03-1.10) | 1.01 (0.95-1.07) | 1.07 (1.04-1.09) | 1.00 (0.96-1.05) |
| MvL | 1.34 (0.67-2.68) | 1.80 (0.48-6.78) | 1.43 (0.94-2.17) | 0.84 (0.40-1.77) | 1.51 (1.10-2.07) | 1.22 (0.72-2.08) |
| HvL | 2.15 (1.20-3.82) | 3.33 (1.06-10.5) | 1.57 (1.05-2.34) | 0.84 (0.43-1.62) | 1.83 (1.35-2.47) | 1.11 (0.66-1.85) |

NOTE. Multivariate analysis of IHC4 and Mammostrat scores entered separately into a Cox regression model with conventional histopathologic variables for TTDR before 5 years and after 5 years for patients in the exploratory subgroups. Values represent estimated hazard ratios and 95% CIs calculated as a ten-unit increase in IHC4 score and high and medium risk compared to low risk as categorised by Mammostrat score. *Adjusted for age, grade, tumour size, nodes positive, treatment and chemo. Abbreviations: IHC4, IHC4 score; MvL, medium v low risk Mammostrat score; HvL, high v low risk Mammostrat score; CI, confidence interval.

**Table 8.4** Wald tests and performance data for assessing the amount of information added by the IHC4 score or the Mammostrat score or both to clinical factors in the first 5 years of follow-up.

| | Edinburgh BCS | | | TEAM | | |
|---|---|---|---|---|---|---|
| Model | Wald test (p-value) | Increase in $R^2$ (%) | Increase in D Statistic | Wald test (p-value) | Increase in $R^2$ (%) | Increase in D Statistic |
| **All ER-Positive** | | (n=1,103) | | | (n=3,766) | |
| C vs. Null (7 df) | 9.5 (<0.001) | 31.1 | 1.51 | 272.1 (<0.001) | 29 | 1.31 |
| C + IHC4 vs. C (1 df) | 12.1 (<0.001) | 6.6 | **0.12** | 61.2 (<0.001) | 5.0 | **0.16** |
| C + Mam vs. C (2 df) | 3.3 (0.04) | 2.5 | 0.07 | 29.2 (<0.001) | 2.5 | 0.08 |
| C + IHC4 + Mam vs. C (3 df) | 5.5 (<0.001) | 5.4 | **0.18** | 78.5 (<0.001) | 6.4 | **0.20** |
| C + IHC4 + Mam vs. C + IHC4 (2 df) | 2.4 (0.09) | 2.8 | 0.06 | 17.3 (<0.001) | 1.4 | 0.04 |
| C + IHC4 + Mam vs. C + Mam (1 df) | 10.1 (0.002) | 2.9 | **0.11** | 47.4 (<0.001) | 3.9 | **0.12** |
| **ER-Positive Endocrine Only** | | (n=831) | | | (n=2,513) | |
| C vs. Null (6 df) | 11.6 (<0.001) | 35.3 | 1.51 | 172.0 (<0.001) | 30.5 | 1.36 |
| C + IHC4 vs. C (1 df) | 6.4 (0.01) | 3.7 | **0.12** | 34.5 (<0.001) | 4.4 | **0.14** |
| C + Mam vs. C (2 df) | 3.2 (0.04) | 3.2 | **0.11** | 10.5 (0.005) | 1.7 | 0.05 |
| C + IHC4 + Mam vs. C (3 df) | 3.8 (0.01) | 6 | **0.21** | 38.7 (<0.001) | 5 | **0.16** |
| C + IHC4 + Mam vs. C + IHC4 (2 df) | 2.6 (0.07) | 2.3 | 0.09 | 4.4 (0.11) | 0.6 | 0.02 |
| C + IHC4 + Mam vs. C + Mam (1 df) | 5.1 (0.02) | 2.8 | **0.10** | 27.1 (<0.001) | 3.3 | **0.11** |

NOTE. Values represent Wald tests (significance level) and the increase in $R^2$ and D statistic for the addition of IHC4 or Mammostrat score or both. Results are given for ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts with follow-up censored at 5 years. Abbreviations: C, clinical model with age, grade, tumour size, nodes positive, treatment and chemotherapy; Null, null model with no covariates; IHC4, IHC4 score; Mam, Mammostrat score; df, degrees of freedom.

**Table 8.5** Period-Specific Multivariate Cox Regression of IHC4 and Mammostrat score Combined

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | 0-5 Years | 5-10 Years | 0-5 Years | 5-10 Years |
| Main Effect* | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| Edinburgh BCS | (97 events) | (61 events) | (70 events) | (45 events) |
| IHC4 | 1.07 (1.03-1.11) | 0.96 (0.91-1.02) | 1.06 (1.01-1.11) | 0.98 (0.92-1.04) |
| MvL | 1.72 (1.06-2.79) | 1.56 (0.84-2.93) | 1.95 (1.10-3.48) | 1.49 (0.71-3.12) |
| HvL | 1.30 (0.72-2.37) | 1.61 (0.76-3.42) | 1.53 (0.77-3.04) | 1.85 (0.81-4.22) |
| TEAM | (416 events) | (132 events) | (242 events) | (74 events) |
| IHC4 | 1.07 (1.05-1.09) | 1.02 (0.98-1.05) | 1.07 (1.04-1.10) | 1.01 (0.96-1.06) |
| MvL | 1.33 (1.01-1.77) | 1.36 (0.85-2.18) | 1.28 (0.89-1.84) | 0.95 (0.50-1.79) |
| HvL | 1.73 (1.33-2.24) | 1.52 (0.97-2.36) | 1.43 (1.02-2.00) | 1.15 (0.66-2.01) |

NOTE. Multivariate analysis of IHC4 and Mammostrat scores entered simultaneously into a Cox regression model with conventional histopathologic variables for TTDR before 5 years and after 5 years for patients in different subgroups. Values represent estimated hazard ratios and 95% CIs calculated as a ten-unit increase in IHC4 score and high and medium risk compared to low risk as categorised by Mammostrat score for all patients with ER-positive breast cancer and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts. *Adjusted for age, grade, tumour size, nodes positive, treatment and chemo. Abbreviations: IHC4, IHC4 score; MvL, medium v low risk Mammostrat score; HvL, high v low risk Mammostrat score; CI, confidence interval.

**Table 8.6** Period-Specific Multivariate Cox Regression of IHC4 and Mammostrat score Combined for Exploratory Subgroup Analysis

| | ER-Positive, Endocrine Only and Node-negative | | ER-Positive, Endocrine Only and Node-Positive | | ER-Positive, Any Treatment and Node-Positive | |
|---|---|---|---|---|---|---|
| | 0-5 Years | 5-10 Years | 0-5 Years | 5-10 Years | 0-5 Years | 5-10 Years |
| Main Effect* | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| Edinburgh BCS | (44 events) | (28 events) | (26 events) | (17 events) | (36 events) | (23 events) |
| IHC4 | 1.10 (1.03-1.17) | 1.03 (0.96-1.11) | 0.97 (0.89-1.06) | 0.89 (0.78-1.00) | 1.03 (0.96-1.11) | 0.92 (0.83-1.02) |
| MvL | 2.08 (1.01-4.25) | 1.68 (0.62-4.54) | 1.59 (0.59-4.29) | 1.99 (0.64-6.19) | 1.20 (0.51-2.82) | 3.06 (1.17-7.98) |
| HvL | 1.48 (0.63-3.48) | 2.79 (1.01-7.71) | 2.16 (0.69-6.74) | 1.38 (0.21-8.85) | 1.58 (0.59-4.21) | 1.30 (0.28-5.95) |
| TEAM | (81 events) | (23 events) | (161 events) | (51 events) | (306 events) | (97 events) |
| IHC4 | 1.10 (1.06-1.15) | 1.03 (0.94-1.12) | 1.06 (1.02-1.09) | 1.01 (0.95-1.08) | 1.06 (1.04-1.08) | 1.00 (0.96-1.05) |
| MvL | 1.20 (0.60-2.40) | 1.78 (0.47-6.67) | 1.31 (0.86-2.01) | 0.82 (0.39-1.75) | 1.41 (1.03-1.93) | 1.22 (0.71-2.08) |
| HvL | 1.66 (0.92-3.00) | 3.20 (1.01-10.1) | 1.36 (0.90-2.05) | 0.81 (0.41-1.61) | 1.63 (1.20-2.20) | 1.10 (0.66-1.86) |

NOTE. Multivariate analysis of IHC4 and Mammostrat scores entered simultaneously into a Cox regression model with conventional histopathologic variables for TTDR before 5 years and after 5 years for patients in different subgroups. Values represent estimated hazard ratios and 95% CIs calculated as a ten-unit increase in IHC4 score and high and medium risk compared to low risk as categorised by Mammostrat score for patients with ER-positive breast cancer treated with endocrine therapy only and either node-positive or node-negative and ER-positive, node-positive patients irrespective of treatment in the Edinburgh BCS and TEAM cohorts. *Adjusted for age, grade, tumour size, nodes positive, treatment and chemo. Abbreviations: IHC4, IHC4 score; MvL, medium v low risk Mammostrat score; HvL, high v low risk Mammostrat score; CI, confidence interval.

#### *8.2.1.3 Comparison of IHC4 and Mammostrat*

The IHC4 score provided more prognostic information beyond that of clinical factors compared to the Mammostrat score for all ER-positive patients in both patient cohorts in the first five years of follow-up (Table 8.4; increase in $R^2$: 6.6% vs 2.5% and D-statistic: 0.12 vs 0.07 in the Edinburgh BCS cohort and increase in $R^2$: 5.0% vs 2.5% and D-statistic: 0.16 vs 0.08 in the TEAM cohort). Likewise for ER-positive patients treated with endocrine therapy in the TEAM cohort, the IHC4 score was the stronger predictor of outcome whereas in the Edinburgh cohort the prognostic information provided by either score was similar (increase in $R^2$: 3.7% vs 3.2% and D-statistic: 0.12 vs 0.11).

#### *8.2.1.4 The addition of both scores to clinical factors*

The scores were entered simultaneously into a multivariate Cox regression model and in the first 5 years of follow-up, the addition of both scores to clinical factors provided statistically significant information ($p<0.05$) for both subsets of patients across both cohorts with increases in $R^2$ between 5 and 6% and increases in D-statistic between 0.16 and 0.21 (Table 8.4). However, both scores only remained significant independent predictors of TTDR restricted to the first 5 years of follow-up when simultaneously entered into a multivariate Cox regression model for all ER-positive patients in the TEAM cohort with a HR for a ten-unit increase in ICH4 score of 1.07 (95% CI 1.05-1.09) and HRs for medium and high versus low Mammostrat score of 1.3 (95% CI 1.0-1.8) and 1.7 (95% CI 1.3-2.2) respectively (Table 8.5). Only the IHC4 score provided significant independent prognostic information on TTDR in the first 5 years of follow-up for all ER-positive patients treated with endocrine therapy in the TEAM cohort and both patient subgroups in the Edinburgh BCS cohort. Although not statistically significant, the Mammostrat score provided some improvement in model discrimination over and above that provided by IHC4 score and clinical factors with an increase in $R^2$ and D-statistic of 2.3% and 0.09 respectively for ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort. There was evidence of an effect of Mammostrat high risk versus low risk after 5 years of survival after adjustment for IHC4 and clinical factors for ER-

positive, node-negative patients treated with endocrine therapy only in the TEAM and Edinburgh BCS cohorts with HRs of 3.2 (95% CI 1.0-10.1) and 2.8 (95% CI 1.0-7.7) respectively (Table 8.6).

#### *8.2.1.5 Test of PH assumption*

Despite the categorisation of the time interval, it is still important to test the PH assumption within these time intervals. If the interval is too large, violation of the PH assumption may still occur and give biased results.

For these data, there was no evidence to reject the PH assumption in each 5 year time interval as indicated by the Grambsch-Therneau test of proportional hazards (data not shown).

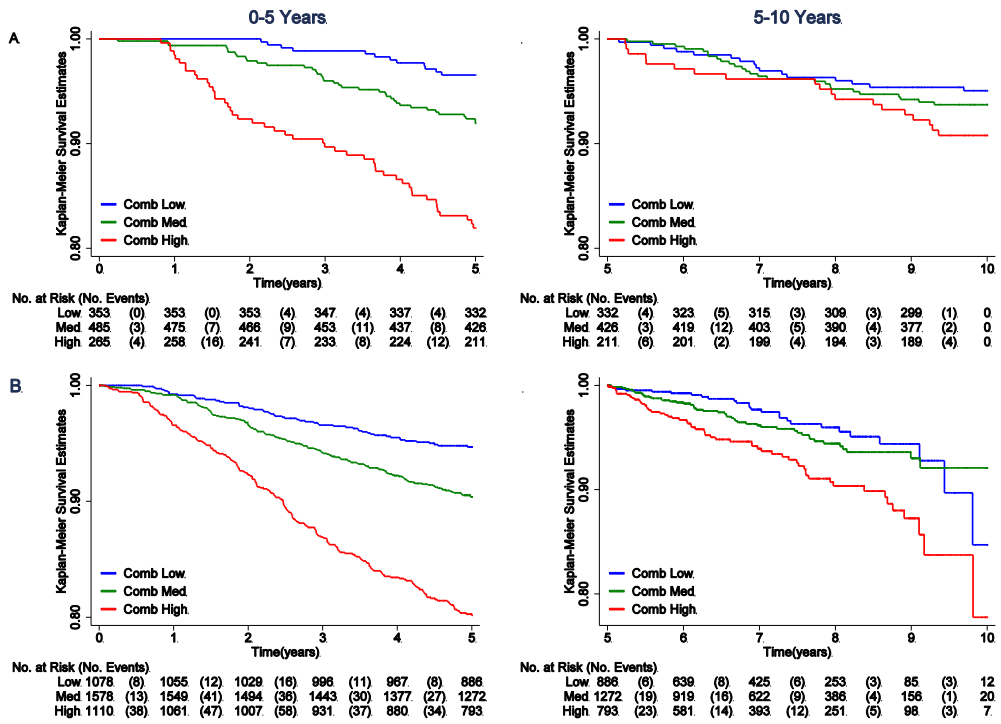
#### *8.2.1.6 Combined IHC4 and Mammostrat Risk Groups*

In section 7.3.2, three risk groups from the combined IHC4 and Mammostrat scores were identified: low risk, IHC4 and Mammostrat low risk; medium risk, IHC4 low risk and Mammostrat high risk and vice versa; high risk, IHC4 and Mammostrat high risk.

Period-specific K-M curves for this combined score are shown for all ER-positive patients and ER-positive patients treated with endocrine therapy in Figure 8.5. Very good separation of the survival curves was observed in the first 5 years of follow up in both the Edinburgh BCS and TEAM cohorts. Good stratification also remained in the 5-10 year interval for all ER-positive patients in the TEAM cohort.

The combined score was an independent predictor of time to distant-recurrence in the first five years of follow-up for both subgroups in the Edinburgh BCS and TEAM cohorts (Table 8.7). Patients with high risk had an increased hazard of 3.3 (95% CI, 1.7-6.5) and 2.9 (95% CI, 2.1-4.0) times that of patients with low risk for all ER-

## All ER-Positive




## ER-Positive, Endocrine Only only

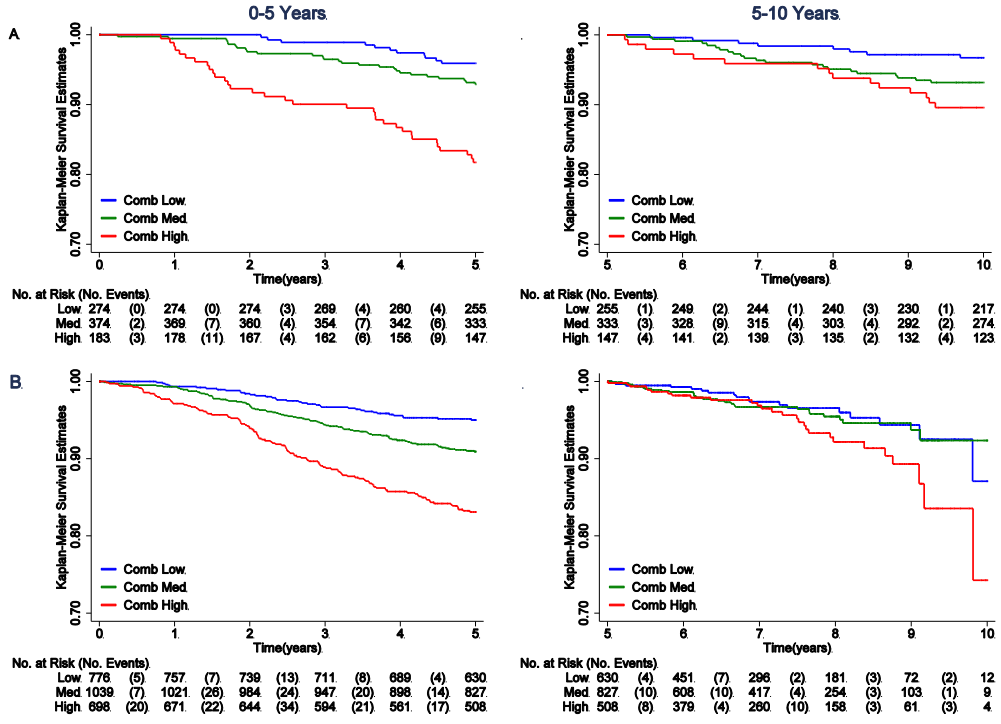



**Figure 8.5** Period-specific Kaplan-Meier curves for time to distant recurrence for the combined score for all ER-positive patients and ER-Positive patients treated with endocrine therapy only. Plots represent (A) Edinburgh BCS and (B) TEAM cohorts. Curves are shown for the period 0-5 years, censoring follow-up of all patients at 5 years after diagnosis. The 5-10 year interval is assessed from the subset of patients who remained distant-recurrence free for at least 5 years and censoring follow-up of patients at 10 years.

**Table 8.7** Period Specific Multivariate Cox Regression of the Combined Score

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | 0-5 Years | 5-10 Years | 0-5 Years | 5-10 Years |
| Main Effect* | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| Edinburgh BCS | (97 events) | (95 events) | (70 events) | (72 events) |
| Comb MvL | 1.94 (1.00-3.77) | 1.32 (0.66-2.65) | 1.50 (0.72-3.10) | 1.87 (0.81-4.28) |
| Comb HvL | 3.32 (1.69-6.53) | 1.22 (0.55-2.70) | 3.14 (1.51-6.54) | 1.75 (0.67-4.57) |
| TEAM | (416 events) | (132 events) | (242 events) | (74 events) |
| Comb MvL | 1.63 (1.18-2.27) | 1.05 (0.64-1.73) | 1.73 (1.15-2.61) | 0.91 (0.49-1.67) |
| Comb HvL | 2.91 (2.10-4.03) | 1.81 (1.10-2.97) | 2.69 (1.79-4.06) | 1.34 (0.72-2.50) |

NOTE: Multivariate analysis of Combined score with conventional histopathologic variables for TTDR before 5 years and after 5 years. Values represent estimated hazard ratios and 95% CIs calculated as high and medium risk compared to low risk as categorised by the combined score *Adjusted for age, grade, tumour size, nodes positive, treatment and chemo. Abbreviations: Comb, combined score; MvL, medium v low risk; HvL, high v low risk; CI, confidence interval.

**Table 8.8** Period-Specific Multivariate Cox Regression of the Combined Score for Exploratory Subgroup Analysis

| | ER-Positive, Endocrine Only and Node-Negative | | ER-Positive, Endocrine Only and Node-Positive | | ER-Positive, Any Treatment and Node-Positive | |
|---|---|---|---|---|---|---|
| | 0-5 Years | 5-10 Years | 0-5 Years | 5-10 Years | 0-5 Years | 5-10 Years |
| Main Effect* | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| Edinburgh BCS | (44 events) | (28 events) | (26 events) | (17 events) | (36 events) | (23 events) |
| Comb MvL | 2.43 (0.79-7.44) | 1.87 (0.81-4.28) | 0.95 (0.34-2.66) | 0.97 (0.33-2.84) | 1.23 (0.50-3.04) | 0.98 (0.38-2.55) |
| Comb HvL | 5.70 (1.88-17.31) | 1.75 (0.67-4.57) | 1.46 (0.48-4.56) | 0.31 (0.05-1.89) | 1.64 (0.62-4.35) | 0.65 (0.18-2.37) |
| TEAM | (81 events) | (23 events) | (161 events) | (51 events) | (306 events) | (97 events) |
| Comb MvL | 2.05 (0.93-4.55) | 2.77 (0.60-12.79) | 1.67 (1.03-3.69) | 0.69 (0.34-1.40) | 1.55 (1.07-2.25) | 0.82 (0.47-1.42) |
| Comb HvL | 3.51 (1.60-7.69) | 3.86 (0.82-18.19) | 2.49 (1.53-4.04) | 1.08 (0.53-2.20) | 2.75 (1.90-3.97) | 1.40 (0.80-2.44) |

NOTE: Multivariate analysis of Combined score with conventional histopathologic variables for TTDR before 5 years and after 5 years for patients in the exploratory subgroups. Values represent estimated hazard ratios and 95% CIs calculated as high and medium risk compared to low risk as categorised by the combined score *Adjusted for age, grade, tumour size, nodes positive, treatment and chemo. Abbreviations: Comb, combined score; MvL, medium v low risk; HvL, high v low risk; CI, confidence interval.

positive patients in the Edinburgh BCS and TEAM cohorts respectively. For all ER-positive patients in the TEAM cohort, the combined score remained a significant independent predictor of survival in the 5-10 year interval (p=0.01). Patients with high risk had 1.8 (95% CI, 1.1-3.0) times the hazard of patients with a low risk combined score.

*Exploratory Subgroups*

In the exploratory subgroups (Table 8.8), the combined score was an independent predictor of time-to-distant recurrence restricted to the first five years of follow-up in all exploratory subgroups in the TEAM cohort. In the Edinburgh BCS cohort, the combined score was an independent predictor of outcome in the first five years of follow-up in those ER-positive, node-negative patients treated with endocrine therapy only. Although not statistically significant, in node-positive subgroups in the Edinburgh BCS cohort, the combined score showed a trend towards a protective effect of higher scores in the 5-10 year interval with hazard ratio estimates of 0.3 (95% CI, 0.1-1.9) and 0.7 (95% CI, 0.2-2.35) for ER-positive, node-positive patients treated with endocrine therapy and any treatment respectively.

#### *8.2.1.7 Two-Year Time Intervals*

The 5 year intervals are clinically relevant, however perhaps smaller intervals could provide more insight into the pattern over time.

The estimated hazard ratios for all ER-positive patients for two year intervals are displayed in Figure 8.6. The smaller time intervals results in a significantly reduced number of events in each interval and therefore wide confidence intervals, especially in the smaller Edinburgh BCS cohort. The IHC4 and Mammostrat scores remained significant predictors of TTDR up to 6 years in the TEAM cohort, after which the 95% confidence interval crosses the value of 1, corresponding to a null effect. In the Edinburgh BCS cohort, the IHC4 score is only

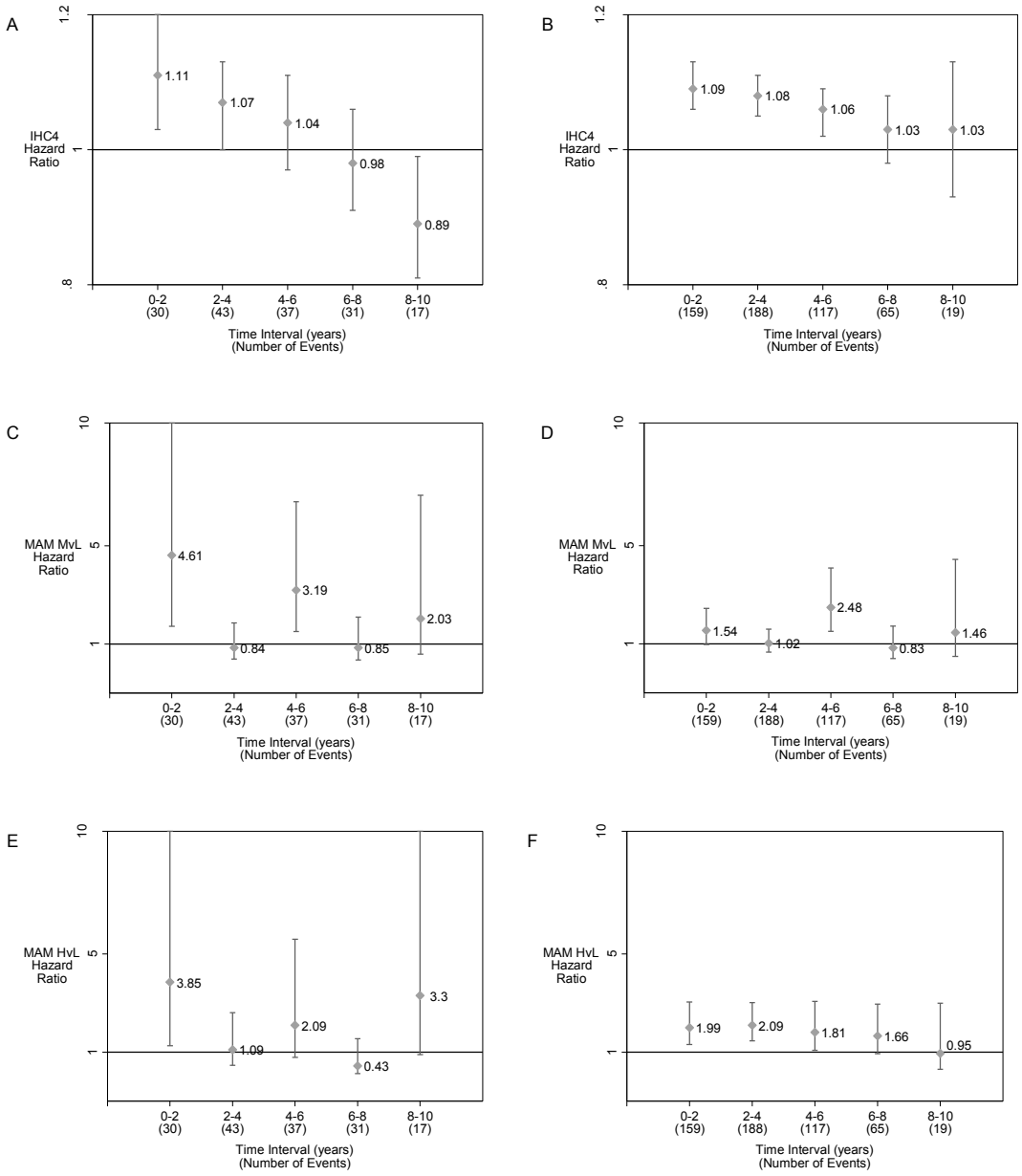



**Figure 8.6** All ER-positive patients. Two-yearly adjusted hazard ratio estimates and 95% confidence intervals for a ten-unit increase in IHC4 score and medium and high Mammostrat score compared to low Mammostrat score. Plots represent (left-column) Edinburgh BCS cohort and (right column) TEAM cohort. Adjusted for age, grade, tumour size, number of positive nodes, treatment and chemotherapy.

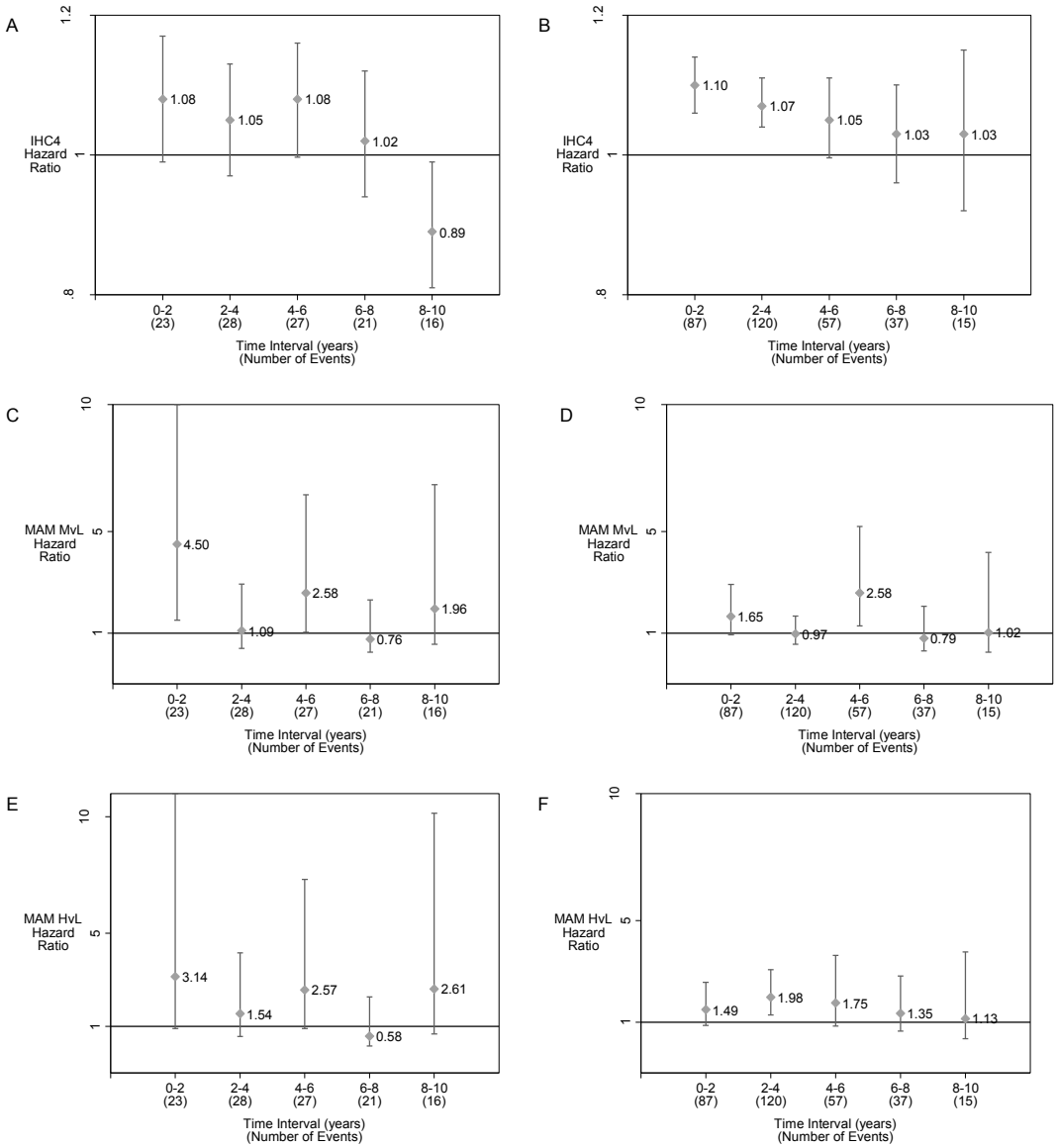



**Figure 8.7** ER-positive patients treated with endocrine therapy only. Two-yearly adjusted hazard ratio estimates and 95% confidence intervals for a ten-unit increase in IHC4 score and medium and high Mammostrat score compared to low Mammostrat score. Plots represent (left-column) Edinburgh BCS cohort and (right column) TEAM cohort. Adjusted for age, grade, tumour size, number of positive nodes, treatment and chemotherapy.

statistically significant in the first four years of follow-up. Up to 6 years, the effect of ICH4 is estimated to be above 1, i.e. higher IHC4 score is associated with an increased risk in TTDR. After 6 years and especially in the 8-10 year period, the HR is estimated to be below one indicating a reduced risk of TTDR with higher IHC4 score (HR (95% CI): 0.9 (0.8-1.0)). Mammostrat score was prognostic of outcome in the 0-2 year and 4-6 year time intervals only. However wide confidence intervals were associated with the Mammostrat score estimates in the Edinburgh BCS cohort.

For all ER-positive patients treated with endocrine therapy only (Figure 8.7), IHC4 score was not prognostic for any of the two-year time intervals in the Edinburgh BCS cohort except for the final 8-10 year interval where higher scores were associated with a reduced hazard, which was also seen in all ER-positive patients. IHC4 score was prognostic up to 4 years in the TEAM cohort. Mammostrat score was prognostic up to two years in the Edinburgh BCS cohort and between 2-6 years in the TEAM cohort.

Due to the already small sizes of the exploratory subgroups, two-year intervals were not considered.

*Combined Score*

The estimated hazard ratios for the two-yearly intervals are displayed in Figure 8.8 for all ER-positive patients and ER-positive patients treated with endocrine therapy only. The combined score was prognostic in the 4-6 year time interval only in the Edinburgh BCS cohort for both patient subgroups. Due to those assigned to low risk having no events in the first two-year time interval, hazard ratios were not estimated for this time period. In the TEAM cohort, the combined score was prognostic up to 6 years in both patient subgroups. The effect was strongest in the 4-6 year interval with hazard ratio estimates of 4.1 (95% CI, 2.1-7.7) and 5.3 (95% CI, 2.0-14.0) for all ER-positive patients and ER-positive patients treated with endocrine therapy respectively. However, the confidence intervals are wide.

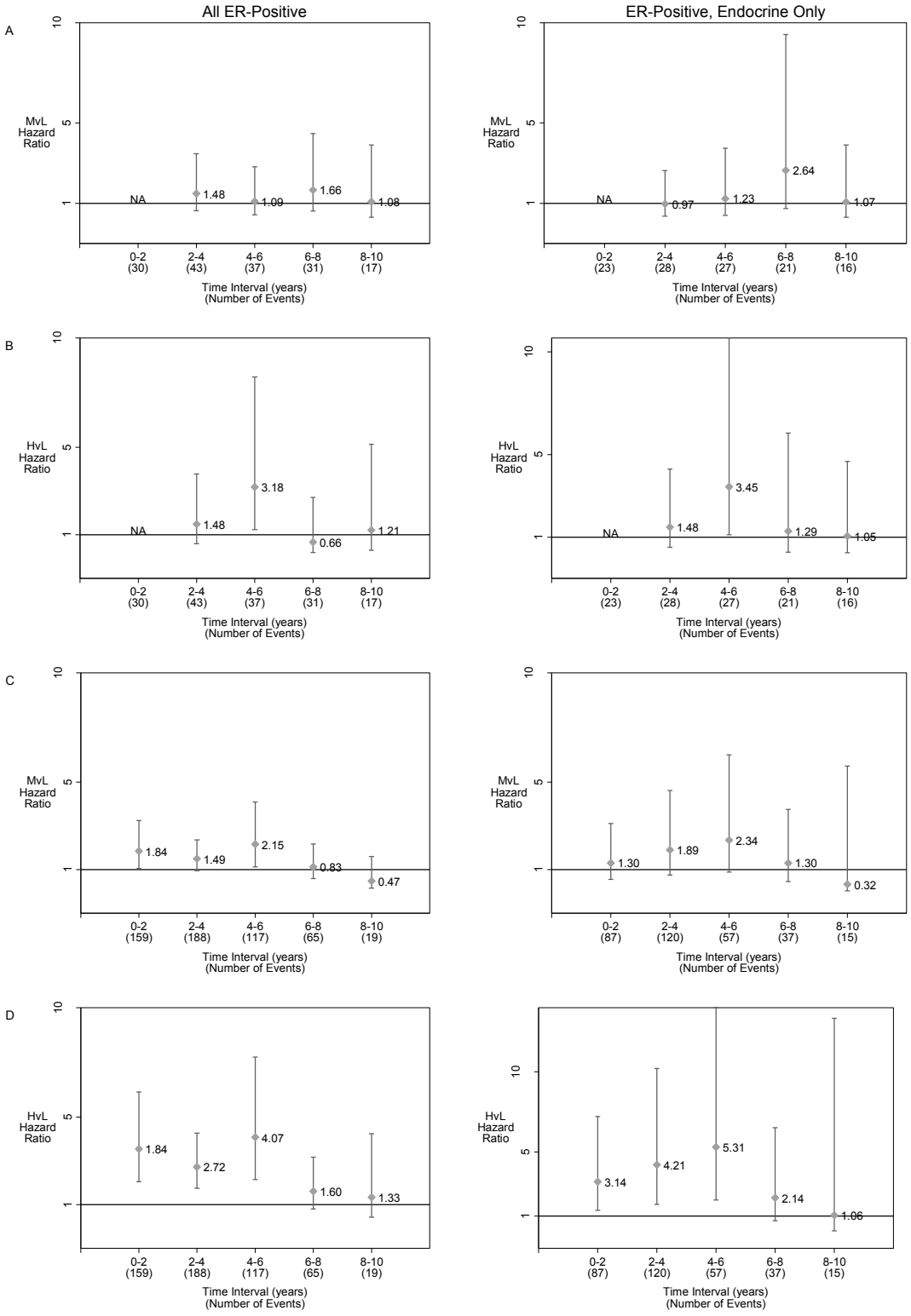



**Figure 8.8** Combined Score. Two-yearly adjusted hazard ratio estimates and 95% confidence intervals for medium and high risk combined score compared to low risk combined score for all ER-positive patients (left-column) and ER-positive patients treated with endocrine therapy only (right-column). Plots represent (A and B) Edinburgh BCS cohort and (C and D) TEAM cohort. Adjusted for age, grade, tumour size, number of positive nodes, treatment and chemotherapy. Not applicable (NA) due to zero events in at least one risk group of the combined score.

### 8.2.2 Summary

A summary of whether IHC4 or Mammostrat were significant predictors of outcome in the 0-5 year and 5-10 year intervals is displayed in Table 8.9. In the first five years of follow-up, IHC4 and Mammostrat were independent predictors of outcome when analysed in separate Cox regression models for all ER-positive patients and ER-positive patients treated with endocrine therapy only.

**Table 8.9** Significant predictors of TTDR in the 0-5 and 5-10 year time intervals by patient subgroup

| | 0-5 Years | | | | 5-10 Years | | | |
|---|---|---|---|---|---|---|---|---|
| | Edinburgh BCS | | TEAM | | Edinburgh BCS | | TEAM | |
| | IHC4 | MAM | IHC4 | MAM | IHC4 | MAM | IHC4 | MAM |
| **Separate IHC4 & MAM** | | | | | | | | |
| All ER+ | ✓ | ✓ | ✓ | ✓ | × | × | × | × (HvL) |
| ER+, Tam Only | ✓ | ✓ | ✓ | ✓ | × | × | × | × |
| ER+, Tam Only, N- | ✓ | ✓ | ✓ | ✓ | × | × (HvL) | × | × (HvL) |
| ER+, Tam Only, N+ | × | × | ✓ | ✓ | × | × | × | × |
| ER+, N+ | × | × | ✓ | ✓ | × | × | × | × |
| **Simulatenous IHC4 & MAM** | | | | | | | | |
| All ER+ | ✓ | × (MvL) | ✓ | ✓ | × | × | × | × |
| ER+, Tam Only | ✓ | × (MvL) | ✓ | × (HvL) | × | × | × | × |
| ER+, Tam Only, N- | ✓ | × (MvL) | ✓ | × | × | × (HvL) | × | × (HvL) |
| ER+, Tam Only, N+ | × | × | ✓ | × | × | × | × | × |
| ER+, N+ | × | × | ✓ | ✓ | × | × (MvL) | × | × |

Note. Summary of whether IHC4 or Mammostrat (MAM) were independent predictors of TTDR after adjustment for age, grade, tumour size, grade, treatment and chemotherapy. IHC4 & Mammostrat were entered seperately or simultaneously into a Cox regression model. Results given for the main analysis subgroups (all ER-positive (ER+) and all ER-positive treated with tamoxifen (tam) only) and exploratory subgroups (ER-positive, treated with tamoxifen only and either node-negative (N-) or node-positive (N+) or ER-positive, node-positive irresepctive of treatment). Where Mammostrat score had an overall non-significant effect but individual level comparisons (medium vs. low risk (MvL) or high vs.low risk (HvL) were at a significantly increased risk are illsutrated in brackets by (MvL) or (HvL).

Both scores remained independent predictors of outcome when entered simultaneously into a Cox regression model in the first five years of follow-up for all ER-positive patients in the TEAM cohort only. Despite Mammostrat not having an overall significant prognostic effect in the Edinburgh BCS cohort and ER-positive patients treated with endocrine therapy only in the TEAM cohort, those medium risk had a significantly increased hazard compared to those low risk in the Edinburgh BCS cohort and those high risk had a significantly increased hazard compared to those low risk in the TEAM cohort.

Despite a sufficient number of events in the 5-10 year interval neither IHC4 nor Mammostrat score remained an independent predictor of outcome after 5 years of follow-up. However, those classified as Mammostrat high risk were at a significantly increased hazard compared to low risk for all ER-positive patients in the TEAM cohort, and for ER-positive patients treated with endocrine therapy only and node-negative in the Edinburgh BCS and TEAM cohorts.

## 8.3 MFPT Approach

The next approach that was considered was the MFPT approach described in section 4.4.4.

Step 3 of the MFPT approach investigates possible non-proportional effects of variables. For each covariate, the FPT algorithm determines the best fitting FP function to model the time-varying effect adjusting for all other covariates in the model. Then a forward selection procedure is used to add any significant time varying effects to the model.

This was performed using the *stmfpt* command in STATA.

### 8.3.1 Splitting the data

The Cox model is fitted by evaluating the partial likelihood at each failure time. Because the time-varying effects are a function of study time, the observations under risk are split into episodes at specified time points, ideally at each event time.

As the investigation of time-varying effects is most sensible with long-term follow-up and requires large data sets in order to have some power to detect an interaction with time, splitting at each event time may cause technical problems. Sauerbrei 'categorised' survival times in half-year periods up to year 15 and a final period >15 years (Sauerbrei et al., 2007). Table 8.10 gives an example of a patient with an event at 1.8 years. Splitting the data into half year intervals results in creating four observations for this patient with an event in period 3.

**Table 8.10** Example of splitting data in half year intervals

| | ID | Gender | Tumour Size | Entry Time | Exit Time | Event Status | Period |
|---|---|---|---|---|---|---|---|
| Original Data | 1 | M | 10 | 0 | 1.8 | 1 | NA |
| Splitting Data Produces | 1 | M | 10 | 0 | 0.5 | 0 | 0 |
| | 1 | M | 10 | 0.5 | 1 | 0 | 1 |
| | 1 | M | 10 | 1 | 1.5 | 0 | 2 |
| | 1 | M | 10 | 1.5 | 2 | 1 | 3 |

We considered categorisation for half-year periods and also smaller intervals (Table 8.11). The first time-period was never considered smaller than 6 months (due to convergence issues) and a final period of >20 years for the Edinburgh BCS cohort. This creates an extra 18,501 and 44,773 observations in the Edinburgh BCS and TEAM cohorts respectively for half year periods up to year 10.

**Table 8.11** Number of observations created when categorising the time interval

| | Interval | | | | | |
|---|---|---|---|---|---|---|
| | 10-Year Follow-up | | | 25-Year Follow-Up | | |
| | 6 months | 2 months | Every Failure | 6 months | 3 months | Every Failure |
| Edinburgh BCS (n=1,103) | 18,501 (20) | 55,231 (58) | 149,524 (152) | 26,197 (41) | 51,825 (80) | 166,441 (186) |
| TEAM (n=3,766) | 44,773 (20) | 130,257 (58) | 1,478,122 (493) | NA | NA | NA |

Note. Numbers represent number of observations created (number of distinct observed times). Initial period always 0-6 months and a final period of >20 years. NA, not applicable.

## 8.3.2 Results

### *8.3.2.1 10 year follow-up*

In multivariable modelling restricting follow-up to 10 years, IHC4 was determined to have a significant time-by-covariate interaction for all ER-positive patients in both cohorts with the best fitting FP to be log of time (Figure 8.9). The parameter associated with this interaction was negative, suggesting the effect of a unit increase in IHC4 on TTDR decreased over time (Figure 8.10). The decrease over time was more prominent in the Edinburgh BCS cohort,

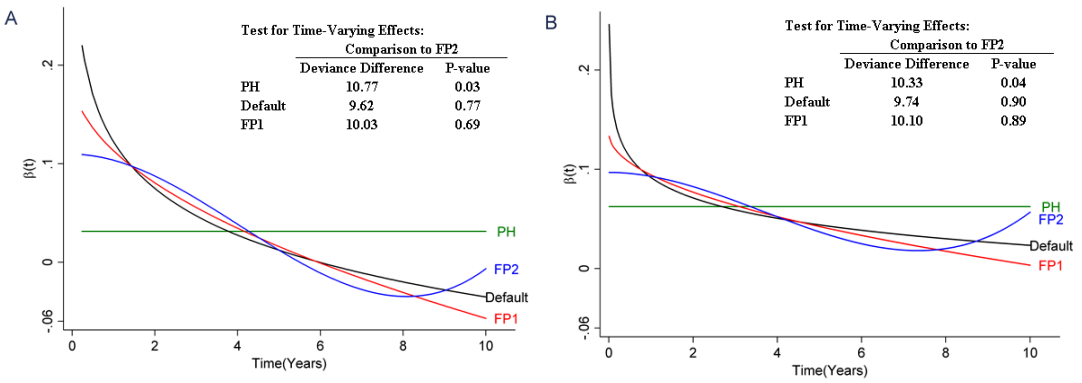



**Figure 8.9** Results of the FPT algorithm for ICH4 score , showing the time-varying effects for proportional hazards (PH), the best fitting FP1 and FP2 model and the default (log (time)) model. Plots represent all ER-positive patients in the (A) Edinburgh BCS and (B) TEAM cohorts.

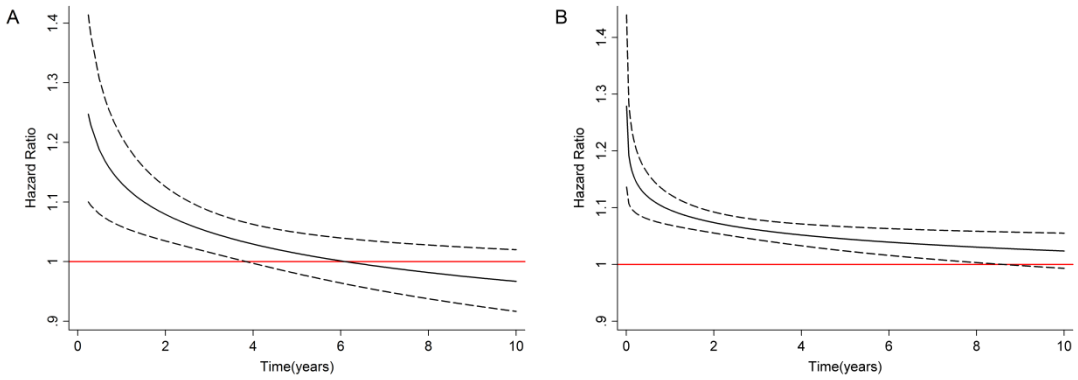



**Figure 8.10** All ER-positive patients, IHC4 Score. Time-dependent adjusted hazard ratio estimate (up to 10 years) with 95% CIs (dashed lines) for a ten-unit increase in IHC4 score for all ER-positive patients in the (A) Edinburgh BCS cohort and (B) TEAM cohorts. Adjusted for age, grade, nodes positive, treatment and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

with the adjusted HR crossing the value 1 (corresponding to a null effect) at approximately 6 years.

IHC4 was also found to have significant time-by-covariate interaction for ER-positive patients treated with endocrine therapy only in the TEAM cohort. However, this significant interaction was only found when smaller time intervals were chosen. This highlights the lack of distinct event times when categorising into half year intervals (only 22

unique failure times). Buchholz found that the MFPT procedure gives similar results for small categorisation intervals and propose that about 50 to 100 distinct event times can give results with sufficient precision (Buchholz, 2010). Even with smaller time-categorisations, IHC4 was not identified to have a significant time-varying effect for ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort resulting in an adjusted hazard ratio estimate of 1.03 (95% CI, 1.00-1.07) for a ten-unit increase in IHC4.

The FPT algorithm did not identify any non-proportional effects of variables when looking at the Mammostrat model (Mammostrat and clinical factors) in both cohorts with follow-up restricted to 10 years (Figure 8.11) with hazard ratio estimates given in Table 8.12.

*Both Scores*

When IHC4 and Mammostrat score were simultaneously entered into a model, the time-varying effect of IHC4 score remained for both cohorts for all ER-positive patients. IHC4 was prognostic of outcome up to 4 and 7 years (where the lower confidence bound crosses 1) for all ER-positive patients in the Edinburgh BCS and TEAM cohorts respectively, after adjustment for Mammostrat score and clinical factors (Figure 8.12). A time-varying effect of

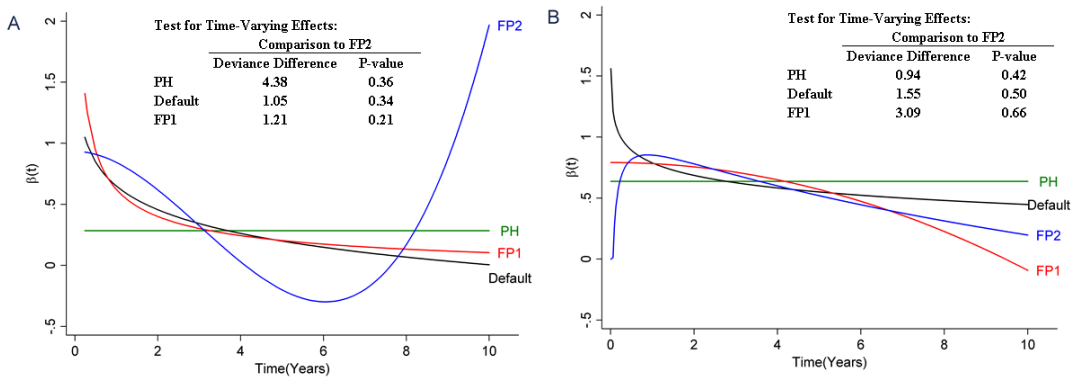



**Figure 8.11** Results of the FPT algorithm for Mammostrat high risk , showing the time-varying effects for proportional hazards (PH), the best fitting FP1 and FP2 model and the default (log (time)) model. Plots represent all ER-positive patients in the (A) Edinburgh BCS and (B) TEAM cohorts.

**Table 8.12** Multivariate Cox regression of Mammostrat score up to 10-years follow-up

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Edinburgh BCS (n=1,103) | TEAM (n=3,766) | Edinburgh BCS (n=831) | TEAM (n=2,513) |
| Main Effect | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| MvL | 1.59 (1.13-2.24) | 1.42 (1.12-1.81) | 1.79 (1.14-2.82) | 1.28 (0.94-1.75) |
| HvL | 1.31 (0.86-2.00) | 1.89 (1.52-2.36) | 1.79 (1.08-2.97) | 1.57 (1.18-2.07) |

NOTE. Multivariate analysis for Mammostrat score with follow-up censored at 10 years. Hazard ratios for risk of TTDR calculated for either high-risk or moderate-risk Mammostrat score compared with low Mammostrat scores at baseline. Values represent estimated hazard ratios and 95% CIs for all ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts. Abbreviations: TTDR, time to distant-recurrence; CI, confidence interval; HvL, high risk v low risk Mammostrat score; MvL, moderate risk v low risk Mammostrat score. * Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

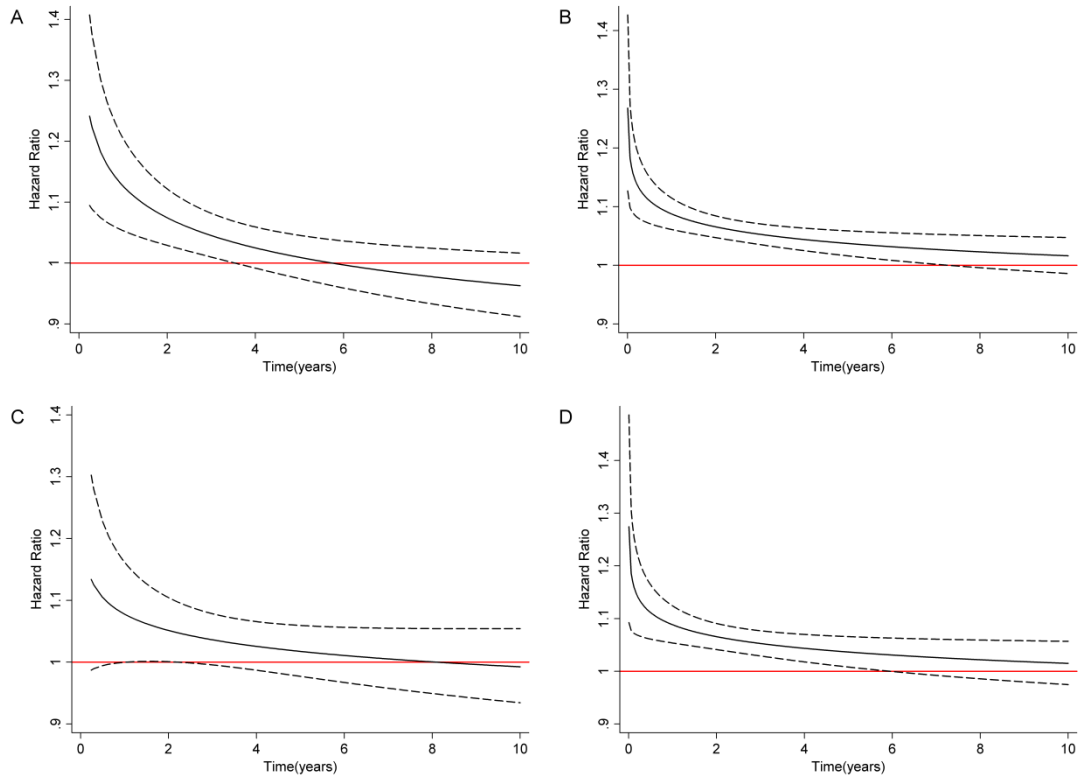



**Figure 8.12** IHC4 Score – Simultaneous Analysis. Time-dependent adjusted hazard ratio estimate (up to 10 years) with 95% CIs (dashed lines) for a ten-unit increase in IHC4 score for all ER-positive patients (A and B) and ER-positive patients treated with endocrine therapy only (C and D) in the Edinburgh BCS (A and C) and TEAM (B and D) cohorts. Adjusted for Mammostrat score, age, grade, nodes positive, treatment and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

IHC4 was identified for ER-positive patients treated with endocrine therapy only in the TEAM cohort and was prognostic up to 6 years of follow-up. However no time-varying effects were identified for ER-positive patients treated with endocrine therapy in the Edinburgh BCS cohort and IHC4 was not an independent predictor of outcome (HR 1.0, 95% CI 1.0-1.1). Table 8.13 gives the hazard ratio estimates for Mammostrat score after adjustment for time-varying IHC4 score. Mammostrat score was an independent predictor of outcome in the first 10-years of follow-up for both patient subgroups in the Edinburgh BCS cohort and all ER-positive patients in the TEAM cohort. Mammostrat score was not an independent predictor of outcome after adjustment for time-varying IHC4 score for ER-

**Table 8.13** Multivariate Cox regression of Mammostrat score after adjustment for time-varying IHC4 score up to 10-years follow-up

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Edinburgh BCS (n=1,103) | TEAM (n=3,766) | Edinburgh BCS (n=831) | TEAM (n=2,513) |
| Main Effect | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| MvL | 1.67 (1.14-2.45) | 1.34 (1.05-1.71) | 1.78 (1.13-2.80) | 1.19 (0.87-1.63) |
| HvL | 1.37 (0.85-2.21) | 1.67 (1.34-2.09) | 1.62 (0.95-2.76) | 1.35 (1.01-1.81) |

NOTE. Multivariate analysis for Mammostrat score with follow-up censored at 10 years. Hazard ratios for risk of TTDR calculated for either high-risk or moderate-risk Mammostrat score compared with low Mammostrat scores at baseline. Values represent estimated hazard ratios and 95% CIs for all ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts. Abbreviations: TTDR, time to distant-recurrence; CI, confidence interval; HvL, high risk v low risk Mammostrat score; MvL, moderate risk v low risk Mammostrat score. * Adjusted for time-varying IHC4 score, age, grade, tumour size, nodes positive, treatment and chemotherapy.

positive patients treated with endocrine therapy only in the TEAM cohort, however those classified as high risk compared to low risk had an increased hazard of 1.4 (95% CI, 1.0-1.8).

#### *8.3.2.2 Full follow-up*

Performing the FPT algorithm on the Edinburgh BCS cohort with full follow-up (max 25 years), ICH4 score was determined to have a non-proportional effect in both patient subgroups (Figure 8.13). For all ER-positive patients, again we see the hazard ratio crossing the value 1, but after approximately 11 years there appears to be a significantly reduced hazard of TTDR corresponding with higher scores.

Mammostrat score was determined to have a significant time-by-covariate interaction for high risk versus low risk. There was uncertainty in the best fitting FP to model the interaction, with log-time chosen for all ER-positive patients and time-cubed chosen for ER-positive patients treated with endocrine therapy only (Figure 8.14). The hazard ratio comparing high risk patients to low risk as categorised by Mammostrat score crosses the value 1 at approximately 7 and 9 years for all ER-positive patients and ER-positive patients treated with endocrine therapy only respectively. However, the confidence intervals are wide due to the small number of high risk patients in the Edinburgh BCS cohort.

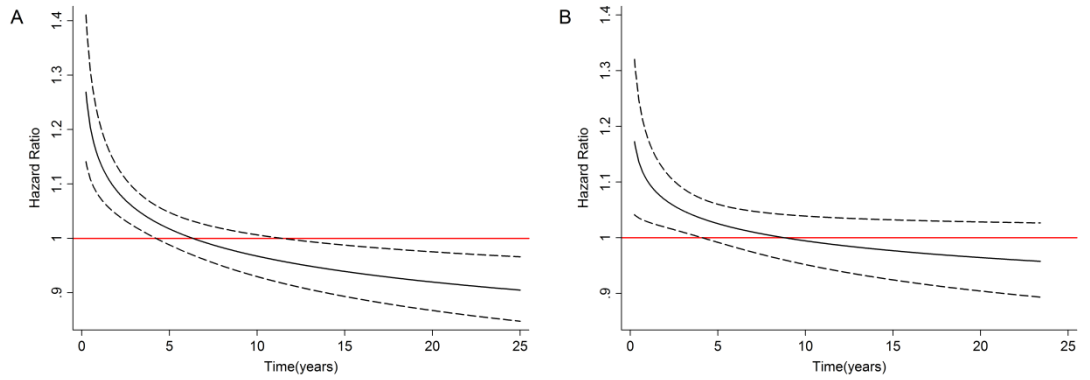



**Figure 8.13** IHC4 Score. Time-dependent adjusted hazard ratio estimate (full follow-up) with 95% CIs (dashed lines) for a ten-unit increase in IHC4 score for (A) all ER-positive patients and (B) ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort. Adjusted for age, grade, nodes positive and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

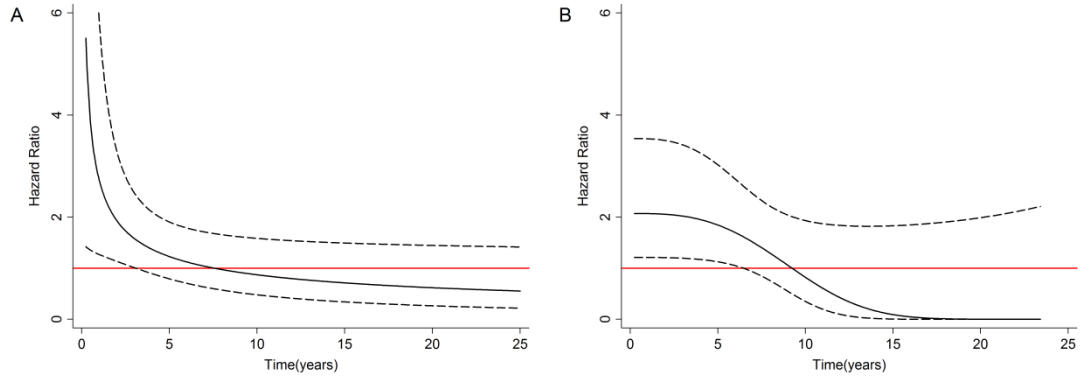



**Figure 8.14** Mammostrat Score. Time-dependent adjusted hazard ratio estimate (full follow-up) with 95% CIs (dashed lines) for high risk compared to low risk Mammostrat score in the Edinburgh BCS cohort for (A) all ER-positive patients and (B) ER-positive patients treated with endocrine therapy only. Adjusted for age, grade, nodes positive and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

*Both Scores*

When IHC4 and Mammostrat were entered simultaneously into a model, only IHC4 score was identified as having a time-varying effect for all ER-positive patients (Figure 8.15A). Suggesting the time-varying effect of high risk Mammostrat score was due to omission of important variables (i.e. ER, PgR, Ki67 and HER2 included in the IHC4 score). Higher IHC4 scores were associated with an increased hazard up to 4 years and then a protective effect for higher scores beyond 11 years of follow-up. Hazard ratio estimates for Mammostrat score

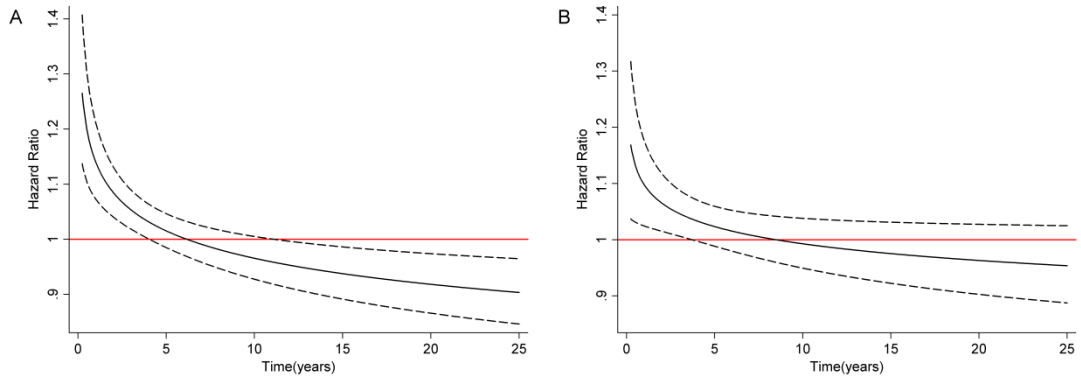



**Figure 8.15** IHC4 Score – Simultaneous Analysis. Time-dependent adjusted hazard ratio (full follow-up) with 95% CIs (dashed lines) for a ten-unit increase in IHC4 score in the Edinburgh BCS cohort for (A) all ER-positive patients and (B) ER-positive patients treated with endocrine therapy only. Adjusted for Mammostrat score, age, grade, nodes positive, treatment and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

after adjustment for time-varying IHC4 and clinical variables were 1.6 (95% CI, 1.1-2.2) and 1.2 (95% CI 0.8-1.9) for medium and high risk compared to low risk for all ER-positive patients.

Neither score was identified as having a significant time-by-covariate interaction by the FPT algorithm for ER-positive patients treated with endocrine therapy-only. Despite the addition of a time-by-covariate interaction for IHC4 not improving the model deviance and therefore not identified in the FPT algorithm, there was still a decrease over time (Figure 8.15B) which was statistically significant as assessed by a Wald test. It demonstrates that IHC4 is prognostic up to 4 years of follow-up only. After adjustment for a time-varying IHC4 score and clinical factors, this gave hazard ratio estimates of 1.62 (95% CI, 1.09-2.42) and 1.30 (95% CI, 0.79-2.13) for medium and high risk compared to low risk Mammostrat score. Not including a time-varying effect for IHC4 results in an adjusted hazard ratio estimate for a ten-unit increase in IHC4 score of 1.02 (95% CI, 0.99-1.06) and very similar Mammostrat score hazard ratio estimates of 1.63 (95% CI, 1.09-2.43) and 1.33 (95% CI, 0.81-2.17) for medium and high risk compared to low risk respectively.

### *8.3.2.3 Exploratory Subgroups*

When IHC4 and Mammostrat were simultaneously entered into the model, IHC4 score was found to have a non-proportional effect in the higher risk node-positive subgroups in the TEAM cohort. Therefore the results are the same as those found when assuming proportional hazards in chapter 7 for ER-positive, node-negative patients treated with endocrine therapy only where both scores remained independent predictors of outcome. In those ER-positive and node-positive, the effect of IHC4 score decreased as follow-up increased (Figure 8.16). IHC4 score was prognostic up to 4 years for those treated with endocrine therapy only and 5 years irrespective of treatment. Mammostrat remained an independent predictor of TTDR in those ER-positive, node-positive patients (irrespective of treatment) in the TEAM cohort with adjusted hazard ratio estimates of 1.35 (95% CI, 1.03-1.78) and 1.47 (95% CI, 1.14-1.91) for medium and high risk compared to low risk Mammostrat score respectively.

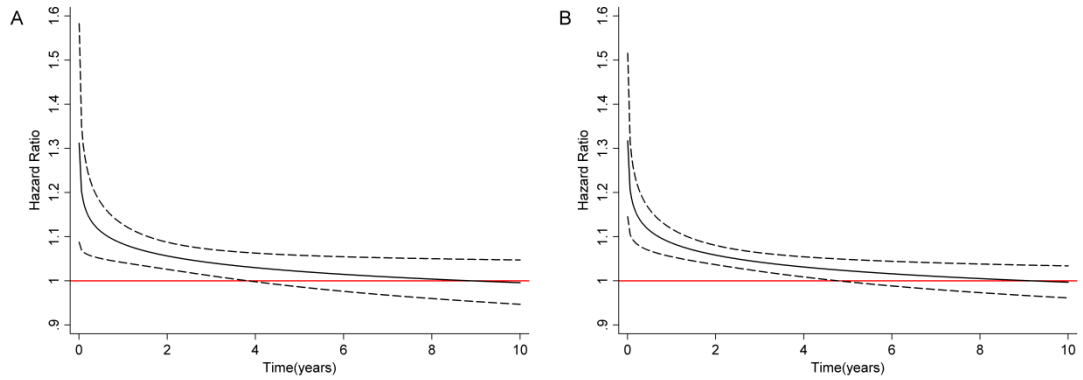



**Figure 8.16** IHC4 Score – Exploratory Subgroups. Time-dependent adjusted hazard ratio with 95% CIs (dashed lines) for a ten-unit increase in IHC4 score for (A) ER-positive, node-positive patients treated with endocrine therapy only and (B) ER-positive, node-positive patients, irrespective of treatment in the TEAM cohort. Adjusted for Mammostrat score, age, grade, nodes positive, treatment and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

In the Edinburgh BCS cohort, for the lower risk patients treated with endocrine therapy only no variables were identified as having a significant time-by-covariate effect, resulting in the same conclusions as from a proportional hazards analysis. IHC4 score was a significant

independent of predictor of outcome for those ER-positive, node-negative and treated with endocrine therapy only with a hazard ratio estimate of 1.06 (95% CI, 1.02-1.11) for a ten-unit increase in IHC4 score. Mammostrat score was not an independent predictor of outcome but those classified as medium risk had an increase hazard of 1.69 (95% CI, 1.02-2.78) compared to those with low risk Mammostrat score. Neither score remained an independent predictor of outcome in those node-positive and treated with endocrine therapy only; however IHC4 score showed a trend towards a protective effect for higher scores with a hazard ratio estimate of 0.95 (95% CI, 0.89-1.01) for a ten-unit increase. Mammostrat score estimates showed a trend toward higher risk associated with higher scores with hazard ratio estimates of 1.51 (95% CI, 0.75-3.03) and 1.74 (95% CI, 0.74-4.08) for medium and high risk compared to low risk Mammostrat score. In the higher risk exploratory subgroup of node-positive patients (irrespective of treatment), IHC4 was identified as having a significant covariate-time interaction (Figure 8.17). The effect of IHC4 decreased over time, higher scores were associated with an increased hazard in the first 3 years of follow-up and a significantly decreased hazard beyond 6.5 years. Mammostrat score was not an independent predictor (after adjustment for time-varying IHC4 score and clinical factors) for those ER-positive, node-positive patients. However there was a trend towards higher risk associated with higher scores with adjusted hazard ratio estimates of 1.55 (95% CI, 0.86-2.79) and 1.41 (95% CI, 0.67-2.98) for medium and high risk compared to low risk Mammostrat score respectively.

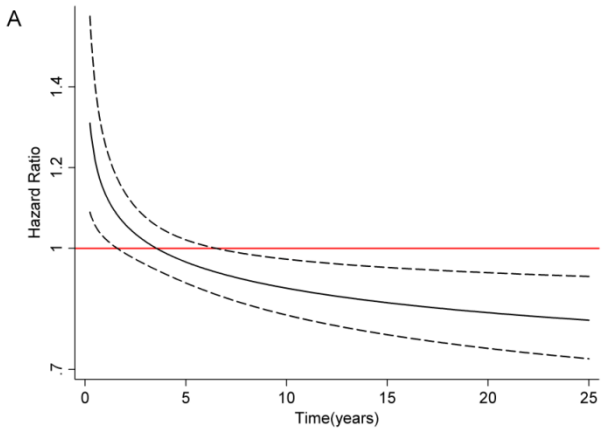



**Figure 8.17** IHC4 Score – Simultaneous Analysis (Exploratory Subgroups). Time-dependent adjusted hazard ratio estimate with 95% CIs (dashed lined) for a ten-unit increase in IHC4 score for ER-positive, node-positive patients in the Edinburgh BCS cohort. Adjusted for Mammostrat score, age, grade, nodes positive and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

#### *8.3.2.4 Combined Score*

There was no evidence to reject the proportional hazards assumption for the combined score, with no significant time-by-covariate interaction identified by the FPT algorithm in the TEAM cohort. The combined score was an independent predictor of outcome up to 10 years follow-up in both patient subgroups with hazard ratio estimates of 1.4 (95% CI, 1.1-1.9) and 2.6 (95% CI, 2.0-3.3) for medium and high risk combined score to low risk combined score for all ER-positive patients. Similarly, hazard ratio estimates were 1.4 (95% CI, 1.0-2.0) and 2.2 (95% CI, 1.6-3.1) for ER-positive patients treated with endocrine therapy only.

Considering full follow-up in the Edinburgh BCS cohort, medium risk compared to low risk combined score was also identified to satisfy the proportional hazards assumption with hazard ratio estimates of 1.4 (95% CI, 1.0-2.1) and 1.4 (95% CI, 0.9-2.3) for all ER-positive patients and ER-positive patients treated with endocrine therapy only, respectively, over the full follow-up. However, high risk compared to low risk was identified to have a significant time-by-covariate interaction in both patient subgroups with the best fitting model to be log-time. The effect was decreasing over time, with the HR estimate crossing the value one at approximately 10 years for all ER-positive patients and 15 years for ER-positive patients

treated with endocrine therapy only (Figure 8.18). In the first year of follow-up we see a large increase in hazard for those classified as high risk compared to low risk, with an estimated adjusted hazard ratio at one year of 5.8 (95% CI, 2.7-12.3) for all ER-positive patients and similarly 5.8 (95% CI, 2.4-13.8) for ER-positive patients treated with endocrine therapy only. However the confidence intervals are very wide. The very large increase in hazard at early time points is due to zero events in the low risk group in the first two years of follow-up compared with 20 in the high risk group.

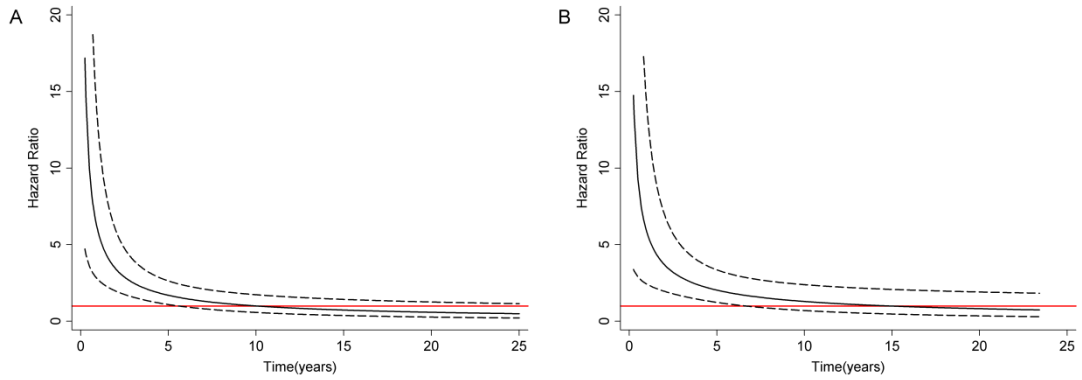



**Figure 8.18** Combined Score. Time-dependent adjusted hazard ratio estimate (full follow-up) with 95% CIs (dashed lines) for high risk compared to low risk combined score in the Edinburgh BCS cohort for (A) all ER-positive patients and (B) ER-positive patients treated with endocrine therapy only. Adjusted for age, grade, nodes positive and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

*Exploratory Subgroups*

No time-varying effects of variables were identified for the exploratory subgroups in the TEAM cohort. Hazard ratio estimates are given in Table 8.14 and the combined score remained an independent predictor of outcome in all exploratory subgroups.

In the Edinburgh BCS cohort, the time-varying effect of high risk combined score remained in the exploratory subgroups (Figure 8.19). The combined score was prognostic of outcome in those ER-positive, node-negative and treated with endocrine therapy only with a hazard ratio estimate of 2.1 (95% CI, 1.1-4.0) for medium compared to low risk combined score and

**Table 8.14** Multivariate Cox regression of the Combined score for the exploratory subgroups in the TEAM cohort

| Main Effect | ER-Positive, Endocrine Only and Node-Negative<br>HR (95% CI) | ER-Positive, Endocrine Only and Node-Positive<br>HR (95% CI) | ER-Positive, Any Treatment and Node-Positive<br>HR (95% CI) |
|---|---|---|---|
| MvL | 2.21 (1.09-4.46) | 1.29 (0.88-1.90) | 1.28 (0.95-1.74) |
| HvL | 3.57 (1.77-7.19) | 1.95 (1.32-2.89) | 2.27 (1.67-3.07) |

NOTE. Multivariate Cox regression analysis for the Combined score in the TEAM cohort. Hazard ratios for risk of TTDR calculated for either high-risk or moderate-risk Combined score compared with low Combined score at baseline. Values represent estimated hazard ratios and 95% CIs for all ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts. Abbreviations: TTDR, time to distant-recurrence; CI, confidence interval; HvL, high risk v low risk Combined score; MvL, moderate risk v low risk Combined score. * Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

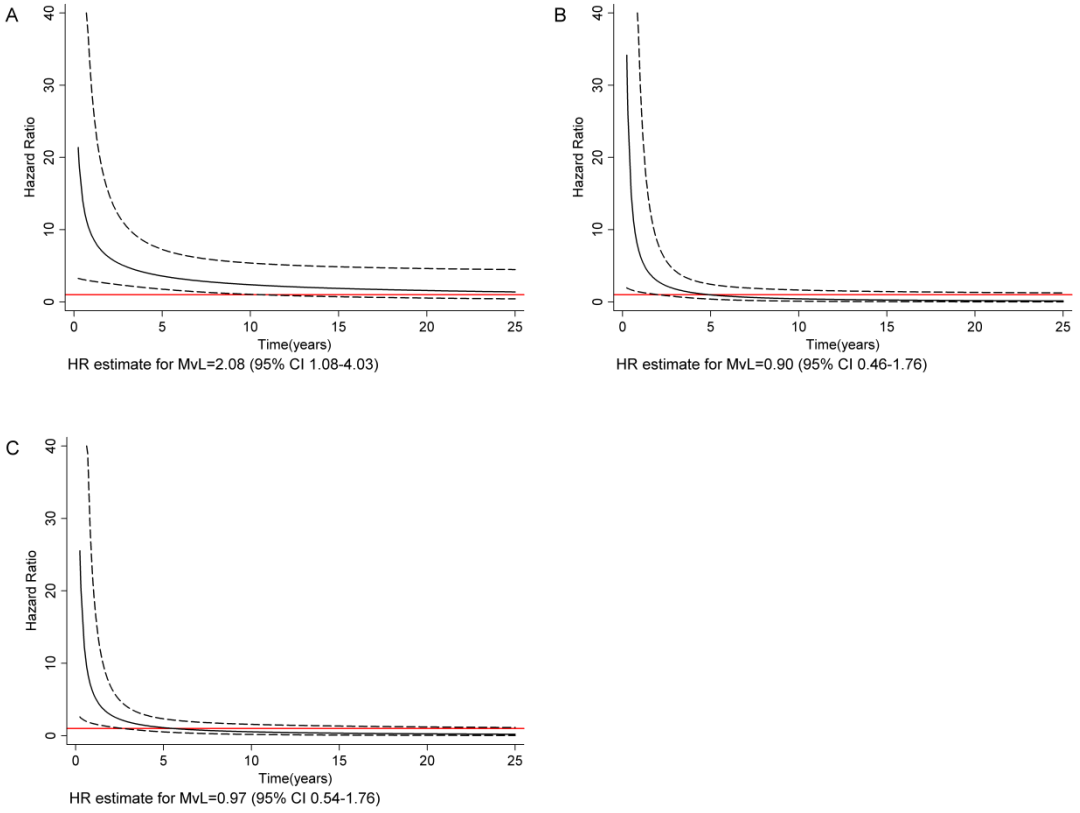



**Figure 8.19** Combined Score (exploratory subgroups). Time-dependent adjusted hazard ratio estimate (full follow-up) with 95% CIs (dashed line) for high risk compared to low risk combined score in the Edinburgh BCS cohort for (A) ER-positive, node-negative and treated with endocrine therapy only, (B) ER-positive, node-positive and treated with endocrine therapy only, and (C) ER-positive, node-positive irrespective of treatment. Adjusted for age, grade, nodes positive and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

those high risk had an increase hazard compared to those classified as low risk up to 10 years. The combined score was not an independent predictor of outcome in those ER-positive, node-positive and treated with endocrine therapy only. In ER-positive, node-positive patients (irrespective of treatment), those high risk had an increased hazard compared to those low risk in the first 3 years of follow-up.

#### *8.3.2.5 Discrimination of Time-Varying Models*

Now we have identified which variables have non-proportional effects and the functional form to model these effects, it is of interest to look at the difference in model discrimination when including time-varying effects and also the comparison between IHC4 and Mammostrat score.

To calculate Royston & Sauerbrei's index of discrimination (D) and $R^2$ including time-varying effects, it is necessary to split the data at each unique event time which allows an interaction between time and the covariate to be fitted.

*10-Year Follow-Up*

The results are given in Table 8.15 for follow-up restricted to 10 years. There is improvement in model discrimination when including a time-varying effect for IHC4 score with increases in D-statistic between 0.08 and 0.32 and increases in $R^2$ between 2.2% and 9.8%. Despite the IHC4 score not being prognostic in ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort ($p=0.07$) there was still improvement in the model discrimination when a time-varying effect was included (increase in D-statistic of 0.11 and $R^2$ of 5.4%).

IHC4 and Mammostrat score were significant independent predictors of TTDR up to 10 years in both cohorts and both patient subgroups ($p<0.05$), except for IHC4 score in ER-positive patients treated with endocrine therapy in the Edinburgh BCS cohort ($p=0.07$). The IHC4 score consistently provided more improvement in model discrimination over clinical

**Table 8.15** Wald tests and performance data for assessing the amount of information added by the time-varying IHC4 score or the Mammostrat score or both to clinical factors in the first 10 years of follow-up.

| Model | Edinburgh BCS | | | TEAM | | |
|---|---|---|---|---|---|---|
| | Wald test (p-value) | % Increase in $R^2$ | Increase in D Statistic | Wald test (p-value) | % Increase in $R^2$ | Increase in D Statistic |
| **All ER-Positive** | | (n=1,103) | | | (n=3,766) | |
| C vs. Null (7 df) | 13.1 (<0.001) | 27.5 | 1.26 | 326.1 (<0.001) | 26.7 | 1.23 |
| C + IHC4 vs. C (1 df) | 4.3 (0.04) | 0.9 | 0.03 | 56.1 (<0.001) | 3.4 | **0.11** |
| C + IHC4 (TV) vs. C + IHC4 (1 df) | 9.2 (0.002) | 9.8 | **0.32** | 9.4 (0.002) | 2.6 | 0.09 |
| C + Mam vs. C (2 df) | 4.1 (0.02) | 1.5 | 0.05 | 32.5 (<0.001) | 2.2 | 0.08 |
| C + IHC4 (TV) + Mam vs. C (4 df) | 5.0 (<0.001) | 11.2 | **0.37** | 86.4 (<0.001) | 7.3 | **0.24** |
| C + IHC4 (TV) + Mam vs. C + IHC4 (TV) (2 df) | 3.5 (0.03) | 0.5 | 0.02 | 20.2 (<0.001) | 1.3 | 0.04 |
| C + IHC4 (TV) + Mam vs. C + Mam (2 df) | 6.0 (0.003) | 9.7 | **0.32** | 52.3 (<0.001) | 5.1 | **0.16** |
| **ER-Positive Endocrine Only** | | (n=831) | | | (n=2,513) | |
| C vs. Null (7 df) | 16.8 (<0.001) | 24.3 | 1.35 | 217.3 (<0.001) | 28.7 | 1.30 |
| C + IHC4 vs. C (1 df) | 3.2 (0.07) | 0.6 | 0.03 | 30.1 (<0.001) | 3.3 | **0.10** |
| C + IHC4 (TV) vs. C + IHC4 (1 df) | 2.2 (0.14) | 5.4 | **0.11** | 5.7 (0.02) | 2.2 | 0.08 |
| C + Mam vs. C (2 df) | 4.0 (0.02) | 1.3 | 0.06 | 9.9 (<0.007) | 1.4 | 0.04 |
| C + IHC4 (TV) + Mam vs. C (4 df) | 3.0 (0.02) | 6.8 | **0.18** | 40.6 (<0.001) | 6.1 | **0.19** |
| C + IHC4 (TV) + Mam vs. C + IHC4 (TV) (2 df) | 3.4 (0.03) | 0.8 | 0.04 | 4.3 (0.12) | 0.6 | 0.01 |
| C + IHC4 (TV) + Mam vs. C + Mam (2 df) | 2.01 (0.13) | 5.5 | **0.12** | 29.8 (<0.001) | 4.7 | **0.15** |

NOTE. Values represent Wald tests (significance level) and the increase in $R^2$ and D statistic for the addition of IHC4 or Mammostrat score or both. IHC4 (TV) refers to including a time-varying effect for IHC4, with an interaction between IHC4 and log (time). Results are given for ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts with follow-up restricted to 10 years. Abbreviations: C, clinical model with age, grade, tumour size, nodal status, treatment and chemotherapy; Null, null model with no covariates; IHC4, IHC4 score; Mam, Mammostrat score; df, degrees of freedom; TV, time-varying.

factors than the Mammostrat score (e.g. increase in D-statistic: 0.20 versus 0.08 and increase in $R^2$: 6.0% versus 2.2%, for all ER-positive patients in the TEAM cohort).

Both scores remained significant independent predictors of TTDR restricted to the first 10 years of follow-up when simultaneously entered into a multivariate Cox regression for all ER-positive patients. Whereas for ER-positive patients treated with endocrine therapy, only Mammostrat score remained an independent predictor of outcome (p=0.03) in the Edinburgh BCS cohort. Oppositely in the TEAM cohort, only IHC4 score remained an independent predictor of outcome where the addition of Mammostrat score did not provide any prognostic information over and above that provided by IHC4 and clinical factors (p=0.1, increase in D-statistic: 0.01, increase in $R^2$: 0.6%). However, after adjustment for IHC4 and clinical factors, those classified as high risk Mammostrat score had an increased hazard of 1.4 (95% CI, 1.0-1.81) compared with low risk Mammostrat score.

*25-Year Follow-Up*

Results are given in Table 8.16 for full-up in the Edinburgh BCS cohort. Again we see a large improvement in model discrimination when a time-varying effect of IHC4 score is included in the model with increases in D-statistic of 0.4 and 0.2 and $R^2$ of 12 and 5% for all ER-positive patients and ER-positive patients treated with endocrine therapy only respectively. The same is not seen however for including a time-varying effect of Mammostrat score, with increases in D-statistic of 0.06 and 0.01 and $R^2$ of 2 and 2.1% for all ER-positive patients and ER-positive patients treated with endocrine therapy only respectively.

Over the full follow-up, both scores remained significant predictors of outcome when simultaneously entered into a multivariate Cox regression for all ER-positive patients only. Only IHC4 score remained an independent predictor of outcome for ER-positive patients treated with endocrine therapy only but this is driven by the high risk Mammostrat score

**Table 8.16** Wald tests and performance data for assessing the amount of information added by the time-varying IHC4 score or the Mammostrat score or both to clinical factors for full-follow-up in the Edinburgh BCS cohort.

| | Edinburgh BCS | | |
|---|---|---|---|
| Model | Wald test (p-value) | % Increase in $R^2$ | Increase in D Statistic |
| **All ER-Positive** | | (n=1,103) | |
| C vs. Null (7 df) | 13.1 (<0.001) | 22.7 | 1.11 |
| C + IHC4 vs. C (1 df) | 2.6 (0.11) | 0.4 | 0.01 |
| C + IHC4 (TV) vs. C + IHC4 (1 df) | 16.7 (<0.001) | 12.0 | **0.38** |
| C + Mam vs. C (2 df) | 3.6 (0.03) | 1.2 | 0.04 |
| C + Mam (TV) vs. C + Mam (1 df) | 4.5 (0.03) | 2.0 | 0.06 |
| C + IHC4 (TV) + Mam vs. C (4 df) | 6.5 (<0.001) | 13.1 | **0.42** |
| C + IHC4 (TV) + Mam vs. C + IHC4 (TV) (2 df) | 3.2 (0.04) | 0.6 | 0.02 |
| C + IHC4 (TV) + Mam vs. C + Mam (2 df) | 9.3 (0.001) | 9.8 | **0.32** |
| **ER-Positive Endocrine Only** | | (n=831) | |
| C vs. Null (7 df) | 16.1 (<0.001) | 24.3 | 1.16 |
| C + IHC4 vs. C (1 df) | 2.5 (0.11) | 0.6 | 0.02 |
| C + IHC4 (TV) vs. C + IHC4 (1 df) | 5.1 (0.02) | 5.4 | **0.17** |
| C + Mam vs. C (2 df) | 3.1 (0.04) | 1.3 | 0.04 |
| C + Mam (TV) vs. C + Mam (1 df) | 3.7 (0.06) | 2.1 | 0.07 |
| C + IHC4 (TV) + Mam vs. C (4 df) | 3.3 (0.01) | 6.9 | **0.22** |
| C + IHC4 (TV) + Mam vs. C + IHC4 (TV) (2 df) | 2.8 (0.06) | 0.8 | 0.03 |
| C + IHC4 (TV) + Mam vs. C + Mam (2 df) | 3.5 (0.03) | 3.5 | **0.11** |

NOTE. Values represent Wald tests (significance level) and the increase in $R^2$ and D statistic for the addition of IHC4 or Mammostrat score or both. (TV) refers to including a time-varying effect for either IHC4 or Mammostrat score. Results are given for ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS cohort with full follow-up. Abbreviations: C, clinical model with age, grade, tumour size, nodal status, treatment and chemotherapy; Null, null model with no covariates; IHC4, IHC4 score; Mam, Mammostrat score; df, degrees of freedom; TV, time-varying.

patients, with patients classified as medium risk having a significantly increased hazard of 1.6 (95% CI, 1.1-2.4) compared with those classified as low risk Mammostrat score.

### 8.3.3 Summary

A summary of time-varying effects identified by the FPT algorithm is displayed in Table 8.17. IHC4 had a strong time-varying effect for all ER-positive patients and ER-positive patients treated with endocrine therapy only in the TEAM cohort. Despite IHC4 being identified as having a time-varying for all ER-positive patients in the Edinburgh BCS cohort,

**Table 8.17** Time-varying effects identified by the FPT algorithm

| | 10-Year Follow-Up | | | | Full Follow-Up | |
|---|---|---|---|---|---|---|
| | Edinburgh BCS | | TEAM | | Edinburgh BCS | |
| | IHC4 | MAM | IHC4 | MAM | IHC4 | MAM |
| **Separate IHC4 & MAM** | | | | | | |
| All ER+ | ✓ | × | ✓ | × | ✓ | ✓ |
| ER+, Endocrine Only | × | × | ✓ | × | ✓ | ✓ |
| **Simultaneous IHC4 & MAM** | | | | | | |
| All ER+ | ✓ | × | ✓ | × | ✓ | × |
| ER+, Endocrine Only | × | × | ✓ | × | × | × |

Note. Summary of whether IHC4 or Mammostrat (MAM) were identified as having a time-varying effect by the FPT algorithm after adjustment for age, grade, tumour size, grade, treatment and chemotherapy. IHC4 & Mammostrat were entered separately or simultaneously into a Cox regression model. Results are given for the main analysis subgroups: all ER-positive (ER+) and all ER-positive treated with endocrine therapy only.

this was not the case for ER-positive patients treated with endocrine therapy only. Figure 8.20 illustrates the results for the different subgroups. Plot A illustrates a time-varying effect of IHC4 compared to a time-constant, with a large over-lap of confidence intervals. Plot B illustrates the comparison of effects for all ER-positive patients and ER-positive patients treated with endocrine therapy only. The effect of a ten-unit increase in IHC4 for the lower risk ER-positive patients treated with endocrine therapy only is not as large at early time-points as the higher risk group of all ER-positive patients. This makes clinical sense with those treated with chemotherapy at an increased risk of early events. However, there is a large overlap in confidence intervals suggesting results are not dissimilar for the two patient subgroups.

A summary of whether IHC4 or Mammostrat were significant predictors of outcome with inclusion of time-varying effects previously identified (Table 8.17) is displayed in Table 8.18. Overall, IHC4 and Mammostrat were independent predictors of outcome when analysed separately or simultaneously after adjustment for clinical factors. Mammostrat was

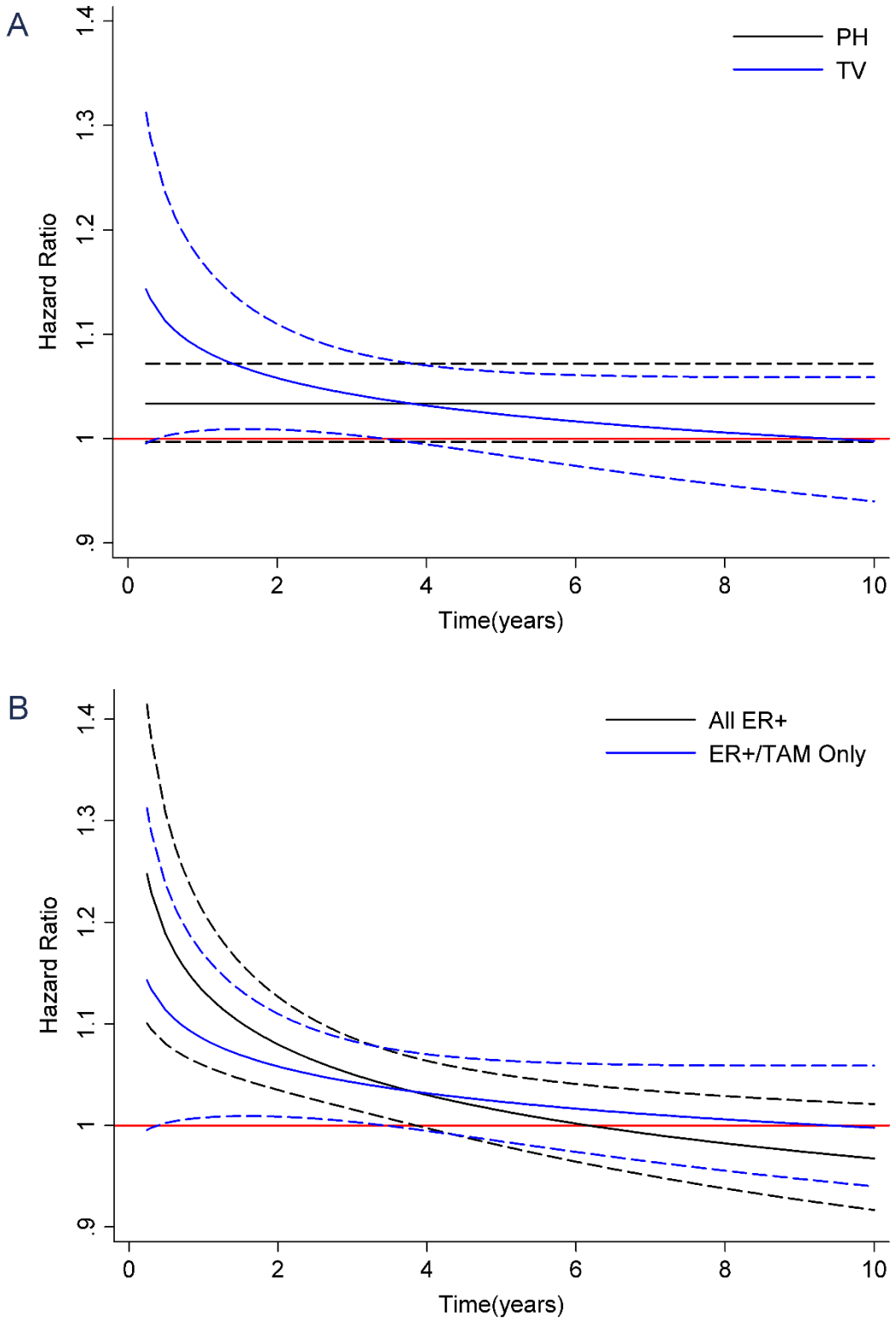



**Figure 8.20** Comparison of adjusted effect and 95% CIs (dashed lines) for a ten-unit increase in IHC4. Plots represent (A) Time-constant (PH) versus time-varying (TV) for ER-positive patients treated with endocrine therapy in the Edinburgh BCS cohort and (B) Time-varying effect for all ER-positive patients versus ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort. Red line at hazard ratio of 1 indicates null effect.

**Table 8.18** Significant predictors of TTDR after identification of time-varying effects

| | 10-Year Follow-Up | | | | Full Follow-Up | |
|---|---|---|---|---|---|---|
| | Edinburgh BCS | | TEAM | | Edinburgh BCS | |
| | IHC4 | MAM | IHC4 | MAM | IHC4 | MAM |
| **Separate IHC4 & MAM** | | | | | | |
| All ER+ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| ER+, Endocrine Only | ✓* | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Simultaneous IHC4 & MAM** | | | | | | |
| All ER+ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| ER+, Endocrine Only | × | ✓ | ✓ | × | ✓ | ✓* |

Note. Summary of whether IHC4 or Mammostrat (MAM) were independent predictors of TTDR after adjustment for age, grade, tumour size, grade, treatment and chemotherapy. IHC4 & Mammostrat were entered separately or simultaneously into a Cox regression model with time-varying effects as previously identified. Results are given for the main analysis subgroups: all ER-positive (ER+) and all ER-positive treated with endocrine therapy only only. * Indicates borderline statistical significance (p-value<=0.07).

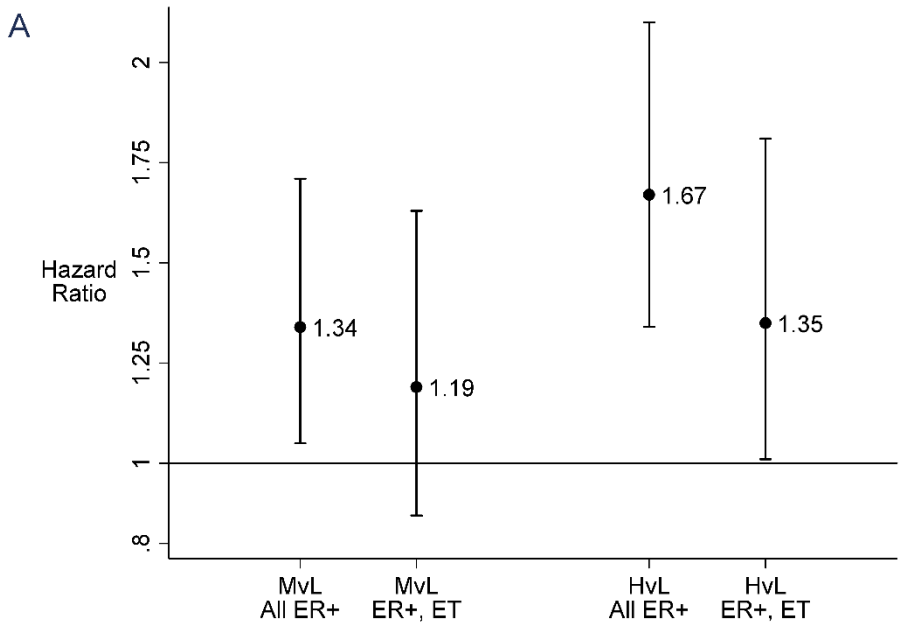



**Figure 8.21** Comparison of estimated effects between subgroups. Hazard ratios and 95% CIs for Mammostrat medium vs. low risk (MvL) and high vs. low risk (HvL) for all ER-positive (ER+) patients and ER-positive patients treated with endocrine therapy only only (ER+, ET) in the TEAM cohort.

overall not an independent predictor of outcome for ER-positive patients treated with endocrine therapy in the TEAM cohort, however those Mammostrat high risk had a significantly increased hazard compared to Mammostrat low risk. The difference in the significance of Mammostrat score between the two subgroups may be due to the reduced number of events in the ER-positive, endocrine therapy only subgroup with overlaps in the estimated confidence intervals (Figure 8.21).

## 8.4 RP Flexible Parametric Models

So far, we have focussed on the Cox proportional hazards model and it's extensions to accommodate non-proportional effects. The literature review identified a parametric approach: flexible parametric models developed by Royston & Parmer (RP flexible parametric models) (see section 4.7).

These models involve modelling on the log-cumulative hazard scale and the baseline hazard is fitted using restricted cubic splines. They are fitted using the *stpm2* command in STATA.

### 8.4.1 Results

#### *8.4.1.1 Baseline hazard*

The first step of fitting an RP model is to determine the degree of flexibility for the baseline hazard function. In both the Edinburgh BCS and TEAM cohorts, the simplest model with 2 degrees of freedom corresponded to the lowest AIC and BIC values (Figure 8.22). The Edinburgh BCS cohort had a consistently lower distant recurrence rate than the TEAM cohort (Figure 8.23) but the trend over time was similar. The recurrence rate peaked at 2 years with approximately 29 distant recurrences per 1000 person-years and a ten-year estimate of 19 per 1000 person-years for the TEAM cohort. In the Edinburgh BCS cohort the recurrence rate peaked at 24 distant recurrences per 1000 person-years at almost 3 years with a ten-year estimate of 12 distant-recurrences per 1000 person-years.

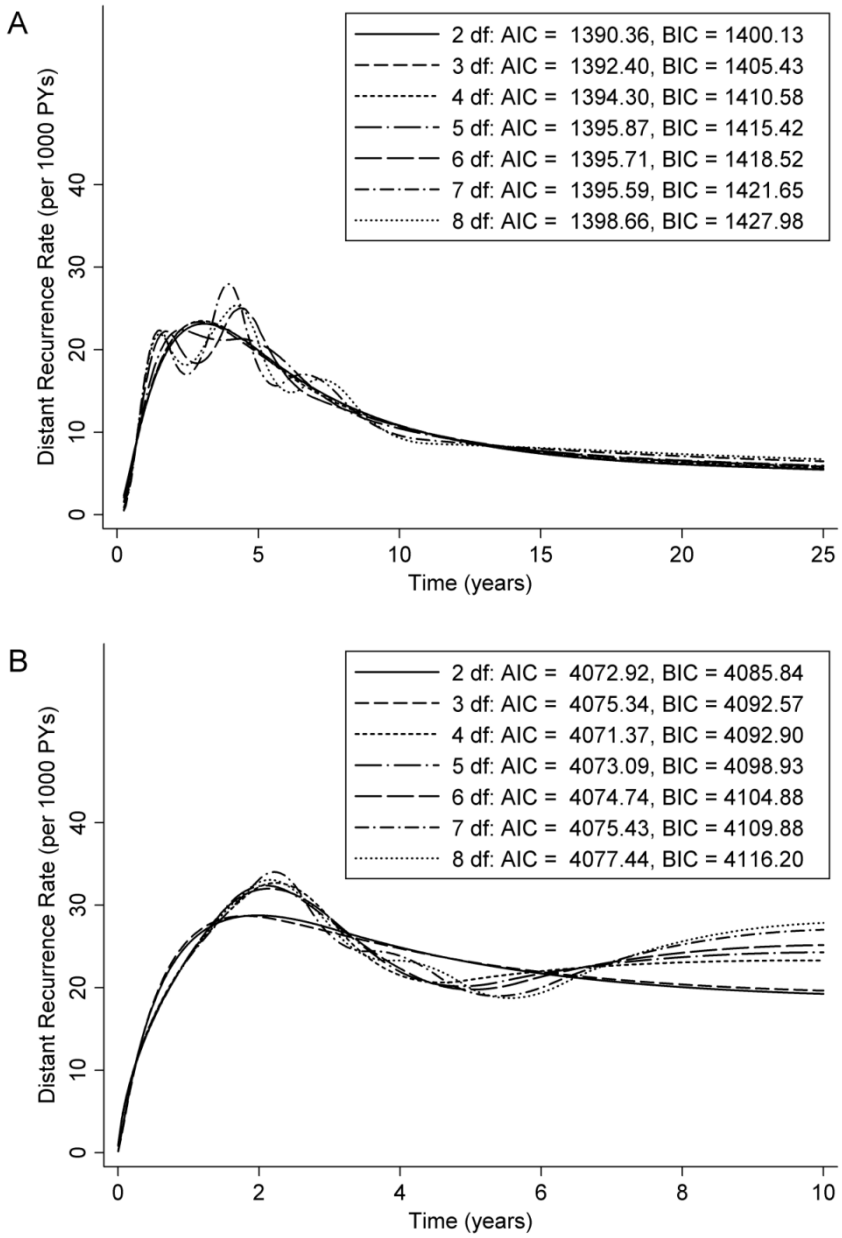



**Figure 8.22** Modelling the baseline hazard (distant recurrence rate per 1000 person-years (PY)) with different degrees of freedom (df) for the (A) Edinburgh BCS and (B) TEAM cohorts.

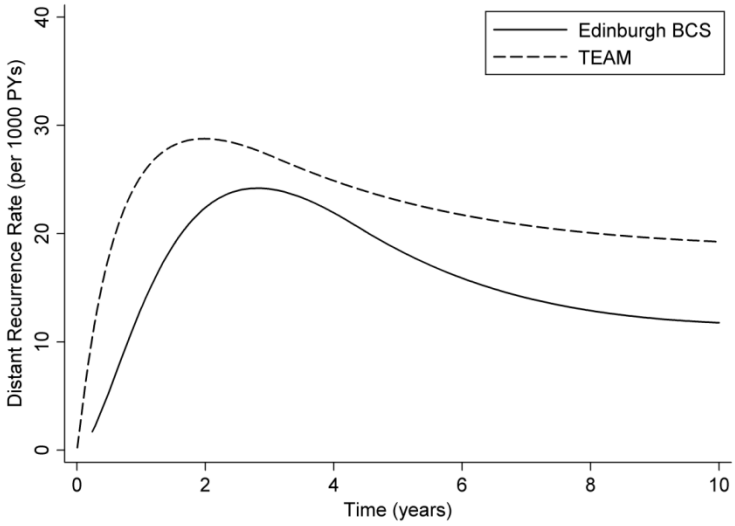



**Figure 8.23** Comparison of baseline hazard rates (per 1000 person-years (PY)) for the Edinburgh BCS and TEAM cohorts.

### *8.4.1.2 Non-proportional effects*

Time-varying effects are included easily by creating an interaction with the covariate and the spline terms. They are chosen using a specified degrees of freedom and a forward selection algorithm. This was performed using the *stpm2t* command in STATA. Once the time-varying effects are identified, the degree of flexibility needed to model the time-varying effect is determined by examining the AIC and BIC values.

*10-Year Follow-up*

The algorithm identified IHC4 to have a time-varying effect for all ER-positive patients in both cohorts with 1 degrees of freedom identified as the best fitting model indicated by the lowest AIC and BIC values (Figure 8.24). The hazard ratio estimate crosses the value 1 at approximately 6.5 years in the Edinburgh BCS cohort (Figure 8.25). The hazard ratio estimate does not cross the value 1 in the TEAM cohort but the lower confidence bound crosses 1 at approximately 9 years. We can see a stronger effect of IHC4 score in the first 2 years, with a two-year hazard ratio estimate of 1.8 (95% CI, 1.3-2.3) and 1.6 (95% CI, 1.4-1.8) in the Edinburgh BCS and TEAM cohorts respectively.

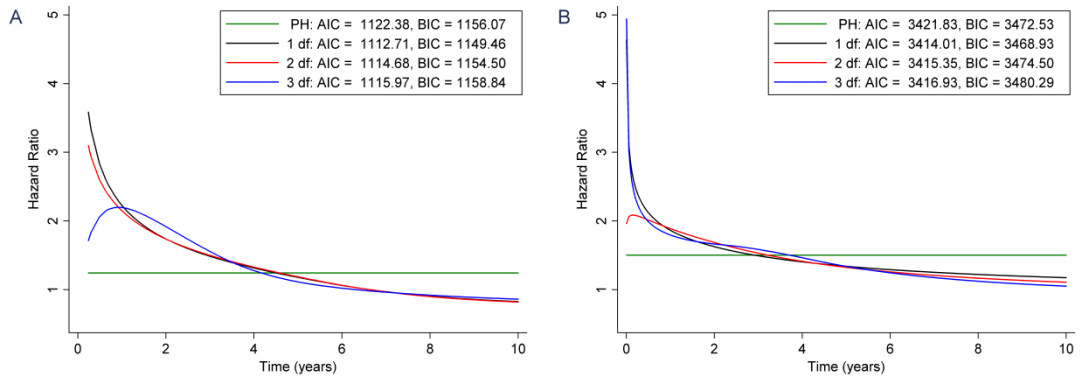



**Figure 8.24** Different degrees of freedom (df) for the time-varying ICH4 score and the proportional hazards estimate (PH). Hazard ratio calculated from between the 75$^{th}$ centile to the 25$^{th}$ centile of IHC4 score. Plots represent all ER-positive patients in the (A) Edinburgh BCS and (B) TEAM cohorts.

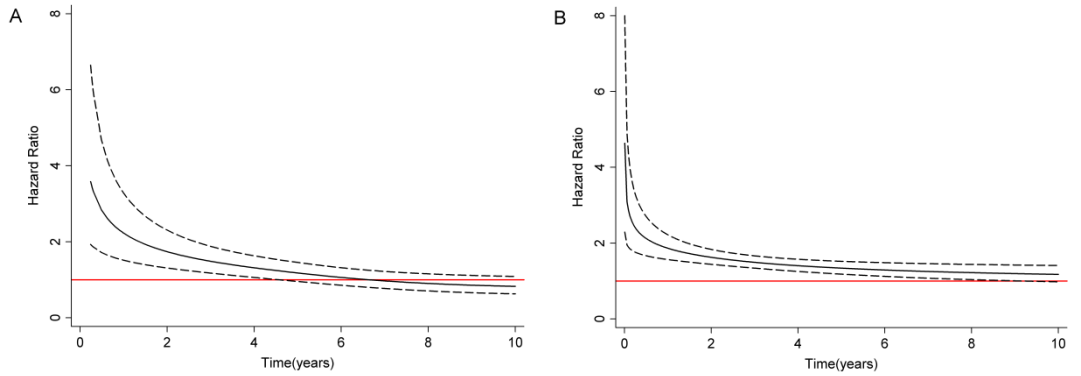



**Figure 8.25** All ER-positive patients. Time-dependent adjusted hazard ratio estimate (up to 10 years) with 95% CIs (dashed lines) for the 75$^{th}$ percentile to the 25$^{th}$ percentile of IHC4 score for all ER-positive patients in the (A) Edinburgh BCS cohort and (B) TEAM cohorts. Adjusted for age, grade, nodes positive, treatment and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

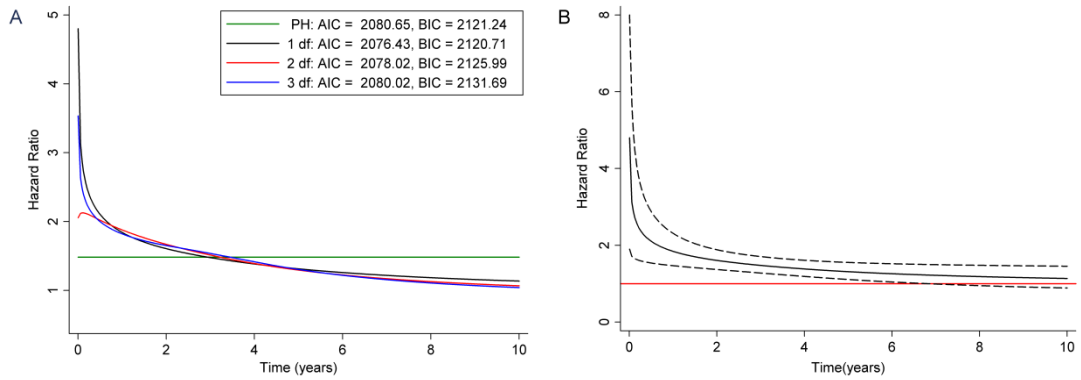



**Figure 8.26** ER-positive patients treated with endocrine therapy. Time-varying effect of IHC4 score for in the TEAM cohort. Plots represent (A) different degrees of freedom (df) for the time-varying ICH4 score and the proportional hazards estimate (PH) and (B) the time-dependent adjusted hazard ratio estimate with 95% CIs (dashed lines) for the 75$^{th}$ percentile to the 25$^{th}$ percentile of IHC4 score. Red line indicating a hazard ratio of 1 corresponding to a null effect.

It is important to note in the extended Cox regression models, the plots of the time-varying effects were for a ten-unit increase in IHC4 score. However due to the parametric nature of the RP models and the time-varying effects modelled by an interaction between the covariate and spline terms, the predictions are based on specific values of IHC4 score. Here we are looking at the 75$^{th}$ compared to the 25$^{th}$ centile of IHC4 score.

Equivalent to the MFPT algorithm, IHC4 score was only identified as having a time-varying effect for ER-positive patients treated with endocrine therapy in the TEAM cohort only (Figure 8.26). This results in a time-constant estimate of 1.04 (95% CI 1.00-1.09) for a ten-unit increase in IHC4 score in the Edinburgh BCS cohort.

Again equivalently to the MFPT approach, no time-varying effects of Mammostrat score was identified in the first 10 years of follow-up for both cohorts. Estimated hazard ratios are given in Table 8.19.

**Table 8.19** Flexible parametric regression of Mammostrat score up to 10-years follow-up

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Edinburgh BCS (n=1,103) | TEAM (n=3,766) | Edinburgh BCS (n=831) | TEAM (n=2,513) |
| Main Effect | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| MvL | 1.59 (1.13-2.24) | 1.42 (1.12-1.81) | 1.79 (1.15-2.82) | 1.28 (0.94-1.75) |
| HvL | 1.31 (0.86-2.00) | 1.90 (1.53-2.37) | 1.79 (1.08-2.97) | 1.58 (1.20-2.09) |

NOTE. Multivariate analysis for Mammostrat score with follow-up censored at 10 years from RP model. Hazard ratios for risk of TTDR calculated for either high-risk or moderate-risk Mammostrat score compared with low Mammostrat scores at baseline. Values represent estimated hazard ratios and 95% CIs for all ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts. Abbreviations: TTDR, time to distant-recurrence; CI, confidence interval; HvL, high risk v low risk Mammostrat score; MvL, moderate risk v low risk Mammostrat score. * Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

The time-varying effect of IHC4 score over the full follow-up in the Edinburgh BCS cohort is shown in Figure 8.27 for both patient subgroups. The time-varying effect is not as prominent in ER-positive patients treated with endocrine therapy and IHC4 is only prognostic of TTDR in the first 4 years of follow-up. For all ER-positive patients, IHC4 score is prognostic for the first 5 years of follow-up, with higher scores corresponding to an increased hazard (4.5 year hazard ratio estimate of 1.2 (95% CI, 1.0-1.5)). After approximately 14.5 years, higher scores correspond to a decreased hazard with a 15 year hazard ratio estimate of 0.8 (95% CI, 0.6-1.0).

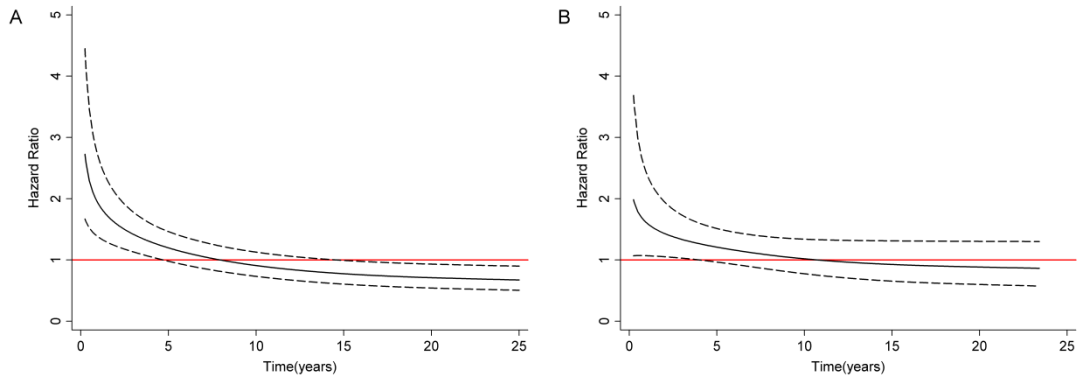



**Figure 8.27** Time-dependent adjusted hazard ratio estimate (full follow-up) with 95% CIs (dashed lines) for 75$^{th}$ centile compared to 25$^{th}$ centile of IHC4 score for (A) all ER-positive patients and (B) ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort. Adjusted for age, grade, nodes positive and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

Mammostrat score was not identified as having a non-proportional effect in either patient subgroup over full follow-up in the Edinburgh BCS cohort as indicated by similar AIC and BIC values between a time-constant and time-varying model (Figure 8.28). Mammostrat score was an independent predictor of outcome for both subgroups with hazard ratio estimates for medium compared to low risk of 1.59 (95% CI, 1.13-2.42) and 1.64 (95% CI, 1.10-2.44) for all ER-positive patients and ER-positive patients treated with endocrine therapy only (Table 8.20).

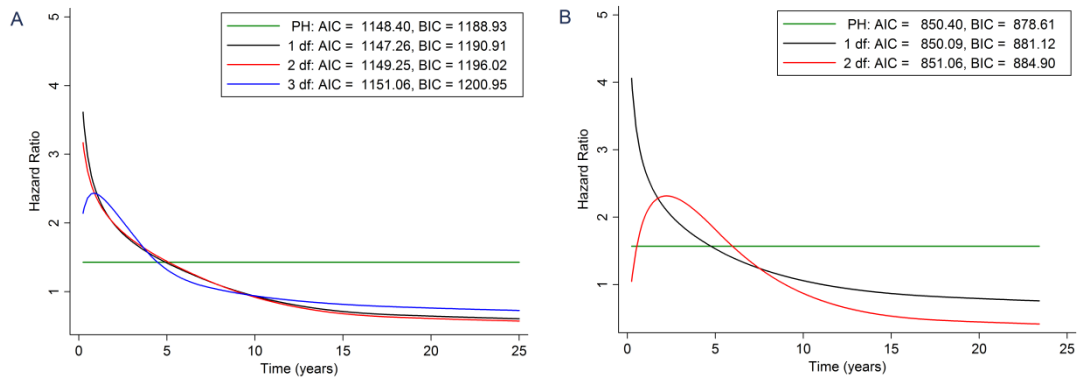



**Figure 8.28** Different degrees of freedom (df) for time-varying Mammostrat high risk and the proportional hazards estimate (PH). Hazard ratio estimate comparing high risk Mammostrat score to low risk Mammostrat score. Plots represent (A) all ER-positive patients and (B) ER-positive patients treated with endocrine therapy in the Edinburgh BCS cohort. Note. Convergence not achieved for 3 degrees of freedom in ER-positive patients treated with endocrine therapy only.

**Table 8.20** Flexible parametric regression of Mammostrat score over full follow-up in the Edinburgh BCS cohort

| | All ER-Positive | ER-Positive Endocrine Only |
|---|---|---|
| Main Effect | HR (95% CI) | HR (95% CI) |
| MvL | 1.59 (1.13-2.42) | 1.64 (1.10-2.44) |
| HvL | 1.31 (0.86-2.00) | 1.44 (0.90-2.31) |

NOTE. Multivariate analysis for Mammostrat score with full follow-up in the Edinburgh BCS cohort from an RP model. Hazard ratios for risk of TTDR calculated for either high-risk or moderate-risk Mammostrat score compared with low Mammostrat scores at baseline. Values represent estimated hazard ratios and 95% CIs for all ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS cohort. Abbreviations: TTDR, time to distant-recurrence; CI, confidence interval; HvL, high risk v low risk Mammostrat score; MvL, moderate risk v low risk Mammostrat score. * Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

*Exploratory Subgroups*

The RP models identified the same time-varying effects as the MFPT algorithm: time-varying effect of IHC4 score in the higher risk node-positive subgroups in the TEAM cohort and a time-varying effect of IHC4 score in the higher risk node-positive (irrespective of treatment) subgroup in the Edinburgh BCS cohort.

### *8.4.1.3 Absolute Predictions of survival*

A key advantage of using a parametric model over the Cox model is it is easy to transform model parameters to make predictions. As well as the predicted hazard ratios over time, which

is a relative measure, you can also look at absolute measures such as survival proportions and hazard rates.

The first column of Figure 8.29 illustrates the difference in predicted hazard rates including a time-constant or a time-varying effect of IHC4. There is little difference for the 75$^{th}$ centile of IHC score but for the 25$^{th}$ centile having a time-constant effect over-estimates the predicted hazard rate at earlier time-points and then later on under-estimates the hazard rate in both cohorts. The second column of Figure 8.29 illustrates the difference in predicted hazard rates between the 75$^{th}$ and 25$^{th}$ centile of IHC4 score, with the difference initially increasing then decreasing over time. In the Edinburgh BCS cohort this difference is only significantly different from zero until approximately 4 years. The difference in predicted survival probability for the 75$^{th}$ to 25$^{th}$ centile of IHC4 score is again more prominent in the TEAM cohort (Figure 8.30), with an estimated reduction in survival of 1.7% (95% CI, 0.7-2.7) at 8 years in the TEAM cohort compared with 0.9% (95% CI, -0.2-1.9) in the Edinburgh BCS cohort.

Absolute predictions for Mammostrat score are shown in Figure 8.31. The issue with low numbers allocated to high risk Mammostrat score in the Edinburgh BCS cohort is highlighted with a larger difference in predicted hazard rates and survival probabilities for medium versus low Mammostrat score compared with high versus low Mammostrat score. In the TEAM cohort, high risk compared to low risk Mammostrat score results in an estimated increase in distant recurrence rate of 4.4 (95% CI, 1.3-7.5) per 1000 person-years and a decrease in survival of 3.0% (95% CI, 1.1-5.0) at 8 years follow-up.

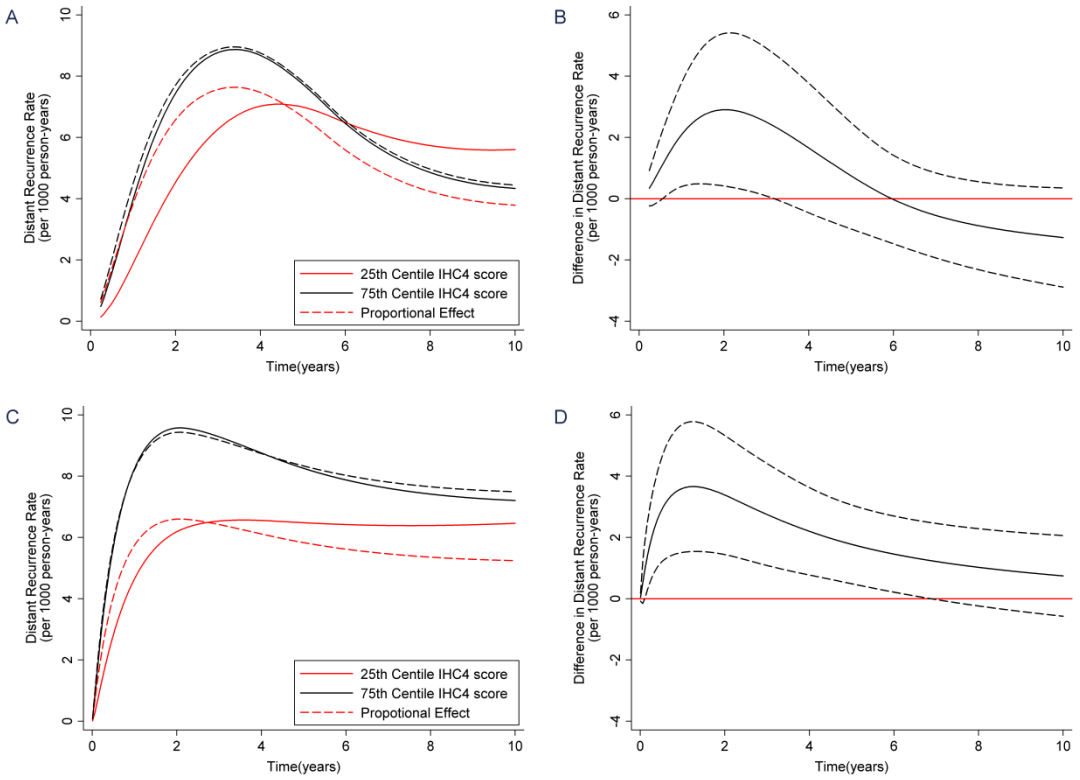



**Figure 8.29** Predicted distant recurrence rates per 1000 person-years for all ER-positive patients in the Edinburgh BCS (A and B) and TEAM (C and D) cohorts. The left column represents the predicted hazard for IHC4 score for the 75$^{th}$ centile and the 25$^{th}$ centile with either a time-constant (proportional) effect (dashed lines) or a time-varying effect (solid lines). The right column represents the difference in predicted distant recurrence rates and 95% confidence interval (dashed lines) between the 75$^{th}$ and 25$^{th}$ centile of IHC4 score. Estimates adjusted for Mammostrat score, age, grade, nodes positive, tumour size, treatment and chemotherapy.

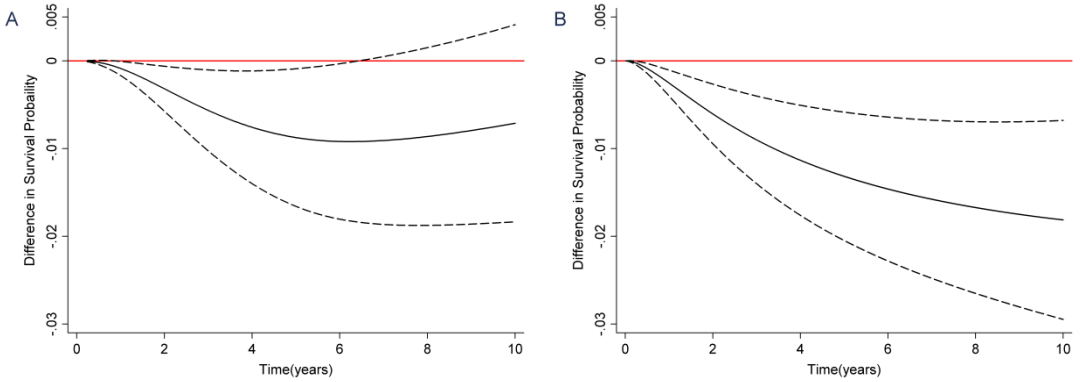



**Figure 8.30** Predicted difference in survival probability and 95% confidence intervals (dashed lines) for the 75$^{th}$ centile compared to 25$^{th}$ centile of IHC4 score for all ER-positive patients in the (A) Edinburgh BCS and (B) TEAM cohorts. Adjusted for Mammostrat score, age, grade, tumour size, nodes positive, treatment and chemotherapy.

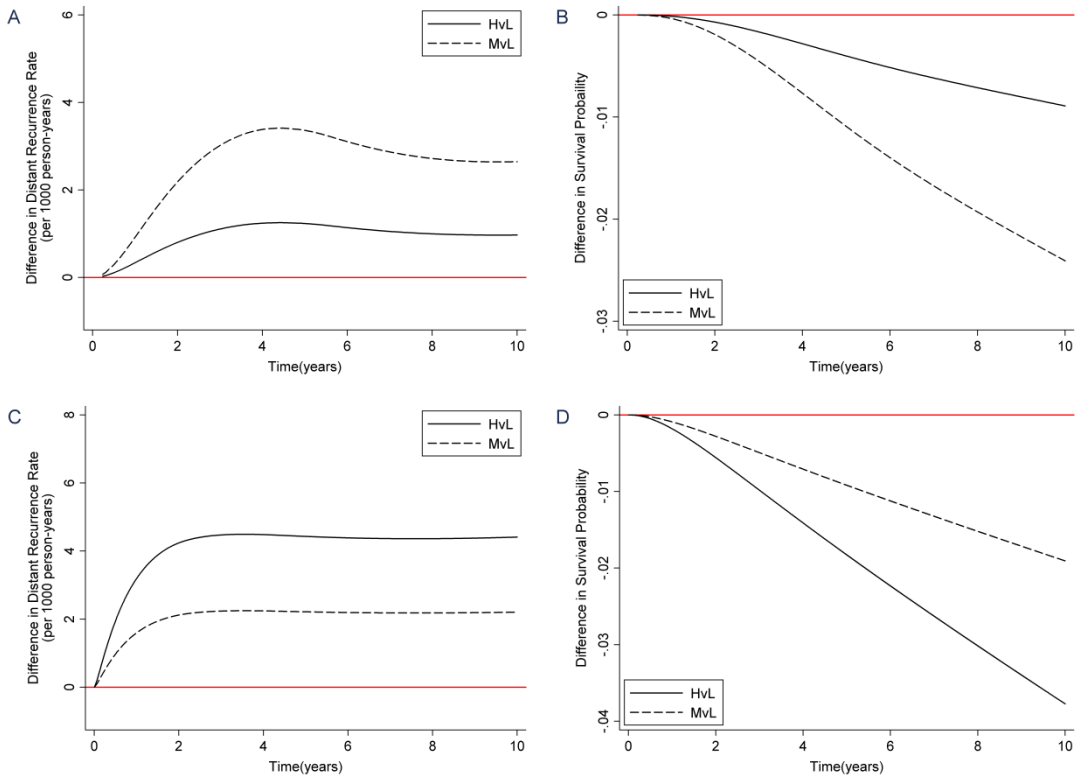



**Figure 8.31** Predicted difference in hazard rates and survival probability for medium and high risk Mammostrat score compared to low risk Mammostrat score for all ER-positive patients in the (A and B) Edinburgh BCS and (C and D) TEAM cohorts. Adjusted for IHC4 score, age, grade, tumour size, nodes positive, treatment and chemotherapy.

#### *8.4.1.4 Restricted mean survival time.*

Including non-proportional effects that are a function of continuous time is advantageous as it prevents categorising of the time-interval. However, this comes with the difficulty of providing an estimate of the effect over time. So far we have plotted the effect over time and read hazard ratio estimates at certain points in time.

When dealing with non-proportional hazards, Royston & Parmer propose to estimate the restricted mean survival time (RMST)(Royston and Parmar, 2011). The restricted mean survival time, $\mu(t^*)$, of a random variable T is the mean of min(T, $t^*$). It may be evaluated as the area under the survival curve S(t) up to $t^*$:

$$\begin{aligned} \mu(t^*) &= E\big(\min(\mathrm{T}, t^*)\big) \\ &= \int_0^{t^*} S(t)dt \ . \end{aligned} \qquad 8.1.$$

When T is time to death, we may think of $\mu(t^*)$ as the '$t^*$-life expectancy'.

Restricted-mean survival time estimates for t* equal to 5 and 10 years are given in Table 8.21. The differences in RMST are more prominent over the 10 years. Being in the 25th centile of IHC4 score increases the time-to distant-recurrence over 10 years by 0.9 (95% CI, 0.3-1.5) years compared to being in the 75th centile of IHC4 score for all ER-positive patients in the Edinburgh BCS cohort. Also, comparing low risk Mammostrat score to medium and high risk increases time-to distant recurrence by 0.9 (95% CI, 0.3-1.5) and 0.6 (95% CI, 0.3-1.5) years respectively. The improvement in time-to distant recurrence holds when IHC4 and Mammostrat are included simultaneously in the model. Larger improvements were seen for all ER-positive patients in the TEAM cohort, with an improvement of 1.0 (95% CI, 0.5-1.5) years for the 25th centile of IHC4 score compared to the 75th centile, and 0.9 (95% CI, 0.3-1.4) years and 1.7 (95% CI, 0.3-1.4) years for low risk Mammostrat score compared to medium and high risk respectively.

**Table 8.21** 5 and 10-year restricted-mean survival estimates for the Edinburgh BCS and TEAM cohorts

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Edinburgh BCS (n=1,103) | TEAM (n=3,766) | Edinburgh BCS (n=831) | TEAM (n=2,513) |
| | Diff in RMST (95% CI) | Diff in RMST (95% CI) | Diff in RMST (95% CI) | Diff in RMST (95% CI) |
| **t* = 5 Years** | | | | |
| Separate* | | | | |
| IHC4 (TV) | 0.64 (0.17-1.11) | 0.38 (0.03-0.73) | 0.51 (0.02-1.00) | 0.53 (0.12-0.94) |
| LvM | 0.57 (0.10-1.05) | 0.27 (-0.09-0.62) | 0.59 (0.11-1.08) | 0.26 (-0.15-0.68) |
| LvH | 0.36 (0.09-1.05) | 0.54 (-0.09-0.63) | 0.53 (0.11-1.08) | 0.51 (-0.16-0.68) |
| Combined* | | | | |
| IHC4 (TV) | 0.60 (0.12-1.08) | 0.30 (-0.04-0.64) | 0.48 (-0.03-0.99) | 0.47 (0.10-0.83) |
| LvM | 0.50 (0.02-0.97) | 0.17 (-0.16-0.51) | 0.60 (0.10-1.11) | 0.15 (-0.21-0.51) |
| LvH | 0.20 (0.02-0.97) | 0.34 (-0.17-0.51) | 0.43 (0.10-1.11) | 0.28 (-0.21-0.51) |
| **t* = 10 Years** | | | | |
| Separate* | | | | |
| IHC4 (TV) | 0.90 (0.31-1.49) | 1.01 (0.47-1.56) | 0.90 (0.27-1.54) | 1.19 (0.59-1.79) |
| LvM | 0.90 (0.32-1.47) | 0.85 (0.30-1.40) | 0.90 (0.28-1.51) | 0.71 (0.10-1.32) |
| LvH | 0.60 (0.31-1.49) | 1.65 (0.30-1.40) | 0.60 (0.27-1.53) | 1.32 (0.11-1.31) |
| Combined* | | | | |
| IHC4 (TV) | 0.91 (0.28-1.53) | 0.82 (0.27-1.36) | 0.91 (0.24-1.58) | 1.06 (0.45-1.66) |
| LvM | 0.95 (0.34-1.55) | 0.62 (0.09-1.16) | 0.95 (0.30-1.59) | 0.46 (-0.14-1.06) |
| LvH | 0.43 (0.32-1.57) | 1.17 (0.09-1.16) | 0.43 (0.28-1.62) | 0.85 (-0.14-1.06) |

NOTE. Difference in RMST for t* equal to 5 and 10 years for Mammostrat score and IHC4 entered separately or simultaneously (combined analysis) into a Royston-Parmer flexible paramteric model. Difference in RMST calculated as the difference in the IQR of the continuous IHC4 score and either high-risk or moderate-risk Mammostrat score compared with low Mammostrat scores at baseline. Abbreviations: RMST, restricted mean survival time; CI, confidence interval; IHC4 (TV), time-varying IHC4 score; LvH, low risk v high risk Mammostrat score; LvM, low risk v moderate risk Mammostrat score. * Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

With longer follow-up available in the Edinburgh BCS cohort, 15 and 20-year RMST differences were available (Table 8.22). The largest improvement in time-to distant recurrence was seen in low risk Mammostrat score compared to high risk over 20 years, with a difference in RMST estimate of 1.9 (95% CI, 0.9-2.8) years.

**Table 8.22** 15 and 20-year restricted-mean survival estimates for the Edinburgh BCS cohort

| | t* = 15 Years | | t* = 20 Years | |
|---|---|---|---|---|
| | All ER-Positive (n=1,103) | ER-Positive Endocrine Only (n=831) | All ER-Positive (n=1,103) | ER-Positive Endocrine Only (n=831) |
| | Diff in RMST (95% CI) | Diff in RMST (95% CI) | Diff in RMST (95% CI) | Diff in RMST (95% CI) |
| Separate* | | | | |
| IHC4 (TV) | 1.14 (0.27-2.01) | 0.97 (0.14-1.81) | 1.23 (0.26-2.21) | 1.05 (0.15-1.95) |
| LvM | 1.48 (0.63-2.32) | 1.26 (0.48-2.04) | 1.85 (0.91-2.79) | 1.42 (0.58-2.25) |
| LvH | 0.96 (0.62-2.33) | 1.14 (0.47-2.05) | 1.22 (0.89-2.81) | 1.29 (0.58-2.26) |
| Combined* | | | | |
| IHC4 (TV) | 1.07 (0.19-1.96) | 0.95 (0.08-1.83) | 1.15 (0.16-2.15) | 1.04 (0.09-1.99) |
| LvM | 1.42 (0.56-2.23) | 1.41 (0.57-2.25) | 1.82 (0.87-2.77) | 1.65 (0.75-2.56) |
| LvH | 0.66 (0.55-2.29) | 0.99 (0.56-2.27) | 0.86 (0.84-2.80) | 1.17 (0.73-2.58) |

NOTE. Difference in RMST for t* equal to 15 and 20 years for Mammostrat score and IHC4 entered separately or simultaneously (combined analysis) into a Royston-Parmer flexible paramteric model. Difference in RMST calculated as the difference in the IQR of the continuous IHC4 score and either high-risk or moderate-risk Mammostrat score compared with low Mammostrat scores at baseline. Abbreviations: RMST, restricted mean survival time; CI, confidence interval; IHC4 (TV), time-varying IHC4 score; LvH, low risk v high risk Mammostrat score; LvM, low risk v moderate risk Mammostrat score. * Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

### *8.4.1.5 Combined Score*

No time-varying effects were identified for the combined score in the TEAM cohort with hazard ratio estimates given in Table 8.23. The combined score was an independent predictor of outcome for both patient subgroups.

**Table 8.23** Flexible parametric regression of Combined score up to 10-years follow-up

| | All ER-Positive | ER-Positive Endocrine Only |
|---|---|---|
| Main Effect | HR (95% CI) | HR (95% CI) |
| MvL | 1.45 (1.10-1.90) | 1.45 (1.04-2.03) |
| HvL | 2.55 (1.95-3.34) | 2.23 (1.59-3.12) |

NOTE. Multivariate analysis for the Combined score with follow-up censored at 10 years from RP model. Hazard ratios for risk of TTDR calculated for either high-risk or moderate-risk Combined score compared with low Combined score at baseline. Values represent estimated hazard ratios and 95% CIs for all ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh BCS and TEAM cohorts. Abbreviations: TTDR, time to distant-recurrence; CI, confidence interval; HvL, high risk v low risk Mammostrat score; MvL, moderate risk v low risk Mammostrat score. * Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

The forward selection algorithm identified the combined score to have a time-varying effect in the Edinburgh BCS cohort. The hazard ratio estimate crosses the value 1 at approximately 9 years for all ER-positive patients and 19 years for ER-positive patients treated with endocrine therapy only (Figure 8.32).

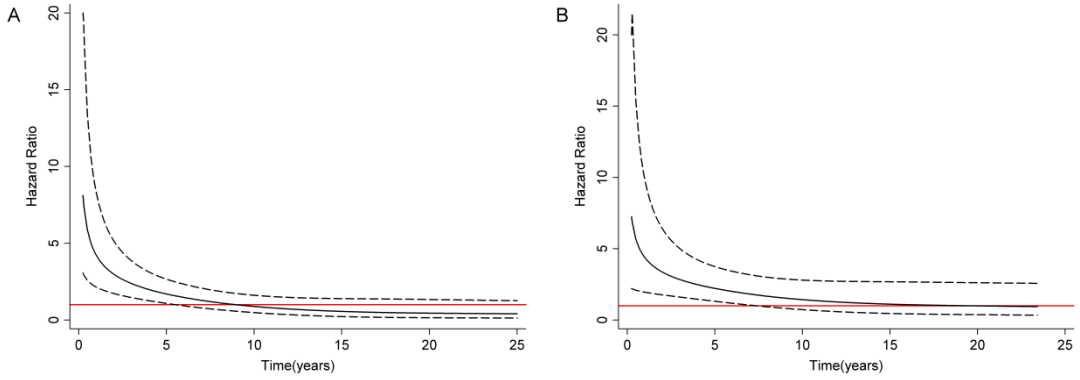



**Figure 8.32** Combined Score. Time-dependent adjusted hazard ratio estimate (full follow-up) in the Edinburgh BCS cohort with 95% CIs (dashed lines) for high risk compared to low risk combined score for (A) all ER-positive patients and (B) ER-positive patients treated with endocrine therapy only. Adjusted for age, grade, nodes positive, treatment and chemotherapy. Red line at hazard ratio of 1 corresponds to a null effect.

Again the same time-varying effects were identified as with the MFPT algorithm for the exploratory subgroups. There was no evidence to reject the PH assumption for the combined score in all exploratory subgroups in the TEAM cohort and was an independent predictor of outcome. The time-varying effect of high risk combined score remained in the Edinburgh BCS cohort. The combined score was prognostic of outcome for ER-positive, node-negative patients treated with endocrine therapy only and those ER-positive, node-positive (irrespective of treatment).

### 8.4.2 Summary

Royston-Parmer flexible parametric models were identified in the literature review as an alternative to Cox regression models. They explicitly model the baseline hazard with the use of restricted cubic splines. We illustrated modelling the baseline hazard using different degrees of freedom. The pattern over time that was observed, with a peak in distant recurrences at approximately 2-3 years is what has been previously observed in breast cancer. The Edinburgh BCS cohort was again highlighted as a lower risk population than the TEAM cohort, with a lower distant recurrence rate across all time points.

RP models allow the inclusion of time-varying effects with an interaction between the covariate and the spline terms. Using a forward selection approach, the RP models identified the same time-varying effects as was identified with the FPT algorithm in the Cox models. One exception was in the Edinburgh BCS cohort considering full follow-up, where the FPT algorithm identified a time-varying effect for Mammostrat high risk whereas a time-constant effect was identified with the RP models. However, this results are similar with a large overlap in confidence intervals (Figure 8.33).

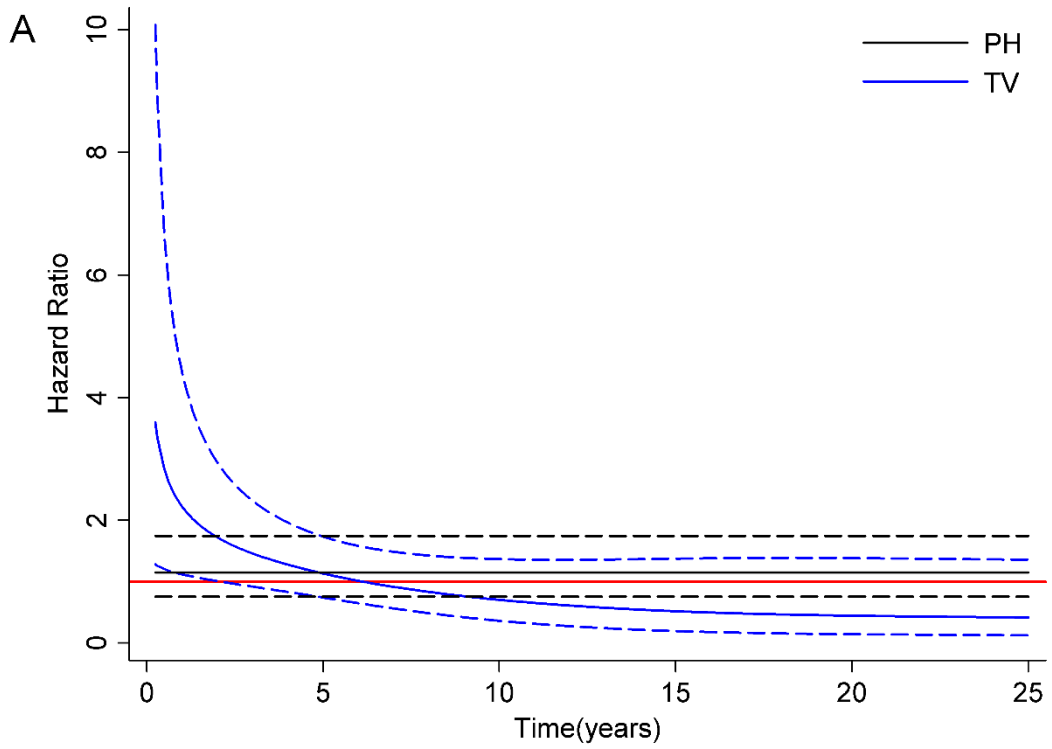



**Figure 8.33** Comparison of time-constant (PH) and time-varying (TV) effect and 95% CIs (dashed lines) for Mammostrat high risk versus low risk for all ER-positive patients in the Edinburgh BCS cohort. Red line at hazard ratio of 1 indicates null effect.

The parametric nature of the RP models allow easy transformation of model parameters to make predictions in particular absolute risk measures rather than relative risk measures (e.g. the hazard ratio). This advantage was illustrated producing plots of differences in hazard rates and survival proportions for IHC4 and Mammostrat score.

A novel measure, the restricted mean survival time, was illustrated as an alternative to the hazard ratio which has the advantage of a more intuitive interpretation and can also be estimated when covariates have non-proportional effects.

## 8.5 Comparison of Approaches

Focussing on the combined model including both scores as well as clinical factors, the comparison of the estimated effects from assuming proportional hazards and including time-varying effects are explored.

### 8.5.1 Extended Cox Model versus Proportional Hazards

IHC4 was identified as having a time-varying effect in both the Edinburgh BCS and TEAM cohorts for all ER-positive patients. When assuming proportional hazards, this results in an initial under-estimation at earlier time-points then an over-estimation of the effect at later time-points (Figure 8.34). You would conclude IHC4 and Mammostrat score are independent predictors of TTDR for all ER-positive patients over the 10 year follow-up in the TEAM cohort. However, when you allow for non-proportional hazards, Mammostrat score remains prognostic of outcome over the 10 years but IHC4 is only prognostic up to approximately 7 years when the lower confidence bound crosses 1. In the Edinburgh BCS cohort, assuming proportional hazards you would conclude only Mammostrat score remains an independent predictor of outcome. Whereas allowing for non-proportional hazards IHC4 score is prognostic of outcome in the first 4 years of follow-up and beyond 11 years of follow-up where increasing IHC4 score has a protective effect.

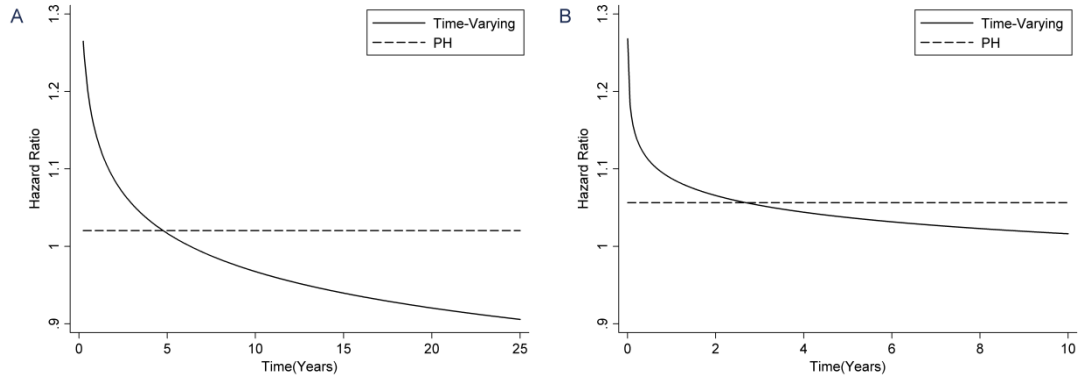



**Figure 8.34** Comparison of extended Cox versus Cox PH model. Plots represent adjusted hazard ratio estimate from an extended Cox model (solid line) and a Cox model assuming proportional hazards (dashed line) for a ten unit increase in IHC4 score in the (A) Edinburgh BCS and (B) TEAM cohorts.

Categorising time results in unrealistic jumps in hazard ratio estimates at cut-points. The five-yearly intervals give some indication of the pattern over time in the larger TEAM cohort but unreliable estimates were observed in the Edinburgh BCS cohort (Figure 8.35). Similarly with two-yearly intervals, estimates are unreliable especially towards the end of follow-up where numbers of events are small (Figure 8.36).

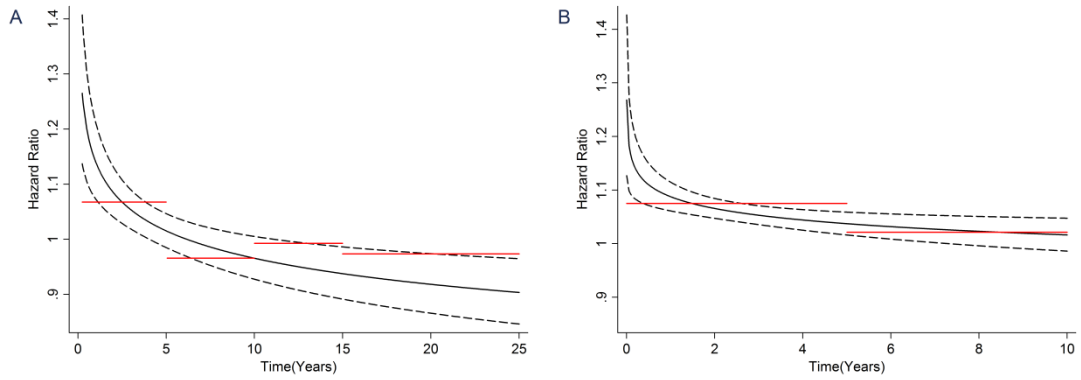



**Figure 8.35** Comparison of extended Cox model and piece-wise constant effects. Adjusted hazard ratio estimate from an extended cox model with 95% confidence intervals (dashed lines) and piece-wise constant effects with 5 year time intervals (red horizontal lines).Plots represent (A) Edinburgh BCS and (B) TEAM cohorts. Note. A final interval of 10 years was used for the Edinburgh BCS cohort due to the small number of events towards the end of follow-up.

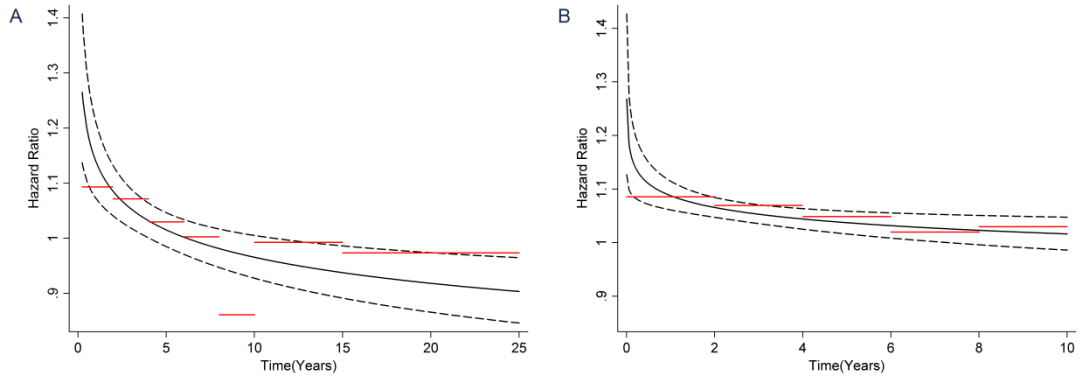



**Figure 8.36** Comparison of extended Cox model and two-yearly intervals.Adjusted hazard ratio estimate from an extended cox model with 95% confidence intervals (dashed lines) and piece-wise constant effects with 2 year time intervals (red horizontal lines). Plots represent (A) Edinburgh BCS and (B) TEAM cohorts. Note. Final two intervals of 5 and 10-years in the Edinburgh BCS cohort due to small number of events towards end of follow-up.

### 8.5.2 RP Models versus Extended Cox Models

In both the flexible parametric RP models and the extended Cox model, the forward selection algorithm for the RP models and the FPT algorithm identified the same time-varying effects for IHC4 score. Both showed a decreasing effect over time for IHC4 and in both approaches the simplest model was chosen, log (time) for the extended Cox model and 1 degrees of freedom for the spline term in the RP model. This resulted in almost identical estimates in the TEAM cohorts and only slight differences in the Edinburgh BCS cohort at early time-points (Figure 8.37).

Over the 10-year follow-up, no time-varying effects were identified for Mammostrat score. The time-constant hazard ratios were almost identical from both models (Table 8.24).

However, differences in the selection of time-varying effects for Mammostrat score was observed in the Edinburgh BCS cohort considering the full follow-up. The FPT algorithm identified a time-varying effect for high risk Mammostrat score whereas in the RP models it was identified as time-constant. The small number of patients allocated to high risk Mammostrat score and after adjustment for IHC4 score the time-varying effect did not remain, suggesting some over-fitting of the model using the FPT algorithm.

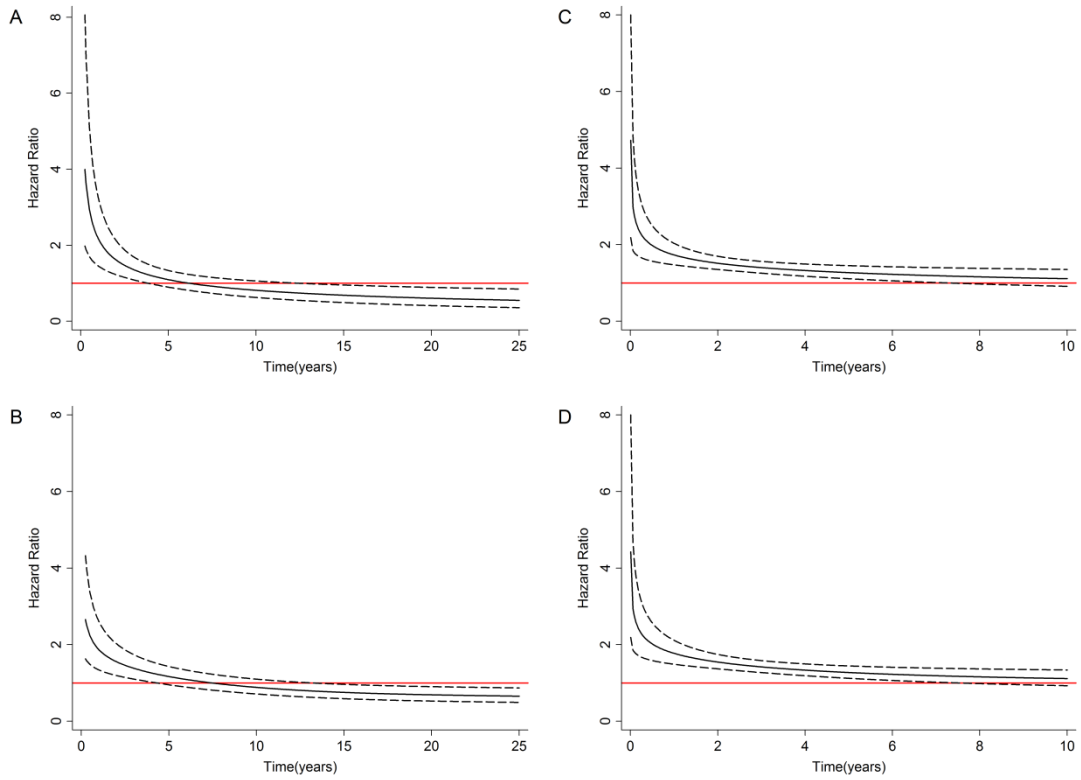



**Figure 8.37** Comparison of extended Cox and RP models. Adjusted hazard ratio estimate with 95% CIs (dashed lines) for the 75$^{th}$ percentile to the 25$^{th}$ percentile of IHC4 score for all ER-positive patients from an extended Cox model (A and C) and a RP model (B and D) in the Edinburgh BCS (left-column) and TEAM (right-column) cohorts.

**Table 8.24** Comparison of hazard ratio estimates for Mammostrat score from an extended Cox or RP model

| | Edinburgh BCS | | TEAM | |
|---|---|---|---|---|
| | Extended Cox | RP | Extended Cox | RP |
| Main Effect | HR (95% CI) | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| MvL | 1.67 (1.14-2.45) | 1.67 (1.14-2.45) | 1.34 (1.05-1.71) | 1.34 (1.05-1.71) |
| HvL | 1.37 (0.85-2.21) | 1.39 (0.86-2.23) | 1.67 (1.34-2.09) | 1.68 (1.34-2.11) |

NOTE. Multivariate analysis for Mammostrat score with follow-up censored at 10 years from extended Cox or RP model. Hazard ratios for risk of TTDR calculated for either high-risk or moderate-risk Mammostrat score compared with low Mammostrat scores at baseline. Values represent estimated hazard ratios and 95% CIs for all ER-positive patients in the Edinburgh BCS and TEAM cohorts. Abbreviations: TTDR, time to distant-recurrence; RP, Roytson-Parmer model; CI, confidence interval; HvL, high risk v low risk Mammostrat score; MvL, moderate risk v low risk Mammostrat score. * Adjusted for time-varying IHC4 score, age, grade, tumour size, nodes positive, treatment and chemotherapy.

An advantage for the use of RP models over the MFPT approach in extended Cox models is avoiding categorising of the time interval to determine the time-varying effects. This avoids the loss of power associated with categorisation and the subjective choice of time-intervals. The time taken to run the FPT algorithm in STATA is also time-consuming compared to the forward selection algorithm in RP models. Running the *stmfpt* command with time categorised into 2 monthly periods, resulting in 134,000 observations in the TEAM cohort, takes approximately 36 times more computer time than the *stpm2t* command for the RP models.

Another advantage for RP models is the easy calculation of absolute risk estimates, such as the differences in predicted hazard rates and survival probabilities.

The main advantage for the MFPT approach is that it resembles a closed test algorithm: the algorithm runs a sequence of tests maintaining the overall type I error rate at a pre-specified nominal level, such as 5%. So it simultaneously determines whether time-varying effects should be added into the model and the best fitting FP to model the time-varying effect. Whereas in the RP models, you need to pre-specify the degrees of freedom for identifying time-varying effects and then simplify the functional form of any significant time-varying effects.

A disadvantage of the RP models is due to the model being parametric; difficulties with convergence can arise in smaller datasets when more complex models are being fitted, such as more degrees of freedom for the baseline hazard or time-varying effects and when more than one time-varying effect is selected.

## 8.6 Conclusions from various subgroup analyses

Differences were observed on the prognostic ability of IHC4 and Mammostrat score depending on the patient subgroup and cohort (Table 8.25). Overall, IHC4 score had a strong time-varying effect in the higher risk patient subgroups. Higher scores were associated with

**Table 8.25** Conclusions on whether IHC4 and/or Mammostrat were prognostic of TTDR in various subgroups.

| Subgroup | **Edinburgh BCS** | **TEAM** |
|---|---|---|
| All ER-Positive | - IHC4 was identified as having a significant time-by-covariate interaction, and the effect of IHC4 was decreasing over time. IHC4 score was prognostic up to 4 years and then beyond 11 years where higher scores were associated with a protective effect.<br>- Mammostrat score satisfied the PH assumption (after adjustment for time-varying IHC4 score) and was an independent predictor of outcome. | - IHC4 was identified as having a significant time-by-covariate interaction, and the effect of IHC4 was decreasing over time. IHC4 score was prognostic up to 7 years follow-up.<br>- Mammostrat score satisfied the PH assumption and was an independent predictor of outcome. |
| ER-Positive Endocrine Therapy Only | - IHC4 was identified as having a significant time-by-covariate interaction, and the effect of IHC4 was decreasing over time. IHC4 was prognostic up to 4 years follow-up.<br>- Mammostrat score satisfied the PH assumption (after adjustment for time-varying IHC4 score) and was an independent predictor of outcome (borderline significance). | - IHC4 was identified as having a significant time-by-covariate interaction, and the effect of IHC4 was decreasing over time. IHC4 score was prognostic up to 6 years follow-up.<br>- Mammostrat score satisfied the PH assumption but was not an independent predictor of outcome (after adjustment for time-varying IHC4 score). However, those classified as Mammostrat high risk had a significant increased hazard compared to those with low risk Mammostrat score. |

| Subgroup | **Edinburgh BCS** | **TEAM** |
|---|---|---|
| ER-Positive, Node-Negative, Endocrine Therapy Only | - No time-varying effects were identified.<br>- IHC4 score was an independent predictor of TTDR.<br>- Mammostrat score did not remain an independent predictor of outcome but there was some evidence of an effect for Mammostrat medium risk compared to low risk. | - No time-varying effects were identified.<br>- IHC4 and Mammostrat score were independent predictors of TTDR. |
| ER-Positive, Node-Positive, Endocrine Therapy Only | - No time-varying effects were identified.<br>- Neither score was a significant independent predictor of outcome.<br>- There was a trend towards higher IHC4 scores associated with a protective effect and higher Mammostrat scores associated with an increased hazard. | - IHC4 was identified as having a significant time-by-covariate interaction, and the effect of IHC4 was decreasing over time. IHC4 was prognostic up to 4 years follow-up.<br>- Mammostrat score did not remain an independent predictor of TTDR. |
| ER-Positive, Node-Positive, Irrespective of Treatment | - IHC4 was identified as having a significant time-by-covariate interaction, and the effect of IHC4 was decreasing over time. Higher scores were associated with an increased hazard in the first 3 years of follow-up and a significantly decreased hazard beyond 6.5 years.<br>- Mammostrat score was not an independent predictor of TTDR. There was a trend towards higher risk associated with higher scores. | - IHC4 was identified as having a significant time-by-covariate interaction, and the effect of IHC4 was decreasing over time. IHC4 was prognostic up to 5 years follow-up.<br>- Mammostrat score remained an independent predictor of TTDR. |

an increased risk of early recurrence but were associated with a significantly reduced risk of recurrence after 11 years. Mammostrat score satisfied the PH assumption and provided prognostic information over the full 10 year follow-up in the TEAM cohort and 25 years in the Edinburgh BCS cohort.

In an exploratory analysis of a combined ICH4/Mammostrat score, the combined score was prognostic of outcome (up to 10 years) in all subgroups in the TEAM cohort. In the Edinburgh BCS cohort, the combined score had a time-varying effect in all patient subgroups and identified a group of patients with a very high risk of distant recurrence within the first two years since diagnosis.

## 8.7 Discussion

In this chapter, the impact of follow-up duration on IHC4 and Mammostrat score was explored.

Piece-wise constant effects, splitting time into intervals, is a commonly used approach to deal with non-proportional hazards. The difficulties associated with this approach were highlighted, in particular, the loss of power from categorisation which resulted in wide confidence intervals for estimates and a lack of power to detect significant effects at later time-points. However, it is a good starting point for detecting non-proportional effects and can help aid interpretation if non-proportional effects are identified. It is recommended that cut-points should be pre-specified and have a clinical meaning. Intervals should contain an adequate number of individuals and, more importantly, events. The intervals should not be too large as to still violate the PH assumption and this should be assessed. It is important not to choose intervals based on minimising p-values as this will bias results.

Creating a continuous time-by-covariate interaction avoids the loss of power from time-categorisation but difficulties arise in determining the functional form of the time-varying covariate. The MFPT algorithm is one approach to deal with this and has the advantage of simultaneously determining variables with non-proportional effects and selecting the

functional form whilst keeping the overall type I error rate at a pre-specified level. However, to determine the functional form of a time-varying covariate, a split at each event time is required as the Cox model is fitted by evaluating the partial likelihood at each failure time. In larger datasets, splitting at each failure time is problematic and requires categorisation of survival times. Consideration to the choice of the time intervals must be given; Buchholz found that the MFPT procedure gives similar results for small categorisation intervals and propose that about 50 to 100 distinct event times can give results with sufficient precision (Buchholz, 2010).

Another method that was applied was the RP flexible parametric model which models the baseline hazard and time-varying effects with restricted cubic splines. This approach has the advantage that due to the model being parametric, splitting at event times is not required to model the time-varying effects and allows absolute predictions of survival and hazard rate. A novel measure, restricted mean survival time (RMST) which is the area under the survival curve up to a selected time point (t) can also be calculated easily from this method. It has a more intuitive interpretation than the hazard ratio and can be thought of as a 't-year life expectancy'. For example, treatment A increases your life expectancy by 2 years compared with treatment B. However, difficulties occur with convergence when fitting these models to small samples or fitting complex models with more than one time-varying effect.

The MFPT and RP models gave very similar results with almost identical hazard ratio estimates for variables satisfying the proportional hazards assumption. The estimated time-dependent hazard ratios were also similar between approaches.

Our analyses identified a strong time-varying effect of IHC4 score. The prognostic effect of IHC4 score on TTDR decreased with increasing follow-up time. IHC4 score appeared to be prognostic of early distant recurrence only (0-5 years) when categorising time into 5-yearly intervals. Previous analysis by Sgroi and colleagues also confirmed a significant prognostic

ability for IHC4 for early distant recurrence only (0-5 years) (Sgroi et al., 2013). However, looking at the score continuously over time, IHC4 score was prognostic of outcome up to approximately 7 years in the TEAM cohort but only up to 4 years in the Edinburgh BCS cohort for all ER-positive patients. The longer follow-up in the Edinburgh BCS cohort identified a protective effect of higher IHC4 scores beyond 11 years of follow-up.

The prognostic effect of Mammostrat score was consistent over time. It identified a group of patients with an increased risk of distant recurrence over full follow-up in the TEAM cohort (10 years) and the Edinburgh BCS cohort (25 years). These results suggest the possible use of Mammostrat score to predict the risk of late recurrence, which will need to be investigated further on other patient cohorts with long follow-up.

The performance of both scores was good, especially in the first 5 years of follow-up, with the combination of both scores significantly improving the ability to discriminate between events and non-events when compared to clinical factors only and good calibration between observed and predicted 5 year risk of TTDR. The IHC4 score provided more prognostic information on TTDR than the Mammostrat score in the first 5 years of follow-up except for all ER-positive patients in the larger TEAM cohort where the addition of both scores provided statistically significant information.

In a post-hoc analysis of a combined IHC4/Mammostrat score, this risk score was prognostic of outcome in all patient subgroups in the TEAM cohort and identified a group of patients at an increased risk of recurrence over the 10-year follow-up. In the Edinburgh BCS cohort, the combined score identified a group of patients with a high risk of distant recurrence in the first five years since diagnosis, especially within the first two years.

# Chapter 9: Analysis of Individual Biomarkers

## 9.1 Introduction

IHC4 score was identified as having a strong time-dependent effect with the prognostic effect of IHC4 decreasing over time. A combined IHC4 and Mammostrat score was prognostic of outcome and identified a group of patients at an increased risk of recurrence.

In this chapter, the individual biomarkers from IHC4 and Mammostrat residual risk panels were analysed to determine which have a prognostic effect on outcome and which were driving the time-dependency.

The assumptions of linearity and proportional hazards in the Cox proportional hazards model was assessed for the individual markers. A multivariable model was developed using the novel MFPT approach to identify which markers were prognostic of outcome and whether any of the markers demonstrated a time-varying effect. The markers were also assessed using RP flexible parametric models to determine any differences or advantages in using a more flexible approach of analysis.

## 9.2 Materials

This analysis will focus on the two main patient subgroups, all ER-positive patients and patients treated with endocrine therapy only in the Edinburgh BCS and TEAM cohorts.

## 9.3 Biomarkers

The nine IHC biomarkers considered in this analysis are ER (histoscores), PgR (histoscores), Ki67 (percentage of positive cells), HER2 (negative/positive) and SLC7A5, CEACAM5, NDRG1, HTF9C and p53 were considered as dichotomous (negative/positive).

## 9.4 Correlation

Initially, correlations between variables were estimated by Spearman's rank correlation coefficient for continuous–continuous associations, Pearson's chi-squared test for categorical-categorical associations and the likelihood ratio test statistic from a regression of a continuous variable with a categorical variable as a predictor in the model. Results are given for all ER-positive patients in Table 9.1 for the Edinburgh BCS cohort and Table 9.2 for the TEAM cohort. Strong evidence of modest correlations between the majority of variables was observed.

## 9.5 Assumptions of Linearity and Proportional Hazards

### 9.5.1 Linearity on the Log-hazard

The functional form of the continuous covariates (ER, PgR and Ki67) were initially assessed by plotting the Martingale residuals from a null Cox regression model against the continuous covariates (Figure 9.1). The assumption of linearity on the log-hazard appears to be reasonable.

Next, in a univariate Cox regression model the fractional polynomial algorithm identified ER and Ki67 to be non-linear in the TEAM cohort. The best fitting FP to model the non-linear effects of ER and Ki67 are illustrated in Figure 9.2. There appears to be some over-fitting of the non-linear effects, with an unrealistic decrease in the hazard associated with higher Ki67 due to a lack of patients with a high percentage of positive Ki67 cells.

### 9.5.2 Proportional Hazards

The proportional hazards assumption was assessed from a univariate Cox regression model by plotting the Schoenfeld residuals against survival time, displayed in Figure 9.3 for the Edinburgh BCS cohort and Figure 9.4 for the TEAM cohort, and with the Grambsch-Thernaue test of proportional hazards (Table 9.3). Several of the markers showed evidence of a non-proportional effect in the Edinburgh BCS cohort, in particular PgR, Ki67, SLC7A5, NDRG1

**Table 9.1** Associations between biomarkers and clinical variables in the Edinburgh BCS cohort

| **Continuous-Continuous Associations - Spearmans rho (p-value)** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | ER | PgR | Ki67 | Age | Tumour Size | | |
| PgR | 0.14 (<0.001) | | | | | | |
| Ki67 | -0.30 (<0.001) | -0.05 (0.10) | | | | | |
| Age | 0.27 (<0.001) | -0.13 (<0.001) | -0.10 (0.001) | | | | |
| Tumour Size | -0.10 (<0.001) | 0.002 (0.94) | 0.23 (<0.001) | 0.04 (0.21) | | | |
| Nodes Pos. | 0.003 (0.93) | 0.10 (<0.001) | 0.11 (<0.001) | 0.06 (0.06) | 0.27 (<0.001) | | |
| **Categorical-Categorical Associations - Pearson's chi-squared (p-value)** | | | | | | | |
| | HER2 | SLC7A5 | CEACAM5 | NDRG1 | HTF9C | P53 | |
| SLC7A5 | 5.4 (0.02) | | | | | | |
| CEACAM5 | 9.0 (0.003) | 5.3 (0.02) | | | | | |
| NDRG1 | 1.5 (0.22) | 55.5 (<0.001) | 0.07 (0.78) | | | | |
| HTF9C | 16.4 (<0.001) | 26.2 (<0.001) | 2.9 (0.09) | 2.08 (0.15) | | | |
| P53 | 11.8 (0.003) | 27.5 (<0.001) | 0.06 (0.81) | 9.5 (0.002) | 28.0 (<0.001) | | |
| Grade | 23.1 (<0.001) | 100.6 (<0.001) | 4.2 (0.13) | 68.6 (<0.001) | 15.3 (<0.001) | 93.4 (<0.001) | |
| **Continuous-Categorical Associations - Likelihood ratio statistic (p-value)** | | | | | | | |
| | HER2 | SLC7A5 | CEACAM5 | NDRG1 | HTF9C | P53 | Grade |
| ER | 3.4 (0.07) | 18.7 (<0.001) | 0.8 (0.36) | 0.6 (0.43) | 17.2 (<0.001) | 12.0 (<0.001) | 55.3 (<0.001) |
| PgR | 1.9 (0.16) | 17.4 (<0.001) | 5.5 (0.02) | 7.0 (0.008) | 11.5 (<0.001) | 2.0 (0.16) | 47.1 (<0.001) |
| Ki67 | 34.3 (<0.001) | 75.5 (<0.001) | 1.9 (0.17) | 24.4 (<0.001) | 32.0 (<0.001) | 49.7 (<0.001) | 207.3 (<0.001) |
| Age | 4.9 (0.03) | 0.4 (0.53) | 1.2 (0.27) | 5.8 (0.02) | 0.2 (0.70) | 12.4 (<0.001) | 5.8 (0.05) |
| Tumour Size | 3.8 (0.05) | 9.9 (0.002) | 3.8 (0.05) | 38.3 (<0.001) | 3.3 (0.07) | 5.7 (0.02) | 73.5 (<0.001) |
| Nodes Pos. | 1.3 (0.25) | 7.7 (0.005) | 0.2 (0.69) | 8.6 (0.003) | 0.7 (0.41) | 9.4 (0.002) | 5.7 (0.06) |

**Table 9.2** Associations between biomarkers and clinical variables in the TEAM cohort.

| **Continuous-Continuous Associations - Spearmans rho (p-value)** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | ER | PgR | Ki67 | Age | Tumour Size | | |
| PgR | 0.25 (<0.001) | | | | | | |
| Ki67 | -0.14 (<0.001) | -0.08 (<0.001) | | | | | |
| Age | 0.12 (<0.001) | 0.05 (0.004) | 0.05 (0.003) | | | | |
| Tumour Size | -0.09 (<0.001) | -0.06 (<0.001) | 0.07 (<0.001) | 0.12 (<0.001) | | | |
| Nodes Pos. | -0.02 (0.26) | -0.02 (0.26) | -0.04 (0.02) | 0.004 (0.83) | 0.2 (<0.001) | | |
| **Categorical-Categorical Associations - Pearson's chi-squared (p-value)** | | | | | | | |
| | HER2 | SLC7A5 | CEACAM5 | NDRG1 | HTF9C | P53 | |
| SLC7A5 | 115.0 (<0.001) | | | | | | |
| CEACAM5 | 15.1 (<0.001) | 8.7 (0.003) | | | | | |
| NDRG1 | 20.1 (<0.001) | 74.6 (<0.001) | 1.6 (0.21) | | | | |
| HTF9C | 40.2 (<0.001) | 122.83 (<0.001) | 12.8 (<0.001) | 4.1 (0.04) | | | |
| P53 | 47.0 (<0.001) | 145.0 (<0.001) | 0.03 (0.85) | 7.7 (0.006) | 52.9 (<0.001) | | |
| Grade | 92.9 (<0.001) | 281.3 (<0.001) | 0.13 (0.94) | 83.2 (<0.001) | 21.7 (<0.001) | 115.3 (<0.001) | |
| **Continuous-Categorical Associations - Likelihood ratio statistic (p-value)** | | | | | | | |
| | HER2 | SLC7A5 | CEACAM5 | NDRG1 | HTF9C | P53 | Grade |
| ER | 39.2 (<0.001) | 63.3 (<0.001) | 20.2 (<0.001) | 3.6 (0.06) | 51.8 (<0.001) | 26.7 (<0.001) | 59.5 (<0.001) |
| PgR | 126.0 (<0.001) | 89.8 (<0.001) | 3.2 (0.07) | 48.3 (<0.001) | 59.8 (<0.001) | 8.1 (0.005) | 62.1 (<0.001) |
| Ki67 | 74.6 (<0.001) | 233.0 (<0.001) | 0.0 (0.97) | 47.2 (<0.001) | 36.1 (<0.001) | 117.7 (<0.001) | 265.2 (<0.001) |
| Age | 8.1 (0.004) | 3.5 (0.06) | 21.9 (<0.001) | 0.4 (0.55) | 0.1 (0.76) | 8.3 (0.004) | 2.3 (0.32) |
| Tumour Size | 0.5 (0.50) | 0.3 (0.58) | 0.2 (0.68) | 10.0 (0.002) | 0.3 (0.61) | 10.0 (0.002) | 17.2 (<0.001) |
| Nodes Pos. | 7.0 (0.008) | 1.1 (0.31) | 0.02 (0.90) | 0.01 (0.92) | 5.2 (0.02) | 5.0 (0.03) | 5.6 (0.06) |

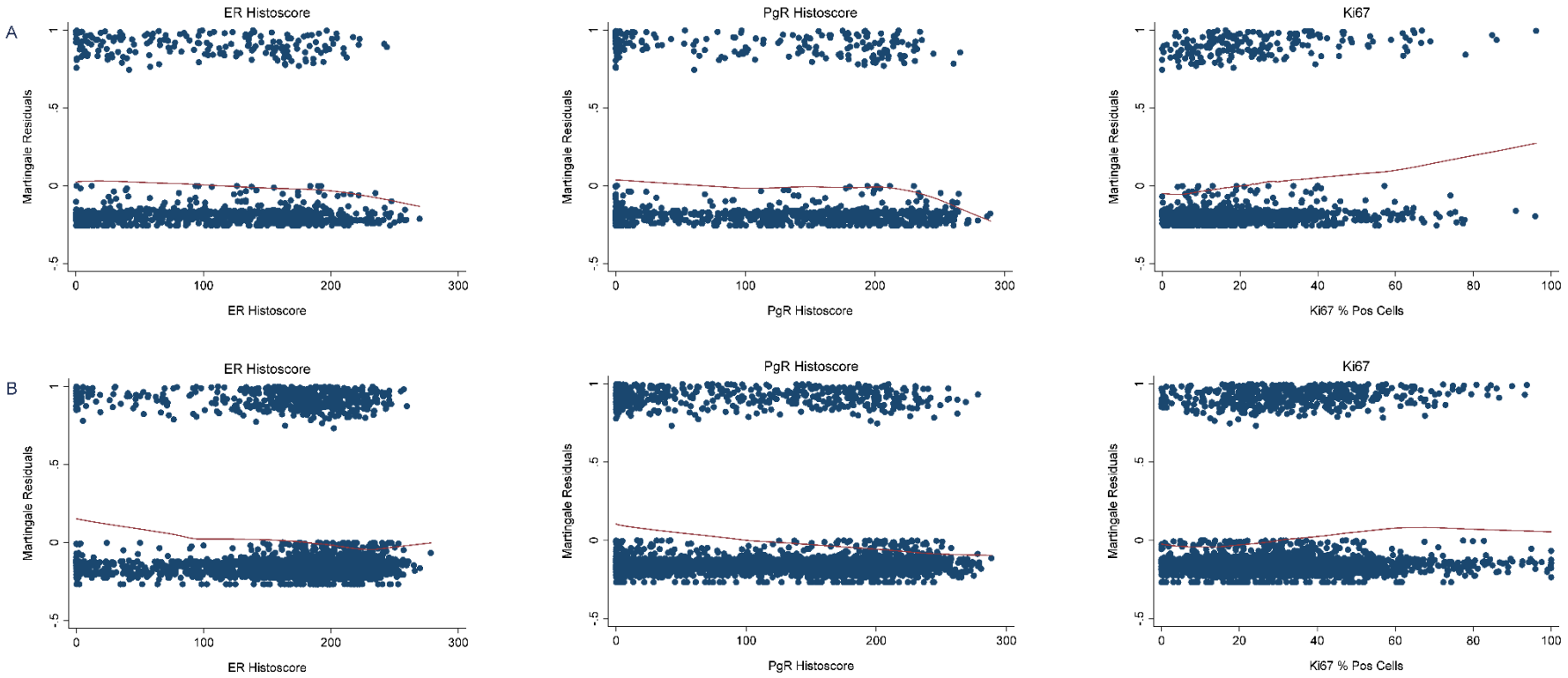



**Figure 9.1** Plots of martingale residuals versus continuous covariates. The red line is a smoothed curve produced by locally weighted regression. Plots represent (A) Edinburgh BCS and (B) TEAM cohort.

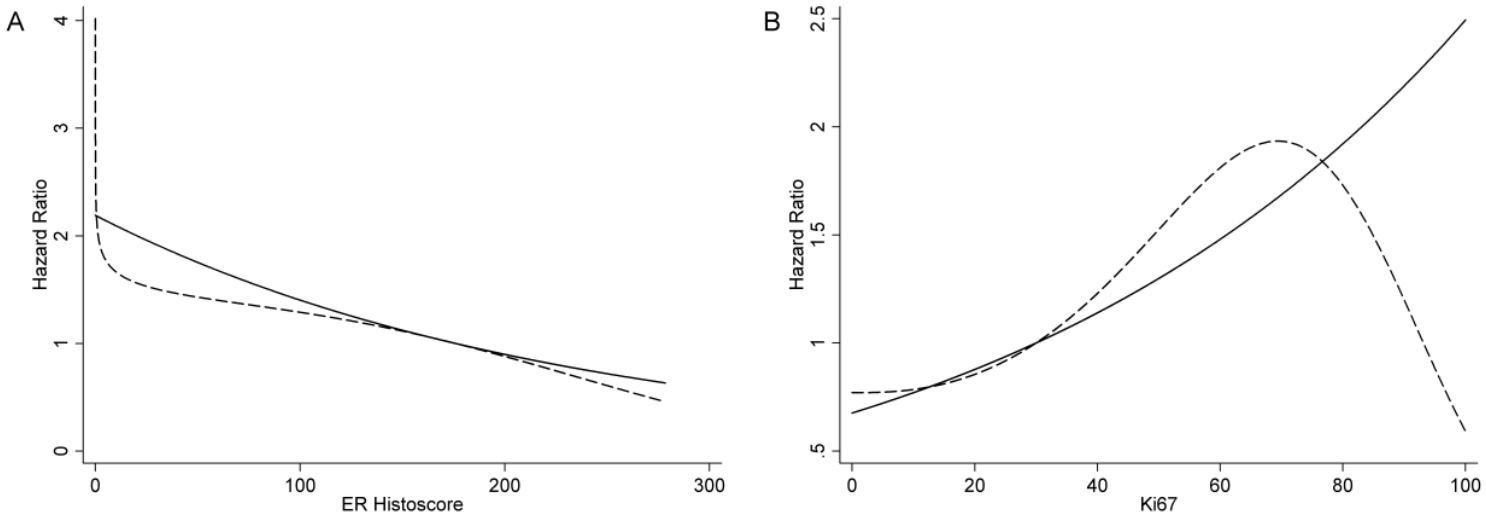



**Figure 9.2** Comparison of non-linear (dashed line) and linear (solid line) effects of (A) ER histoscores and (B) Ki67 in the TEAM cohort.

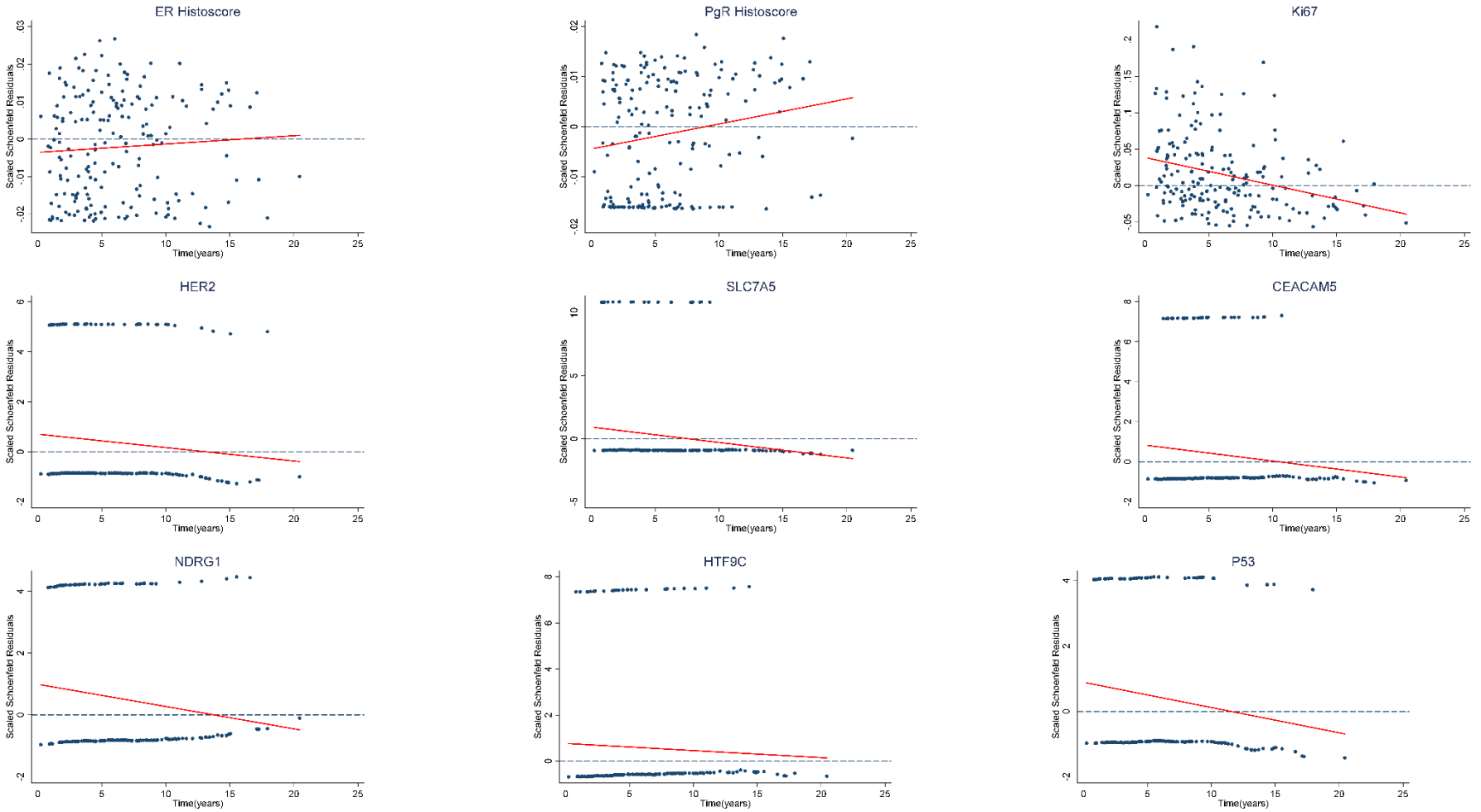



**Figure 9.3** Plots of Schoenfeld residuals for each explanatory variable versus survival time in the Edinburgh BCS cohort. Red (solid) line indicating the fitted linear regression line and the blue (dashed) line a reference line at zero.

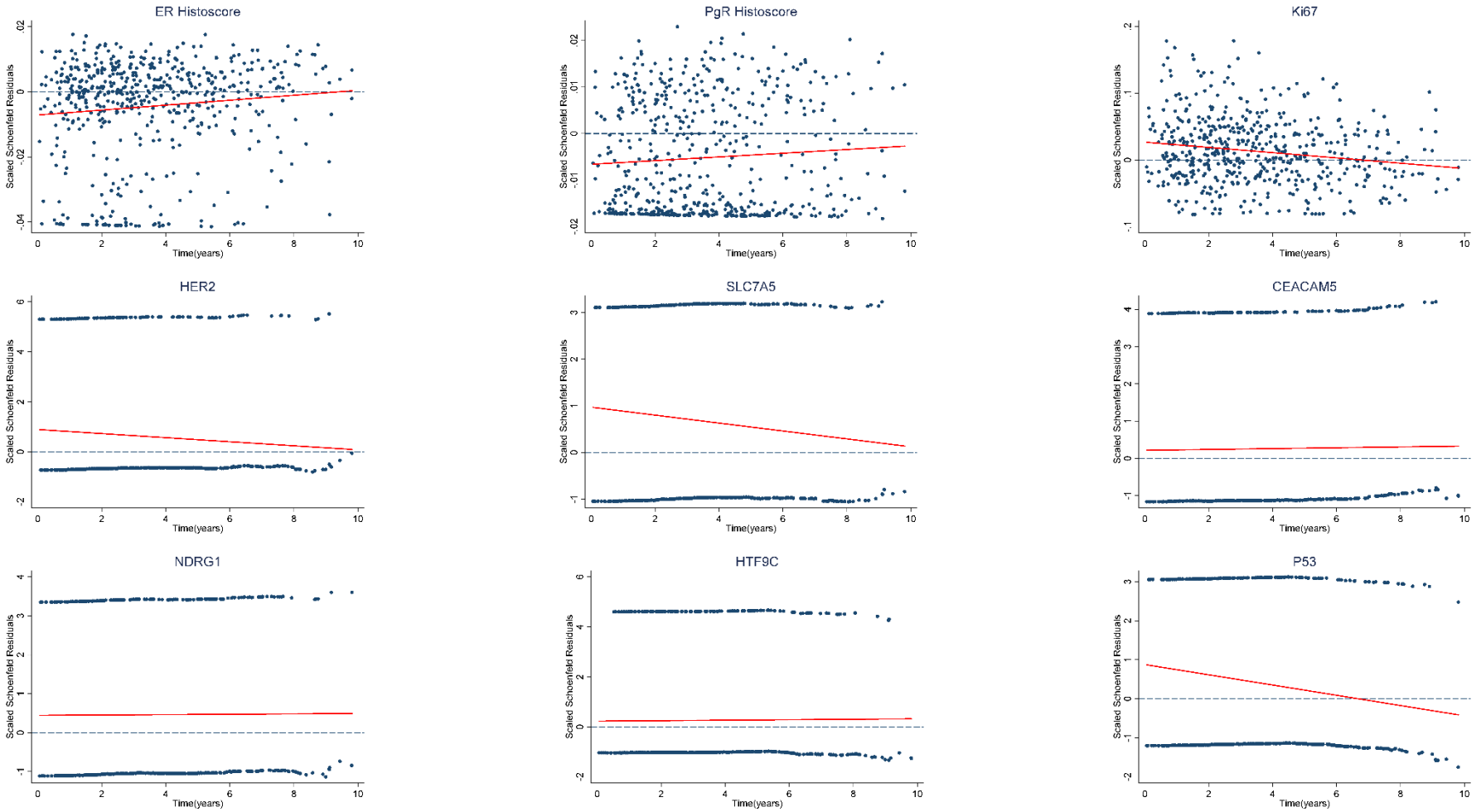



**Figure 9.4** Plots of Schoenfeld residuals for each explanatory variable versus survival time in the TEAM cohort. Red (solid) line indicating the fitted linear regression line and the blue (dashed) line a reference line at zero.

**Table 9.3** Grambsch-Therneau test of non-proportional hazards

| | P-value of Grambsch-Therneau Test | |
|---|---|---|
| Variable | Edinburgh BCS | TEAM |
| ER | 0.37 | 0.01 |
| PgR | 0.01 | 0.10 |
| HER2 | 0.21 | 0.10 |
| Ki67 | <0.001 | <0.001 |
| SLC7A5 | 0.04 | 0.04 |
| CEACAM5 | 0.11 | 0.80 |
| NDRG1 | 0.07 | 0.91 |
| HTF9C | 0.53 | 0.84 |
| P53 | 0.05 | 0.001 |

Note. Test of non-proportional hazards from a univariate Cox regression analysis.

and p53. In the TEAM cohort, ER, Ki67, SLC7A5 and p53 showed evidence of a non-proportional effect.

## 9.6 Multivariable Model

The multivariable fractional polynomial time approach (MFPT) was used to determine which variables are associated with TTDR, the functional form of the variables and any variables with non-proportional effects.

Step 1 uses the MFP algorithm to determine variables associated with the outcome over the full follow-up and the functional form to model the variables, calling this the time-fixed model.

Step 2 involves restricting the follow-up to contain the first half of the events, i.e. 3 years in the TEAM cohort and 5 years in the Edinburgh BCS cohort. The MFP algorithm is then run to determine if any covariates have a significant short-term effect and to determine the functional form of the variables.

Step 3 is then used to determine non-proportional effects of the variables identified in step 1 and step 2 of the MFPT approach using the FPT algorithm.

A 10% significance level was used for the selection of prognostic variables and 5% significance level for the selection of time-varying effects.

### 9.6.1 10 year follow-up

The MFPT approach was applied to the Edinburgh BCS cohort with follow-up censored at 10 years and the TEAM cohort.

#### *9.6.1.1 Biomarkers only*

The MFPT approach was initially performed including the 9 biomarkers only, not including the clinical variables.

*Edinburgh BCS Cohort*

The results of the model building for the Edinburgh BCS cohort are given in Table 9.4. All included variables had a linear effect on the log-hazard with Ki67 having a strong non-proportional effect. An additional marker, PgR, was identified as prognostic of outcome when considering ER-positive patients treated with endocrine therapy only. The final model contained markers from both the IHC4 and Mammostrat risk panels.

**Table 9.4** MFPT Results for the Edinburgh BCS Cohort (10 year follow-up) – biomarker only model

| | Edinburgh BCS | | | | | |
|---|---|---|---|---|---|---|
| | All ER-Positive | | | ER-Positive Endocrine Only | | |
| Variables | STEP 1 | STEP 2 | STEP 3 | STEP 1 | STEP 2 | STEP 3 |
| ER | | | | | | |
| PgR | | | | * | * | * |
| Ki67 | * | * | * | * | * | * |
| HER2 | | | | | | |
| p53 | * | * | | * | * | |
| NDRG1 | * | * | | * | * | |
| HTF9C | | | | | | |
| CEACAM5 | * | * | | * | * | |
| SLC7A5 | | | | | | |

Note: All variables included in steps 1 and 2 were determined to be linear. Step 2 includes variables selected in step 1 (using the whole time-period) and variables selected with a short-term effect (patients censored at 4 year follow-up). Step 3 indicates the variables in step 2 which have evidence of a non-proportional effect, with the best fitting transformation to be log(time) for Ki67 and time$^3$ for PgR.

*TEAM Cohort*

Results of the model building in the TEAM cohort are given in Table 9.5. All markers except HTF9C were identified as being prognostic of outcome over the full follow-up for all ER-positive patients. In the univariate analysis, section 9.5.1, ER and Ki67 were identified to be non-linear but there appeared to be some over-fitting. In the multivariable model all variables were determined to have a linear effect on the log-hazard.

Fewer markers were identified as having a prognostic effect on TTDR for ER-positive patients treated with endocrine therapy only. Ki67 was not identified as having a prognostic effect over the full follow-up (step 1), but was identified as prognostic of outcome when follow-up was censored at 3 years (step 2). This highlights the importance of considering both short and long-term follow-up. The majority of biomarkers identified by the MFPT approach for ER-positive patients treated with endocrine therapy only were from the IHC4 model, with only SLC7A5 from the Mammostrat model being prognostic of outcome. Again Ki67 was identified as having a strong non-proportional effect, with p53 also having a time-varying effect for all ER-positive patients. The importance of considering other transformations of time was highlighted here, with the best fitting FP to model the interaction between p53 and time was time-squared (Figure 9.5).

**Table 9.5** MFPT Results for the TEAM cohort – biomarker only model

| | TEAM | | | | | |
|---|---|---|---|---|---|---|
| | All ER-Positive | | | ER-Positive Endocrine Only | | |
| Variables | STEP 1 | STEP 2 | STEP 3 | STEP 1 | STEP 2 | STEP 3 |
| ER | * | * | | * | * | |
| PgR | * | * | | * | * | |
| Ki67 | * | * | * | | * | * |
| HER2 | * | * | | * | * | |
| p53 | * | * | * | | | |
| NDRG1 | * | * | | | | |
| HTF9C | | | | | | |
| CEACAM5 | * | * | | | | |
| SLC7A5 | * | * | | * | * | |

Note: All variables included in steps 1 and 2 were determined to be linear. Step 2 includes variables selected in step 1 (using the whole time-period) and variables selected with a short-term effect (patients censored at 3 year follow-up). Step 3 indicates the variables in step 2 which have evidence of a non-proportional effect, with the best fitting transformation to be log(time) for Ki67 and time$^2$ for p53.

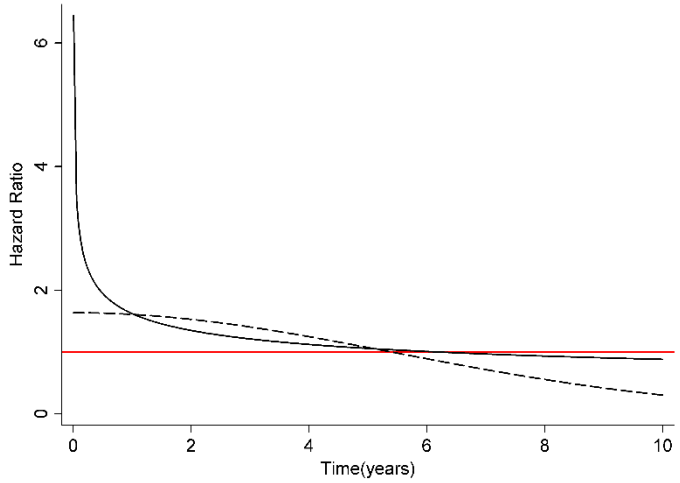



**Figure 9.5** Illustration of different fractional polynomials to model interaction between p53 and time: log-time (solid line) and time-squared (dashed line). Red line at hazard ratio of 1 indicates null effect.

### *9.6.1.2 Biomarkers and clinical variables*

The MFPT algorithm was then performed forcing the clinical variables (age, grade, non-linear tumour size and number of positive nodes, treatment and chemotherapy) into the model.

*Edinburgh BCS Cohort*

Results are given in Table 9.6 for the Edinburgh BCS cohort. Fewer markers remained prognostic of outcome when clinical variables were included in the model. For all ER-positive patients in the Edinburgh BCS cohort, HTF9C was identified as an important predictor by the MFPT model building approach when clinical variables were included in the model, whereas it was not prognostic of outcome when only biomarkers were considered. With the inclusion of the clinical variables, Ki67 was only prognostic of outcome when follow-up was restricted to 4 years (step 2) and again was identified as having a strong non-proportional effect.

**Table 9.6** MFPT Results for the Edinburgh BCS cohort (10 year follow-up) – biomarker and clinical model

| | Edinburgh BCS | | | | | |
|---|---|---|---|---|---|---|
| | All ER-Positive | | | ER-Positive Endocrine Only | | |
| Variables | STEP 1 | STEP 2 | STEP 3 | STEP 1 | STEP 2 | STEP 3 |
| ER | | | | | | |
| PgR | | | | | | |
| Ki67 | | * | * | | * | * |
| HER2 | | | | | | |
| p53 | | | | | | |
| NDRG1 | | | | | * | |
| HTF9C | * | * | | | | |
| CEACAM5 | * | * | | * | * | |
| SLC7A5 | | | | | | |

Note: All variables included in steps 1 and 2 were determined to be linear. Step 2 includes variables selected in step 1 (using the whole time-period) and variables selected with a short-term effect (patients censored at 4 year follow-up). Step 3 indicates the variables in step 2 which have evidence of a non-proportional effect, with the best fitting transformation to be log(time) for Ki67. Clinical variables (age, grade, no. positive nodes, tumour size and chemotherapy) were included in the model.

*TEAM Cohort*

Results from model building using the MFPT approach with biomarkers and clinical variables for the TEAM cohort are given in Table 9.7. Despite a strong time-varying effect of Ki67 in the biomarker only model, Ki67 was not identified as being prognostic of outcome when clinical variables were included in the model. This warranted further explanation due to the previous strong effect of Ki67 in the short-term (up to 3 years follow-up). The MFP algorithm first tests the best fitting degree 2 FP model of Ki67 against not including it in the model and this was non-significant (increase in model deviance 4.4, p-value 0.36). We previously suspected some over-fitting of ER and Ki67 when modelled with a degree 2 FP (Figure 9.2). If we restrict ER and Ki67 to be linear on the log-hazard, Ki67 is prognostic of outcome in the short-term and included in step 2 of the MFPT approach (Table 9.8). This identified an issue with the MFP algorithm, where a variable is rejected due to over-fitting of the non-linear effect in the model.

**Table 9.7** MFPT Results for the TEAM cohort – biomarker and clinical model

| | TEAM | | | | | |
|---|---|---|---|---|---|---|
| | All ER-Positive | | | ER-Positive Endocrine Only | | |
| Variables | STEP 1 | STEP 2 | STEP 3 | STEP 1 | STEP 2 | STEP 3 |
| ER | * | * | | | | |
| PgR | * | * | | * | * | |
| Ki67 | | | | | | |
| HER2 | | | | * | * | |
| p53 | * | * | * | | | |
| NDRG1 | * | * | | | | |
| HTF9C | | | | | | |
| CEACAM5 | * | * | | | | |
| SLC7A5 | * | * | | * | * | |

Note: All variables included in steps 1 and 2 were determined to be linear. Step 2 includes variables selected in step 1 (using the whole time-period) and variables selected with a short-term effect (patients censored at 3 year follow-up). Step 3 indicates the variables in step 2 which have evidence of a non-proportional effect, with the best fitting transformation to be time$^2$ for p53. Clinical variables (age, grade, no. positive nodes, tumour size, treatment and chemotherapy) were included in the model.

**Table 9.8** MFPT Results for the TEAM cohort restricting ER and Ki67 to be linear – biomarker and clinical model

| | TEAM | | | | | |
|---|---|---|---|---|---|---|
| | All ER-Positive | | | ER-Positive Endocrine Only | | |
| Variables | STEP 1 | STEP 2 | STEP 3 | STEP 1 | STEP 2 | STEP 3 |
| ER | * | * | | * | * | |
| PgR | * | * | | * | * | |
| Ki67 | | * | * | | | |
| HER2 | | | | * | * | |
| p53 | * | * | * | | | |
| NDRG1 | * | * | | | | |
| HTF9C | | | | | | |
| CEACAM5 | * | * | | | | |
| SLC7A5 | * | * | | * | * | |

Note: All variables included in steps 1 and 2 were determined to be linear. Step 2 includes variables selected in step 1 (using the whole time-period) and variables selected with a short-term effect (patients censored at 3 year follow-up). Step 3 indicates the variables in step 2 which have evidence of a non-proportional effect, with the best fitting transformation to be log-time for Ki67 and time$^3$ for p53. Clinical variables (age, grade, no. positive nodes, tumour size, treatment and chemotherapy) were included in the model.

Restricting ER and Ki67 to be linear (or considering degree 1 FPs only) in the biomarker only model also resulted in different variable selection (Table 9.9) for ER-positive patients treated with endocrine therapy only. Ki67 was identified as prognostic over the full follow-up (step 1) and HER2 was no longer identified as prognostic of outcome.

**Table 9.9** MFPT Results for the TEAM cohort restricting ER and Ki67 to be linear – biomarker only model

| | TEAM | | | | | |
|---|---|---|---|---|---|---|
| | All ER-Positive | | | ER-Positive Endocrine Only | | |
| Variables | STEP 1 | STEP 2 | STEP 3 | STEP 1 | STEP 2 | STEP 3 |
| ER | * | * | | * | * | |
| PgR | * | * | | * | * | |
| Ki67 | * | * | * | * | * | * |
| HER2 | * | * | | | | |
| p53 | * | * | * | | | |
| NDRG1 | * | * | | | | |
| HTF9C | | | | | | |
| CEACAM5 | * | * | | | | |
| SLC7A5 | * | * | | * | * | |

Note: All variables included in steps 1 and 2 were determined to be linear. Step 2 includes variables selected in step 1 (using the whole time-period) and variables selected with a short-term effect (patients censored at 3 year follow-up). Step 3 indicates the variables in step 2 which have evidence of a non-proportional effect, with the best fitting transformation to be log(time) for Ki67 and time$^2$ for p53.

*Edinburgh BCS Cohort*

The overfitting of Ki67 with a degree 2 FP then prompted further checks on the model building in the Edinburgh BCS cohort. An issue was identified with PgR, with a degree 2 FP resulting in a non-significant increase in model deviance compared to not including the marker in the model (increase in model deviance 4.4, p-value 0.4). Restricting PgR to be linear, resulted in a significant increase in model deviance (increase in model deviance 4.9, p-value 0.03) and was included in step 1 of the MFPT model building for all ER-positive patients (Table 9.10). However, there does not appear to be any serious over-fitting when comparing a linear effect on the log-hazard to a degree 2 FP (Figure 9.6).

**Table 9.10** MFPT Results for the Edinburgh BCS cohort (10 year follow-up) restricting PgR to be linear – biomarker only model

| | Edinburgh BCS | | | | | |
|---|---|---|---|---|---|---|
| | All ER-Positive | | | ER-Positive Endocrine Only | | |
| Variables | STEP 1 | STEP 2 | STEP 3 | STEP 1 | STEP 2 | STEP 3 |
| ER | | | | | | |
| PgR | * | * | * | * | * | * |
| Ki67 | * | * | * | * | * | * |
| HER2 | | | | | | |
| p53 | * | * | | * | * | |
| NDRG1 | * | * | | * | * | |
| HTF9C | | | | | | |
| CEACAM5 | * | * | | * | * | |
| SLC7A5 | | | | | | |

Note: All variables included in steps 1 and 2 were determined to be linear. Step 2 includes variables selected in step 1 (using the whole time-period) and variables selected with a short-term effect (patients censored at 4 year follow-up). Step 3 indicates the variables in step 2 which have evidence of a non-proportional effect, with the best fitting transformation to be log-time for Ki67 and time$^3$ for PgR.

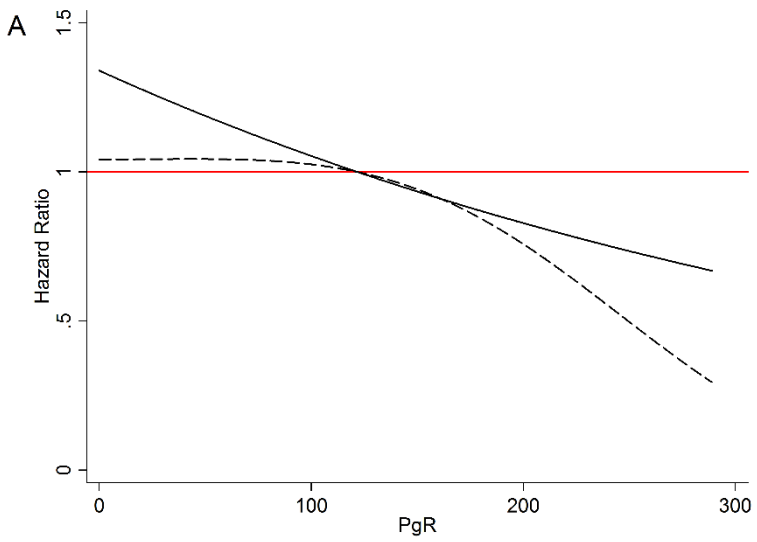



**Figure 9.6** Comparison of non-linear (dashed line) and linear (solid line) effects of PgR in the Edinburgh BCS cohort (10-year follow-up). Red line at hazard ratio of 1 indicates null effect.

Restricting PgR to be linear in the biomarker and clinical model also resulted in differences in variable selection (Table 9.11), with PgR included as prognostic of outcome over the 10-year

follow-up (step 1) and HTF9C no longer being prognostic of outcome for all ER- positive patients (step 1) and NDRG1 not prognostic of outcome in the short-term (step 2) for ER-positive patients treated with endocrine therapy only.

**Table 9.11** MFPT Results for the Edinburgh BCS cohort (10 year follow-up) restricting PgR to be linear – biomarker and clinical model

| | Edinburgh BCS | | | | | |
|---|---|---|---|---|---|---|
| | All ER-Positive | | | ER-Positive Endocrine Only | | |
| Variables | STEP 1 | STEP 2 | STEP 3 | STEP 1 | STEP 2 | STEP 3 |
| ER | | | | | | |
| PgR | * | * | | * | * | |
| Ki67 | | * | * | | * | * |
| HER2 | | | | | | |
| p53 | | | | | | |
| NDRG1 | | | | | | |
| HTF9C | | | | | | |
| CEACAM5 | * | * | | * | * | |
| SLC7A5 | | | | | | |

Note: All variables included in steps 1 and 2 were determined to be linear. Step 2 includes variables selected in step 1 (using the whole time-period) and variables selected with a short-term effect (patients censored at 4 year follow-up). Step 3 indicates the variables in step 2 which have evidence of a non-proportional effect, with the best fitting transformation to be log(time) for Ki67. Clinical variables (age, grade, no. positive nodes, tumour size and chemotherapy) were included in the model.

#### *9.6.1.3 Cox Regression Models*

The estimated coefficients and hazard ratios from the final model identified by the MFPT approach are given in Table 9.12 for the Edinburgh BCS cohort and Table 9.13 for the TEAM cohort. The time-varying effects are displayed in Figure 9.7 and Figure 9.8 for the Edinburgh BCS and TEAM cohorts respectively.

The markers from the Mammostrat model that were independent predictors of outcome (SLC7A5, CEACAM5, NDRG1, p53) in either the Edinburgh BCS and TEAM cohort, having a positive marker increased the hazard compared to being marker negative.

**Table 9.12** Multivariate Cox Regression of biomarkers identified by the MFPT approach in the Edinburgh BCS Cohort (10 year follow-up)

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Coeff (Std. Err.) | HR (95% CI) | Coeff (Std. Err.) | HR (95% CI) |
| **Biomarker Only Model** | | | | |
| ER (50 units) | - | - | - | - |
| PgR (50 units) | -0.14 (0.06) | 0.87 (0.77-0.97) | -.20 (0.07) | 0.82 (0.71-0.94) |
| Ki67 (10 units) | 0.37 (0.08) | 1.45 (1.23-1.70) | 0.37 (0.10) | 1.45 (1.19-1.77) |
| HER2 | - | - | - | - |
| SLC7A5 | - | - | - | - |
| CEACAM5 | 0.45 (0.22) | 1.57 (1.03-2.40) | 0.56 (0.24) | 1.76 (1.09-2.83) |
| NDRG1 | 0.45 (0.18) | 1.56 (1.10-2.22) | 0.50 (0.20) | 1.65 (1.11-2.47) |
| HTF9C | - | - | - | - |
| p53 | 0.32 (0.18) | 1.37 (0.96-1.96) | 0.50 (0.21) | 1.66 (1.11-2.48) |
| PgR*time3 | 0.0003 (0.0002) | 1.0003 (0.9999-1.0007) | 0.0004 (0.0002) | 1.0004 (0.9999-1.0008) |
| Ki67*log(time) | -0.16 (0.06) | 0.85 (0.76-0.95) | -0.18 (0.07) | 0.84 (0.73-0.96) |
| **Biomarker and Clinical Model*** | | | | |
| ER (50 units) | - | - | - | - |
| PgR (50 units) | -0.10 (0.05) | 0.90 (0.82-0.99) | -0.11 (0.06) | 0.89 (0.80-1.00) |
| Ki67 (10 units) | 0.30 (0.09) | 1.35 (1.13-1.60) | 0.29 (0.10) | 1.34 (1.10-1.63) |
| HER2 | - | - | - | - |
| SLC7A5 | - | - | - | - |
| CEACAM5 | 0.41 (0.22) | 1.51 (0.98-2.31) | 0.54 (0.25) | 1.72 (1.06-2.79) |
| NDRG1 | - | - | - | - |
| HTF9C | - | - | - | - |
| p53 | - | - | - | - |
| Ki67*log(time) | -0.18 (0.06) | 0.84 (0.74-0.94) | -0.19 (0.07) | 0.83 (0.72-0.95) |

NOTE. Multivariate Cox regression of the biomarkers in the final model identified by the MFPT algorithm. Coefficients (standard error) and hazard ratios (95% CIs) for risk of TTDR calculated for a 50 unit increase in ER or PgR, a 10 unit increase in Ki67 and positive compared to negative for all other markers. Values calculated for all ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh cohort with follow-up censored at 10 years. Abbreviations: TTDR, time to distant-recurrence; Coeff, coefficient; Std. Err, standard error; CI, confidence interval.
* Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

**Table 9.13** Multivariate Cox Regression of biomarkers identified by the MFPT approach in the TEAM Cohort.

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Coeff (Std. Err.) | HR (95% CI) | Coeff (Std. Err.) | HR (95% CI) |
| **Biomarker Only Model** | | | | |
| ER (50 units) | -0.08 (0.03) | 0.93 (0.86-0.99) | -0.14 (0.05) | 0.87 (0.80-0.96) |
| PgR (50 units) | -0.20 (0.03) | 0.82 (0.78-0.86) | -0.17 (0.03) | 0.84 (0.79-0.90) |
| Ki67 (10 units) | 0.11 (0.04) | 1.12 (1.05-1.20) | 0.12 (0.05) | 1.13 (1.03-1.24) |
| HER2 | 0.22 (0.11) | 1.25 (1.00-1.55) | | |
| SLC7A5 | 0.36 (0.36) | 1.43 (1.19-1.73) | 0.46 (0.12) | 1.58 (1.24-2.01) |
| CEACAM5 | 0.30 (0.10) | 1.34 (1.11-1.63) | - | - |
| NDRG1 | 0.26 (0.09) | 1.30 (1.08-1.56) | - | - |
| HTF9C | - | - | - | - |
| p53 | 0.46 (0.12) | 1.59 (1.25-2.02) | - | - |
| Ki67*log(time) | -0.07 (0.03) | 0.93 (0.88-0.98) | -0.06 (0.04) | 0.94 (0.88-1.01) |
| p53*time$^2$ | -0.01 (0.005) | 0.99 (0.98-1.00) | - | - |
| **Biomarker and Clinical Model*** | | | | |
| ER (50 units) | -0.08 (0.04) | 0.93 (0.86-0.99) | -0.12 (0.05) | 0.89 (0.80-0.97) |
| PgR (50 units) | -0.17 (0.03) | 0.84 (0.80-0.89) | -0.12 (0.04) | 0.89 (0.83-0.95) |
| Ki67 (10 units) | 0.09 (0.04) | 1.09 (1.01-1.18) | - | - |
| HER2 | - | - | 0.27 (0.16) | 1.31 (0.96-1.77) |
| SLC7A5 | 0.32 (0.10) | 1.38 (1.13-1.68) | 0.32 (0.13) | 1.38 (1.07-1.78) |
| CEACAM5 | 0.27 (0.10) | 1.31 (1.07-1.60) | - | - |
| NDRG1 | 0.32 (0.10) | 1.38 (1.14-1.67) | - | - |
| HTF9C | - | - | - | - |
| p53 | 0.43 (0.11) | 1.54 (1.23-1.92) | - | - |
| Ki67*log(time) | -0.08 (0.03) | 0.92 (0.87-0.98) | - | - |
| p53*time$^3$ | -0.002 (0.0007) | 0.998 (0.997-0.999) | - | - |

NOTE. Multivariate Cox regression of the biomarkers in the final model identified by the MFPT algorithm. Coefficients (standard error) and hazard ratios (95% CIs) for risk of TTDR calculated for a 50 unit increase in ER or PgR, a 10 unit increase in Ki67 and positive compared to negative for all other markers. Values calculated for all ER-positive patients and ER-positive patients who received no chemotherapy in the TEAM cohort. Abbreviations: TTDR, time to distant-recurrence; Coeff, coefficient; Std. Err, standard error; CI, confidence interval. * Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

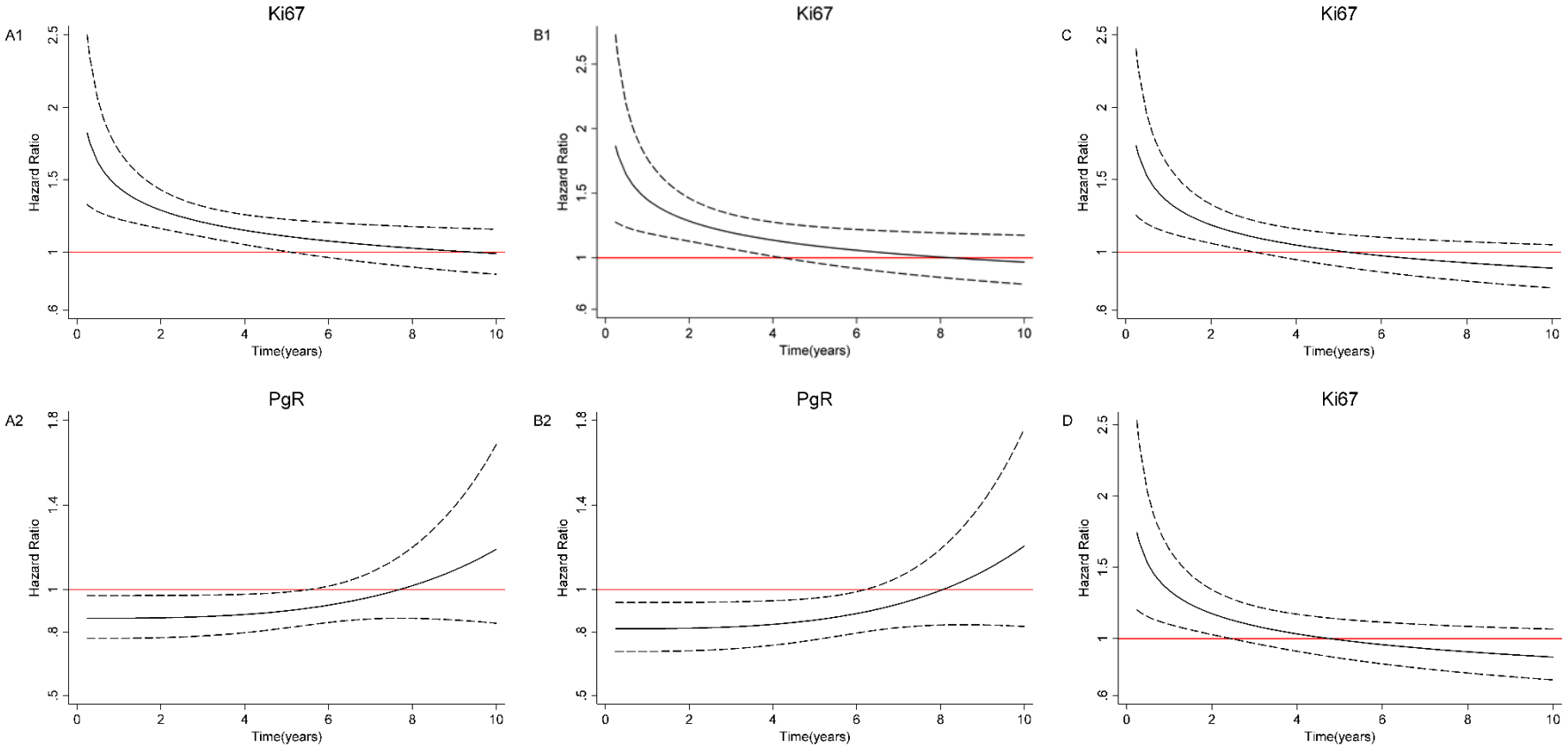



**Figure 9.7** Edinburgh BCS cohort. Time-varying adjusted hazard ratios for all ER-positive patients (A and C) and ER-positive patients treated with endocrine therapy only (B and D) in the Edinburgh BCS cohort from the biomarker only model (A and B) and the biomarker and clinical model (age, grade, tumour size, number positive nodes and chemotherapy) (C and D).

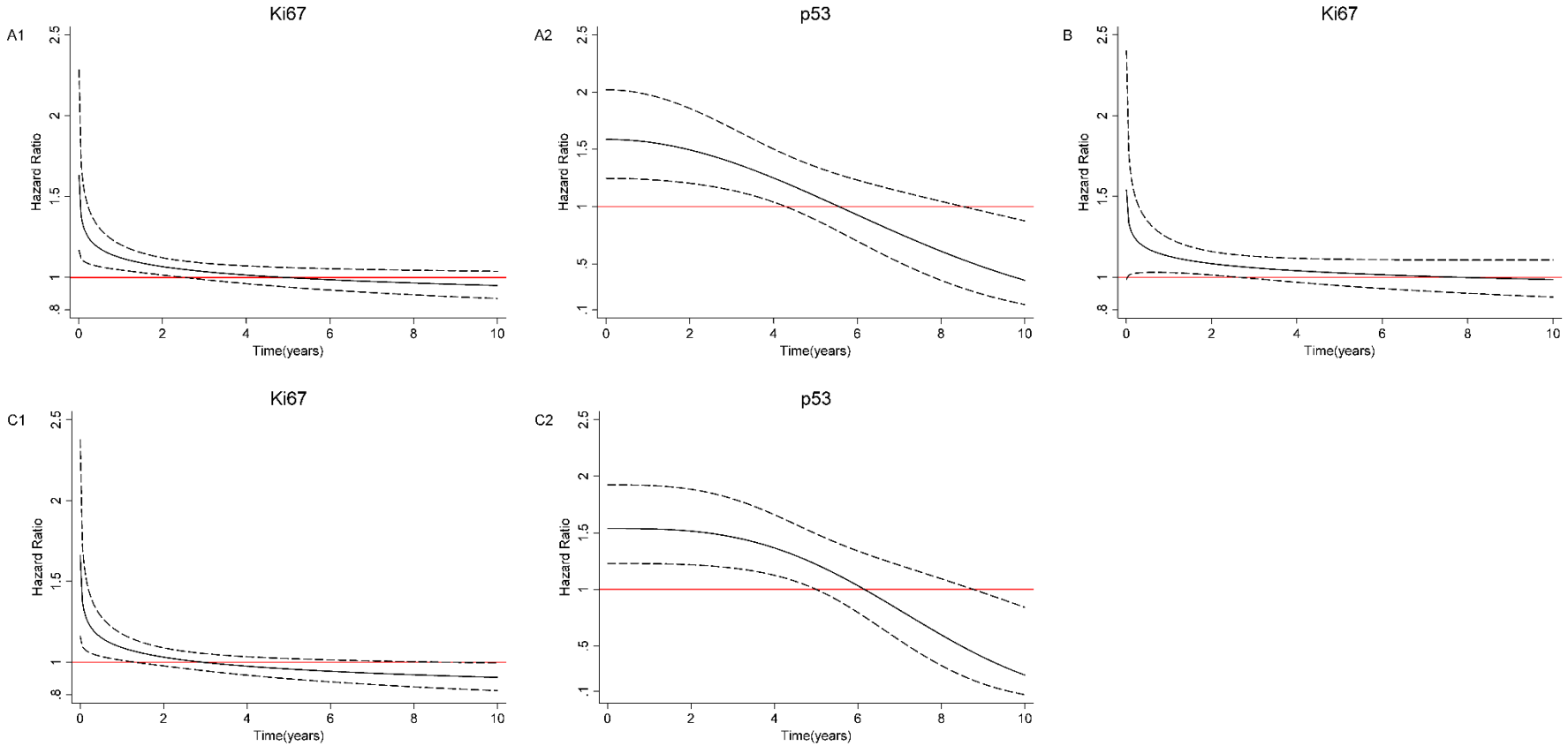



**Figure 9.8** TEAM cohort. Time-varying adjusted hazard ratios for all ER-positive patients (A and C) and ER-positive patients treated with endocrine therapy only (B) in the TEAM cohort from the biomarker only model (A and B) and the biomarker and clinical model (age, grade, tumour size, number positive nodes, treatment and chemotherapy) (C).

In both cohorts, a higher percentage of positive Ki67 cells was associated with an increased risk of TTDR in the short-term, with the effect decreasing over time. PgR was also identified as having a time-varying effect in the biomarker only model in the Edinburgh BCS cohort, where higher PgR histoscores were associated with a decreased hazard up to approximately 6 years compared with lower histoscores. P53 was identified as having a time-varying effect in both the biomarker only and biomarker and clinical model for all ER-positive patients in the TEAM cohort. P53-positive was associated with an increased hazard up to approximately 4 years follow-up compared to p53-negative, the effect decreased over time and p53-positive was associated with a decreased hazard beyond approximately 8 years.

ER was only identified as an important predictor of outcome in the TEAM cohort, with higher ER histoscores associated with a decreased hazard compared to lower histoscores.

One issue with the MFPT approach is in the final model there can be non-significant predictors due to having 3 steps in the model building. For example, in the biomarker and clinical model for all ER-positive patients in the Edinburgh BCS cohort, CEACAM5 was identified as prognostic of outcome in step 1 (considering full follow-up) with a hazard ratio estimate of 1.53 (95% CI, 1.00-2.34). With the addition of Ki67 in step 2 and a time-varying effect included in the model, CEACAM5 becomes borderline significant at the 5% level with a hazard ratio estimate of 1.51 (95% CI, 0.98-2.31).

#### *9.6.1.4 Discrimination*

The discrimination was compared using Royston & Sauerbrei's $R^2$ and D-statistic between the model identified in step 1 of the MFPT approach; variables prognostic of outcome over the full follow-up, and the final model; variables prognostic of both short and long-term outcome and the inclusion of any time-varying effects.

The biomarker only model was considered with and without adjustment for clinical variables as well as the biomarker and clinical model to compare whether the additional markers in the

biomarker only model improved the models ability to discriminate between events and non-events.

Results are given in Table 9.14 for the Edinburgh BCS cohort with follow-up censored at 10 years. There was improved prognostic separation between step 1 of the MFPT model building approach and the final model for all 3 models and both patient subgroups with increases in $R^2$ between 3.5 and 8.0% and increases in D-statistic between 0.11 and 0.28. The extra biomarkers in the adjusted biomarker model did not improve the prognostic ability compared to the biomarker and clinical model with similar $R^2$ and D-statistics for all ER-positive patients ($R^2$: 29.3 versus 28.4%, D-statistic: 1.32 versus 1.29). There was a larger improvement in model discrimination for ER-positive patients treated with endocrine therapy only for the adjusted biomarker model, however this was not statistically significant ($R^2$: 32.2 versus 30%, D-statistic: 1.41 versus 1.34).

**Table 9.14** Edinburgh BCS cohort (10-year follow-up). Comparison of discrimination of model identified in step 1 of the MFPT approach and the final model.

| | Edinburgh BCS | | | | | |
|---|---|---|---|---|---|---|
| | $R^2$ (%) | | | D-Statistic | | |
| | Step 1 | Final | Increase | Step 1 | Final | Increase |
| **All ER-Positive** | | | | | | |
| Biomarker Only Model | 8.0 | 15.3 | 7.3 | 0.61 | 0.87 | 0.27 |
| Adjusted Biomarker Model | 24.3 | 29.3 | 5.0 | 1.16 | 1.32 | 0.16 |
| Biomarker and Clinical Model | 24.3 | 28.4 | 4.1 | 1.16 | 1.29 | 0.13 |
| **ER-Positive Endocrine Only** | | | | | | |
| Biomarker Only Model | 9.8 | 17.7 | 8.0 | 0.67 | 0.95 | 0.28 |
| Adjusted Biomarker Model | 26.7 | 32.2 | 5.4 | 1.24 | 1.41 | 0.17 |
| Biomarker and Clinical Model | 26.5 | 30.0 | 3.5 | 1.23 | 1.34 | 0.11 |

Note. Values represent $R^2$ and D-statistic for the models identified by step 1 and the final model of the MFPT approach, with a difference in D of at least 0.1 indicating improved prognostic separation. Results are given for all ER-positive patients and ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort with follow-up censored at 10 years. The models represent: biomarker only model, model selected by MFPT including biomarkers only; Adjusted biomarker model, model selected by MFPT including biomarkers only then adjusting for clinical variables (age, grade, tumour size, nodes positive, chemotherapy); Biomarker and clinical model, model selected by MFPT including biomarkers and clinical variables.

The results for the TEAM cohort are given in Table 9.15. There were no significant improvements in model discrimination between the model identified by step 1 of the MFPT approach and the final model with increases in $R^2$ between 0 and 2.6% and increases in D-statistic between 0 and 0.08. There were very similar results between the adjusted biomarker model and the biomarker and clinical model.

**Table 9.15** TEAM cohort. Comparison of discrimination of model identified in step 1 of the MFPT approach and the final model.

| | TEAM | | | | | |
|---|---|---|---|---|---|---|
| | $R^2$ (%) | | | D-Statistic | | |
| | Step 1 | Final | Increase | Step 1 | Final | Increase |
| **All ER-Positive** | | | | | | |
| Biomarker Only Model | 17.2 | 19.8 | 2.6 | 0.93 | 1.02 | 0.08 |
| Adjusted Biomarker Model | 32.9 | 35.2 | 2.2 | 1.43 | 1.51 | 0.07 |
| Biomarker and Clinical Model | 32.9 | 34.8 | 2.0 | 1.43 | 1.50 | 0.06 |
| **ER-Positive Endocrine Only** | | | | | | |
| Biomarker Only Model | 14.5 | 15.7 | 1.2 | 0.84 | 0.88 | 0.04 |
| Adjusted Biomarker Model | 33.1 | 33.6 | 0.5 | 1.44 | 1.46 | 0.02 |
| Biomarker and Clinical Model | 33.3 | 33.3 | 0.0 | 1.45 | 1.45 | 0.00 |

Note. Values represent $R^2$ and D-statistic for the models identified by step 1 and the final model of the MFPT approach, with a difference in D of at least 0.1 indicating improved prognostic separation. Results are given for all ER-positive patients and ER-positive patients treated with endocrine therapy only in the TEAM cohort. The models represent: biomarker only model, model selected by MFPT including biomarkers only; Adjusted biomarker model, model selected by MFPT including biomarkers only then adjusting for clinical variables (age, grade, tumour size, nodes positive, chemotherapy); Biomarker and clinical model, model selected by MFPT including biomarkers and clinical variables.

### 9.6.2 Full follow-up

The MFPT approach was also applied to the Edinburgh BCS cohort considering full follow-up (maximum 25 years).

### *9.6.2.1 Biomarkers only*

The results of the model building for the Edinburgh BCS cohort considering full follow-up are given in Table 9.16. For all ER-positive patients, the same biomarkers were not identified when follow-up was censored at 10 years (see Table 9.10). HTF9C was prognostic of outcome and p53 was not identified as being prognostic over the full follow-up (step 1) or the five-year follow-up (step 2). PgR and CEACAM5 were not prognostic of outcome over the full follow-up period (step 1), but were identified as being prognostic of outcome when follow-up was censored at 5 years (step 2). Despite only being prognostic in the short-term (adjusted HR for the 0-5 year interval 1.73 (95% CI, 1.03-2.90) compared to the 0-25 year interval 1.31 (95% CI, 0.87-1.98)), a non-proportional effect was not identified for CEACAM5 by the FPT algorithm.

**Table 9.16** MFPT Results for the Edinburgh BCS Cohort with Full follow-up – Biomarker only model

| | Edinburgh BCS | | | | | |
|---|---|---|---|---|---|---|
| | All ER-Positive | | | ER-Positive Endocrine Only | | |
| Variables | STEP 1 | STEP 2 | STEP 3 | STEP 1 | STEP 2 | STEP 3 |
| ER | | | | | | |
| PgR | | * | * | * | * | |
| Ki67 | * | * | * | * | * | * |
| HER2 | | | | | | |
| p53 | | | | * | * | |
| NDRG1 | * | * | | * | * | |
| HTF9C | * | * | | | | |
| CEACAM5 | | * | | | * | |
| SLC7A5 | | | | | | |

Note: All variables included in steps 1 and 2 were included as linear. Step 2 includes variables selected in step 1 (using the whole time-period) and variables selected with a short-term effect (patients censored at 5 year follow-up). Step 3 indicates the variables in step 2 which have evidence of a non-proportional effect, with the best fitting transformation to be log(time) for Ki67 and PgR.

Similarities were observed between the variables identified by the MFPT algorithm for ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort with full

follow-up and follow-up censored at 10 years. The same markers were identified but PgR was not identified as having a time-varying effect when full follow-up was considered.

#### *9.6.2.2 Biomarkers and Clinical Variables*

Results of the MFPT algorithm including the clinical variables (age, grade, non-linear tumour size and number of positive nodes, treatment and chemotherapy) for the Edinburgh BCS cohort with full follow-up are given in Table 9.17. Fewer biomarkers remained prognostic of outcome when clinical variables were included in the model. HTF9C was the only marker that was prognostic over the full follow-up (step 1) with the inclusion of clinical variables. PgR, Ki67 and CEACAM5 was then included in the model in step 2, as well as SLC7A5 for ER-positive patients treated with endocrine therapy only. PgR and Ki67 were identified as having time-varying effects.

**Table 9.17** MFPT Results for the Edinburgh BCS Cohort with Full follow-up

| | Edinburgh BCS | | | | | |
|---|---|---|---|---|---|---|
| | All ER-Positive | | | ER-Positive Endocrine Only | | |
| Variables | STEP 1 | STEP 2 | STEP 3 | STEP 1 | STEP 2 | STEP 3 |
| ER | | | | | | |
| PgR | | * | * | | * | |
| Ki67 | | * | * | | * | * |
| HER2 | | | | | | |
| p53 | | | | | | |
| NDRG1 | | | | | | |
| HTF9C | * | * | | * | * | |
| CEACAM5 | | * | | | * | |
| SLC7A5 | | | | | * | |

Note: All variables included in steps 1 and 2 were included as linear. Step 2 includes variables selected in step 1 (using the whole time-period) and variables selected with a short-term effect (patients censored at 5 year follow-up). Step 3 indicates the variables in step 2 which have evidence of a non-proportional effect, with the best fitting transformation to be log-time for Ki67 and PgR. Clinical variables (age, grade, no. positive nodes, tumour size and chemotherapy) were included in the model.

### *9.6.2.3 Cox Regression Models*

The estimated coefficients and hazard ratios from the final model identified by the MFPT approach are given in Table 9.18 with the time-varying effects displayed in Figure 9.9.

**Table 9.18** Multivariate Cox Regression of biomarkers identified by the MFPT approach in the Edinburgh BCS Cohort (full follow-up)

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Coeff (Std. Err.) | HR (95% CI) | Coeff (Std. Err.) | HR (95% CI) |
| **Biomarker Only Model** | | | | |
| ER (50 units) | - | - | - | - |
| PgR (50 units) | -0.22 (0.10) | 0.80 (0.67-0.97) | -0.09 (0.05) | 0.91 (0.83-1.00) |
| Ki67 (10 units) | 0.40 (0.08) | 1.49 (1.28-1.74) | 0.40 (0.09) | 1.49 (1.24-1.79) |
| HER2 | - | - | - | - |
| SLC7A5 | - | - | - | - |
| CEACAM5 | 0.28 (0.21) | 1.32 (0.88-2.00) | 0.34 (0.24) | 1.41 (0.88-2.25) |
| NDRG1 | 0.40 (0.17) | 1.49 (1.07-2.06) | 0.42 (0.19) | 1.52 (1.05-2.21) |
| HTF9C | 0.36 (0.21) | 1.43 (0.95-2.16) | - | - |
| p53 | - | - | 0.39 (0.19) | 1.48 (1.02-2.15) |
| PgR*log(time) | 0.11 (0.05) | 1.11 (1.00-1.24) | - | - |
| Ki67*log(time) | -0.19 (0.05) | 0.83 (0.75-0.91) | -0.19 (0.06) | 0.83 (0.74-0.93) |
| **Biomarker and Clinical Model** | | | | |
| ER (50 units) | - | - | - | - |
| PgR (50 units) | -0.24 (0.10) | 0.78 (0.65-0.95) | -0.09 (0.05) | 0.92 (0.05) |
| Ki67 (10 units) | 0.33 (0.08) | 1.39 (1.18-1.64) | 0.33 (0.10) | 1.39 (0.13) |
| HER2 | - | - | - | - |
| SLC7A5 | - | - | -0.53 (0.31) | 0.59 (0.32-1.08) |
| CEACAM5 | 0.23 (0.21) | 1.26 (0.83-1.90) | 0.38 (0.24) | 1.46 (0.91-2.35) |
| NDRG1 | - | - | - | |
| HTF9C | 0.36 (0.21) | 1.43 (0.95-2.16) | 0.37 (0.24) | 1.45 (0.91-2.33) |
| p53 | - | - | - | - |
| PgR*log(time) | 0.11 (0.05) | 1.12 (1.01-1.25) | - | - |
| Ki67*log(time) | -0.21 (0.05) | 0.81 (0.73-0.90) | -0.20 (0.06) | 0.82 (0.73-0.92) |

NOTE. Multivariate Cox regression of the biomarkers in the final model identified by the MFPT algorithm. Coefficients (standard error) and hazard ratios (95% CIs) for risk of TTDR calculated for a 50 unit increase in ER or PgR, a 10 unit increase in Ki67 and positive compared to negative for all other markers. Values calculated for all ER-positive patients and ER-positive patients who received no chemotherapy in the Edinburgh cohort with full follow-up. Abbreviations: TTDR, time to distant-recurrence; Coeff, coefficient; Std. Err, standard error; CI, confidence interval. * Adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

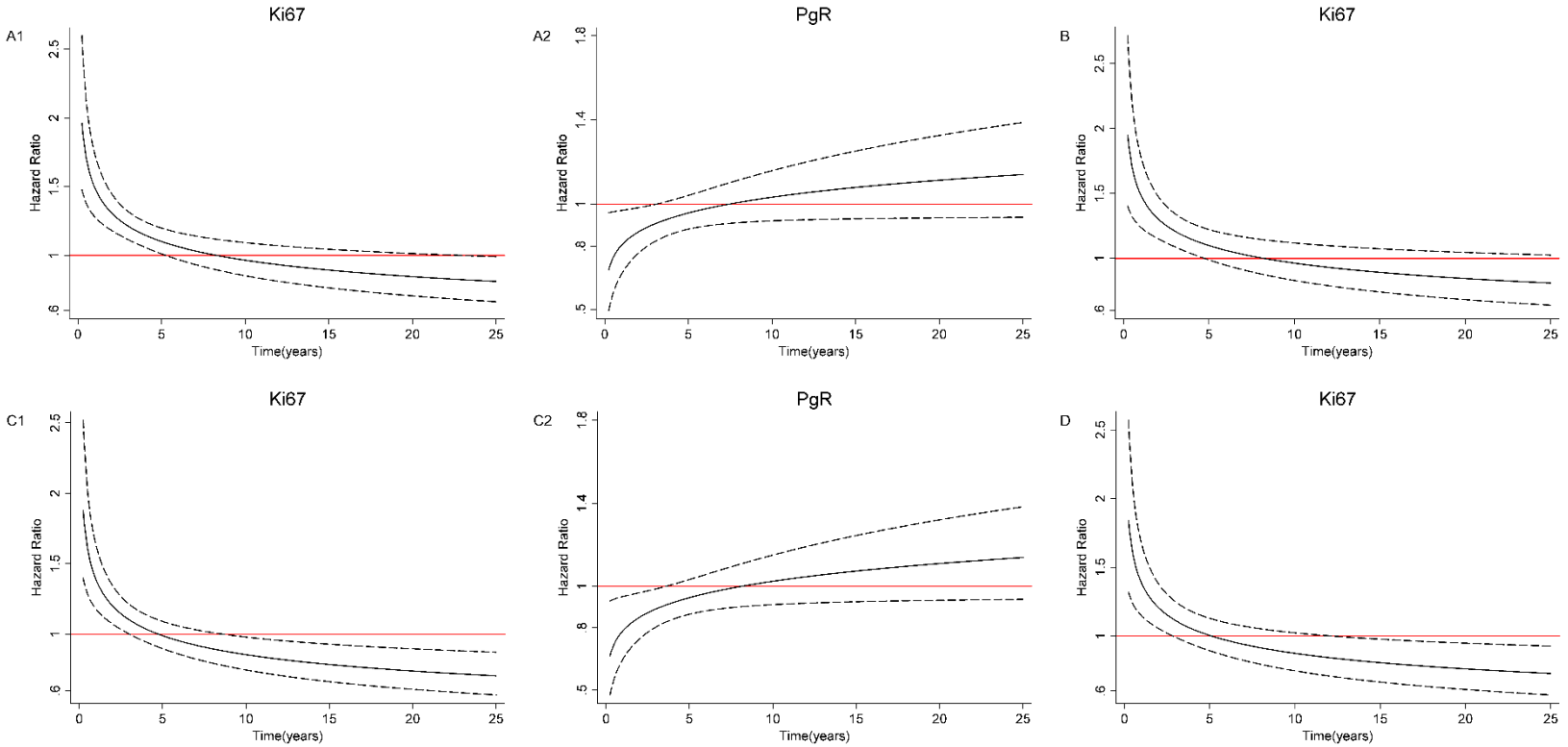



**Figure 9.9** Edinburgh BCS cohort – full follow-up. Time-varying adjusted hazard ratios for a ten-unit increase in Ki67 and fifty-unit increase in PgR. Plots represent (A and C) all ER-positive patients and (B and D) ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort from the biomarker only model (A and B) and the biomarker and clinical model (C and D) (age, grade, tumour size, no. positive nodes and chemo).

### *9.6.2.4 Discrimination*

Again the discrimination was compared between the model identified in step 1 of the MFPT approach; variables prognostic of outcome over the full follow-up, and the final model; variables prognostic of both short and long-term outcome and the inclusion of any time-varying effects. The same 3 models were also compared: the biomarker only model with and without adjustment for clinical variables and the biomarker and clinical model.

**Table 9.19** Edinburgh BCS cohort (full follow-up). Comparison of discrimination of model identified in step 1 of the MFPT approach and the final model.

| | Edinburgh BCS | | | | | |
|---|---|---|---|---|---|---|
| | $R^2$ (%) | | | D-Statistic | | |
| | Step 1 | Final | Increase | Step 1 | Final | Increase |
| **All ER-Positive** | | | | | | |
| Biomarker Only Model | 7.3 | 14.8 | 7.51 | 0.57 | 0.85 | 0.28 |
| Adjusted Biomarker Model | 23.4 | 29.1 | 5.64 | 1.13 | 1.31 | 0.18 |
| Biomarker and Clinical Model | 23.5 | 29.0 | 5.50 | 1.13 | 1.31 | 0.17 |
| **ER-Positive Endocrine Only** | | | | | | |
| Biomarker Only Model | 9.3 | 14.8 | 5.45 | 0.66 | 0.85 | 0.20 |
| Adjusted Biomarker Model | 25.8 | 29.5 | 3.73 | 1.21 | 1.33 | 0.12 |
| Biomarker and Clinical Model | 25.1 | 30.1 | 5.00 | 1.18 | 1.34 | 0.16 |

Note. Values represent $R^2$ and D-statistic for the models identified by step 1 and the final model of the MFPT approach, with a difference in D of at least 0.1 indicating improved prognostic separation. Results are given for all ER-positive patients and ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort with full follow-up. The models represent: biomarker only model, model selected by MFPT including biomarkers only; Adjusted biomarker model, model selected by MFPT including biomarkers only then adjusting for clinical variables (age, grade, tumour size, nodes positive, chemotherapy); Biomarker and clinical model, model selected by MFPT including biomarkers and clinical variables.

There was improved prognostic separation between step 1 of the MFPT model building approach and the final model for all 3 models and both patient subgroups with increases in $R^2$ between 3.7 and 7.5% and increases in D-statistic between 0.12 and 0.28. The extra biomarkers

in the adjusted biomarker model did not improve the prognostic ability compared to the biomarker and clinical model with similar $R^2$ and D-statistics.

## 9.7 RP Flexible Parametric Models

The RP models also allow fitting on different scales, so far we have only considered the hazards scale.

Returning to the mathematical formulas, the basic Weibull model is written as:

$$lnH(t;x) = \ln H_0(t) + x\beta \quad . \qquad 9.1.$$

The survival function S(t) is related to the cumulative hazard function H(t) by H(t)=-ln S(t) allowing us to rewrite (9.1) as:

$$\ln\{-ln\, S(t)\} = \ln\{-\ln S_0(t)\} + x\beta \quad , \qquad 9.2.$$

Which generalises to:

$$g_\theta\{S(t)\} = g_\theta\{S_0(t)\} + x\beta \quad , \qquad 9.3.$$

Where $g_\theta(.)$ is a monotonic increasing function, depending on a parameter θ. Royston and Parmer took $g_\theta(.)$ to be Aranda-Ordaz's (Aranda-Ordaz, 1981) function (Royston and Parmar, 2002b)

$$g_\theta(x) = \ln\left(\frac{x^{-\theta} - 1}{\theta}\right) , \qquad 9.4.$$

where θ > 0. The limit of $g_\theta(x)$ as θ tends to 0 is ln (-ln x), so that with θ = 0 we get the proportional hazards model (9.1).

When θ = 1, the model becomes the proportional odds model:

$$logit\{1 - S(t)\} = logit\{1 - S_0(t)\} + x\beta \quad . \qquad 9.5.$$

The proportional model is structurally similar to a Cox model, but with the feature that the hazard ratio for a covariate converges to 1 as t → ∞. This may be an advantage when we observe a violation of the proportional hazards assumption in the Cox model.

The probit class of models can also be considered with the RP family of survival models. The θ parameter is redundant and $g_\theta(.)$ is defined as minus the probit or inverse normal cumulative distribution function:

$$g_\theta\{S(t;x)\} = -\Phi^{-1}\{S(t;x)\} = -\Phi^{-1}\{S_0(t)\} + x\beta \ . \qquad 9.6.$$

The Aranda-Ordaz definition of RP models provides a large number of potential models. However, with θ not equal to 0 or 1, the interpretation of covariate effects is difficult.

In their book Royston and Lambert consider four key issues in developing a prognostic model (Royston and Lambert, 2011),:

1. Choice of scale (hazards, odds, or probit) and baseline complexity (degrees of freedom for the spline function).
2. Selection of influential variables.
3. Dealing with possible nonlinearity of continuous covariates.
4. Assessing the need to extend the model for time-dependent effects.

### 9.7.1 Choice of scale

The choice of scale for the model was considered using a preliminary model with all 9 biomarkers. The choice of scale and number of knots for the baseline spline function was assessed using the Akaike information criterion (AIC) and Bayes information criterion (BIC) with lower values indicating a better fit of the model.

Results are given in Table 9.20 for the Edinburgh BCS cohort with follow-up censored at 10 years. Except for the AO model, a model with 2 degrees of freedom optimised both AIC and

BIC. The PH model is the poorest fitting model, however the PO and probit only marginally improve the model fit. The scale minimising the AIC and BIC is the AO scale. Similar results were seen were considering the full follow-up for the Edinburgh BCS cohort (Table 9.21).

**Table 9.20** Edinburgh BCS cohort (10-year follow-up). Choice of scale and baseline complexity for multivariable model containing the 9 biomarkers.

| d.f. | PH | | PO | | probit | | AO | |
|---|---|---|---|---|---|---|---|---|
| | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC |
| 1 | 176.8 | 238.1 | 170.1 | 231.3 | 156.3 | 217.6 | 140.0 | 204.3 |
| 2 | 157.8 | 222.1 | 153.9 | 218.2 | 149.7 | 214.0 | 140.3 | 207.7 |
| 3 | 159.2 | 226.5 | 155.2 | 222.6 | 149.9 | 217.3 | 141.6 | 212.0 |
| 4 | 160.6 | 231.0 | 156.6 | 227.1 | 151.8 | 222.2 | 143.0 | 216.5 |
| 5 | 160.4 | 233.9 | 156.5 | 230.0 | 152.7 | 226.2 | 143.6 | 220.1 |

Note. For legibility, 1,000 has been subtracted from all AIC and BIC values. Lowest values of AIC and BIC have been underlined for each scale and the overall lowest AIC and BIC values have been shaded. Abbreviations: d.f, degrees of freedom; PH, proportional hazards; PO, proportional odds; AO, Aranda-Ordaz.

**Table 9.21** Edinburgh BCS cohort (full follow-up). Choice of scale and baseline complexity for multivariable model containing the 9 biomarkers.

| d.f. | PH | | PO | | probit | | AO | |
|---|---|---|---|---|---|---|---|---|
| | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC |
| 1 | 287.8 | 351.0 | 278.4 | 341.7 | 260.0 | 323.3 | 225.3 | 291.7 |
| 2 | 256.3 | 322.7 | 251.7 | 318.2 | 246.0 | 312.5 | 226.4 | 296.0 |
| 3 | 258.2 | 327.9 | 253.6 | 323.3 | 246.9 | 316.5 | 228.0 | 300.8 |
| 4 | 260.2 | 333.0 | 255.6 | 328.4 | 248.9 | 321.7 | 230.0 | 306.0 |
| 5 | 261.7 | 337.6 | 257.0 | 333.0 | 250.9 | 325.8 | 231.4 | 310.5 |

Note. For legibility, 1,000 has been subtracted from all AIC and BIC values. Lowest values of AIC and BIC have been underlined for each scale and the overall lowest AIC and BIC values have been shaded. Abbreviations: d.f, degrees of freedom; PH, proportional hazards; PO, proportional odds; AO, Aranda-Ordaz.

The results for the TEAM cohort are given in Table 9.22. More variation is seen in the degrees of freedom that optimises both AIC and BIC. Again, the PH model has the highest AIC and BIC values, with a probit scale optimising the BIC and AO scale optimising the AIC.

Similar results were observed when clinical variables (age, grade, tumour size, number of positive nodes, treatment and chemotherapy) were included in the model (data not shown).

**Table 9.22** TEAM cohort. Choice of scale and baseline complexity for multivariable model containing the 9 biomarkers.

| d.f. | PH | | PO | | probit | | AO | |
|---|---|---|---|---|---|---|---|---|
| | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC |
| 1 | 801.9 | 887.4 | 787.5 | 873.0 | 766.2 | 851.6 | 763.8 | 853.5 |
| 2 | 779.2 | 868.9 | 771.6 | 861.3 | 766.9 | 856.6 | 763.5 | 857.5 |
| 3 | 781.6 | 875.6 | 774.0 | 868.0 | 767.7 | 861.7 | 765.6 | 863.9 |
| 4 | 778.1 | 876.4 | 770.4 | 868.6 | 762.3 | 860.6 | 761.9 | 864.5 |
| 5 | 780.0 | 882.6 | 772.2 | 874.8 | 764.1 | 866.7 | 763.6 | 870.5 |

Note. For legibility, 3,000 has been subtracted from all AIC and BIC values. Lowest values of AIC and BIC have been underlined for each scale and the overall lowest AIC and BIC values have been shaded. Abbreviations: d.f, degrees of freedom; PH, proportional hazards; PO, proportional odds; AO, Aranda-Ordaz.

### 9.7.2 Selection of variables and functional forms

The three steps of the MFPT approach was followed where RP flexible parametric models were fitted considering each of the different scales where previously the approach was applied using Cox models. The MFP algorithm applied in the first two steps of the MFPT approach can be used with RP models, but the FPT algorithm in step 3 of the MFPT approach is specifically for the Cox model. For step 3 with RP models, the *stpm2t* command was used which tests for time-varying effects using a forward selection criteria and a specified degrees of freedom to model the interaction between the covariate and the spline terms.

*Edinburgh BCS Cohort (10 year follow-up)*

Results are given in Table 9.23 for the Edinburgh BCS cohort with follow-up censored at 10 years. The MFPT approach fitted using a RP model with the hazards, odds or probit scale identified the same markers as were identified using the Cox models (see Table 9.10). The Cox model however identified time-varying effects for both PgR and Ki67, whereas the RP models with hazards or odds scale only identified a time-varying effect for Ki67. There were

**Table 9.23** Final Biomarker Only Model from the MFPT approach fitting using RP models with different scales for the Edinburgh BCS cohort (10-year follow-up)

| | Edinburgh BCS | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | All ER-Positive | | | | ER-Positive Endocrine Only | | | |
| Variables | Hazards | Odds | probit | AO | Hazards | Odds | probit | AO |
| ER | | | | | | | | |
| PgR | * | * | * | * | * | * | * | * |
| Ki67 | * (TV) | * (TV) | * | * | * (TV) | * (TV) | * | * |
| HER2 | | | | * | | | | |
| p53 | * | * | * | * | * | * | * | * |
| NDRG1 | * | * | * | * | * | * | * | * |
| HTF9C | | | | | | | | |
| CEACAM5 | * | * | * | | * | * | * | * |
| SLC7A5 | | | | | | | | |

Note. Final MFPT model fitted using RP models with different choice of scales (hazards, odds, probit and Aranda-Ordaz (AO)) for all ER-positive and ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort with follow-up censored at 10 years. TV indicates a time-varying effect (inclusion of a covariate-time interaction).

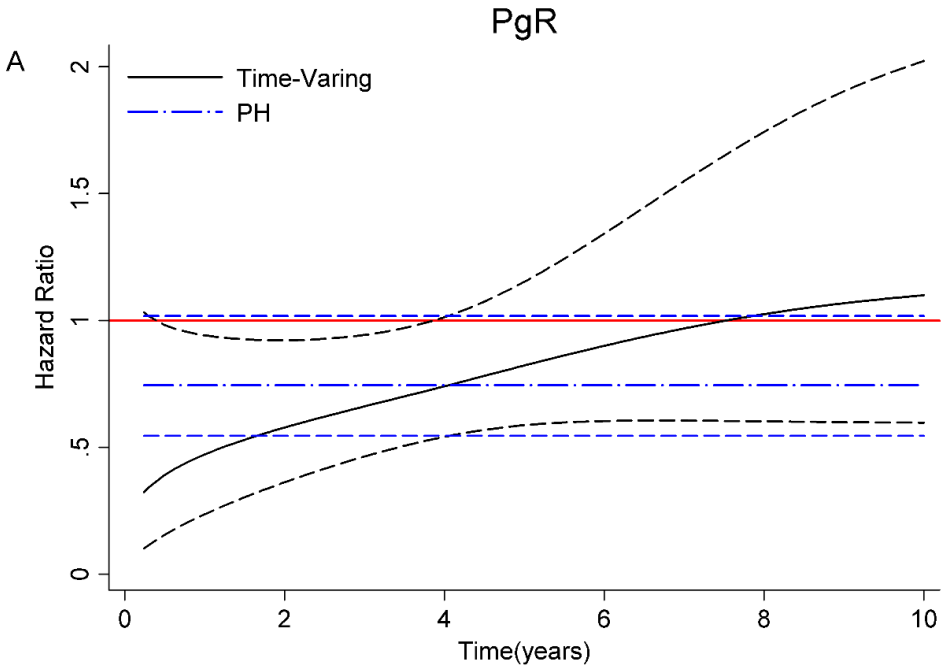



**Figure 9.10** Time-varying (solid black line) and time-constant (dash-dot blue line) adjusted hazard ratio and 95% confidence intervals for PgR. Red line at hazard ratio of 1 indicates a null effect.

issues with convergence when it was attempted to fit more than one time-varying effect. If we force the model to fit (no convergence achieved), including a time-varying effect of PgR did not improve the model fit when compared to time-constant effect (AIC, 141.9 versus 141.5) displayed in Figure 9.10. No time-varying effects were identified with the probit RP model, and the AO model identified different markers when considering all ER-positive patients with HER2 identified as prognostic and CEACAM5 not identified as prognostic of outcome.

*TEAM Cohort*

Results are given in Table 9.24 for the TEAM cohort with all scales selecting the same markers as were chosen with the Cox model (see Table 9.9). For all ER-positive patients, Ki67 and p53 were both identified as having time-varying effects. A disadvantage when modelling on the log cumulative-hazard scale compared with the log-hazard scale in the Cox model is that when there are two variables with time-varying effects, the hazard ratio of the first variable depends on the level of the second variable, displayed in Figure 9.11.There is close agreement between the time-varying effect of Ki67 for those p53 negative and p53 positive but they are not identical as they would be if modelling on the log-hazard scale (as in a Cox model). However, the same conclusions would be drawn.

**Table 9.24** Final Biomarker Only Model from the MFPT approach fitted using RP models with different scales for the TEAM cohort.

| | TEAM | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | All ER-Positive | | | | ER-Positive Endocrine Only | | | |
| Variables | Hazards | Odds | probit | AO | Hazards | Odds | probit | AO |
| ER | * | * | * | * | * | * | * | * |
| PgR | * | * | * | * | * | * | * | * |
| Ki67 | * (TV) | * (TV) | * | * | * (TV) | * | * | * |
| HER2 | * | * | * | * | | | | |
| p53 | * (TV) | * (TV) | * | * | | | | |
| NDRG1 | * | * | * | * | | | | |
| HTF9C | | | | | | | | |
| CEACAM5 | * | * | * | * | | | | |
| SLC7A5 | * | * | * | * | * | * | * | * |

Note. Final MFPT model fitted using RP models with different choice of scales (hazards, odds, probit and Aranda-Ordaz (AO)) for all ER-positive and ER-positive patients treated with endocrine therapy only in the TEAM cohort. TV indicates a time-varying effect effect (significant covariate-time interaction).

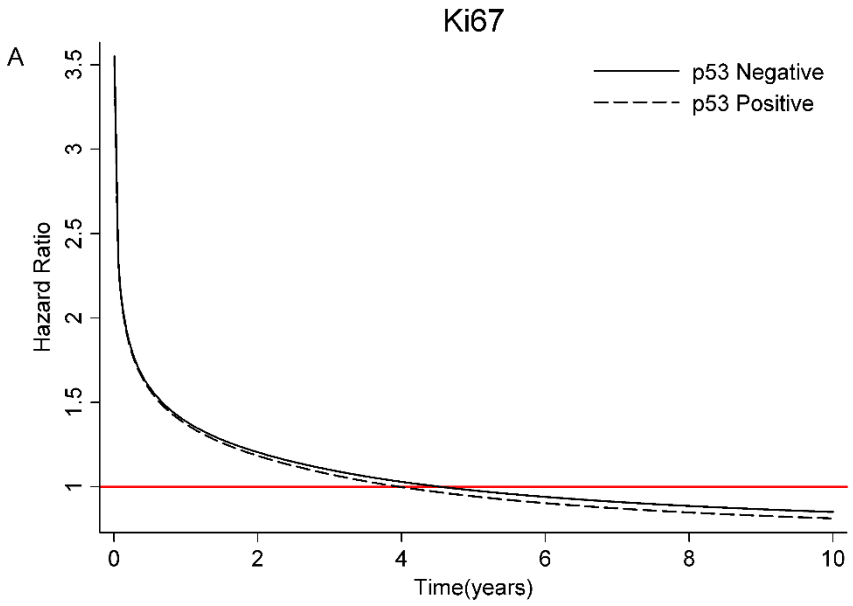



**Figure 9.11** Time-varying adjusted hazard ratio for Ki67 for those p53-negative (solid line) and p53-positive (dashed line) ER-positive patients in the TEAM cohort. Red line at hazard ratio of 1 indicates a null effect.

*Edinburgh BCS Cohort (full follow-up)*

Results are given in Table 9.25 for the Edinburgh BCS cohort considering full follow-up. The hazards and odds scales select the same markers as the Cox model (see Table 9.16) for all ER-positive patients, with the probit model selecting an additional marker, p53, and the AO model selecting fewer markers with HTF9C and CEACAM5 not included in the model. Time-varying effects of Ki67 and PgR were identified with the hazards scale, but with the odds scale, only Ki67 was identified as having a time-varying effect with PgR not reaching the strict 5% level but was of borderline significance (increase in model deviance including PgR as time-varying: 3.5, p-value 0.06).

**Table 9.25** Final Biomarker Only Model from the MFPT approach fitting using RP models with different scales for the Edinburgh BCS cohort (full follow-up)

| | Edinburgh BCS | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | All ER-Positive | | | | ER-Positive Endocrine Only | | | |
| Variables | Hazards | Odds | probit | AO | Hazards | Odds | probit | AO |
| ER | | | | | | | | |
| PgR | * (TV) | * | * | * | * | * | * | * |
| Ki67 | * (TV) | * (TV) | * | * | * (TV) | * (TV) | * (TV) | * |
| HER2 | | | | | | | | |
| p53 | | | * | * | * | * | * | * |
| NDRG1 | * | * | * | * | * | * | * | * |
| HTF9C | * | * | * | | | | | |
| CEACAM5 | * | * | * | | * | * | * | * |
| SLC7A5 | | | | | | | | |

Note. Final MFPT model fitted using RP models with different choice of scales (hazards, odds, probit and Aranda-Ordaz (AO)) for all ER-positive and ER-positive patients treated with endocrine therapy only in the Edinburgh BCS cohort with full follow-up. TV indicates a time-varying effect (inclusion of a covariate-time interaction).

### 9.7.3 Comparison of Estimated Effects

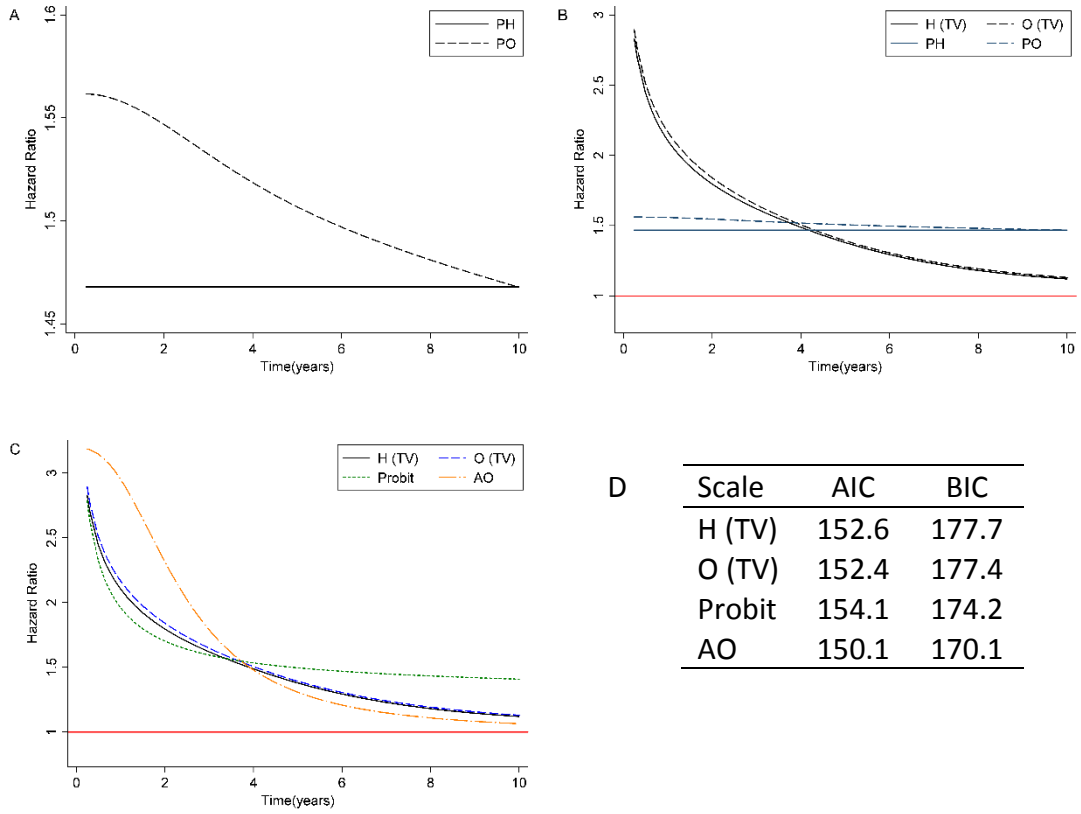
An issue with using the different scales is the interpretation of the coefficients. A comparison of the estimated coefficients is given in Table 9.26. There are similar estimates for the hazards and odds scales, with the exponential of the coefficients from the hazard model giving the well-known hazard ratio estimate but not necessarily an easy interpretation with the need for an understanding of the hazard (the rate at which events happen). The introduction of a time-varying effect in the hazards model also further complicates the interpretation of the coefficients. The exponential of the coefficients from the odds model give the odds ratio estimate, however odds are not always understood correctly and have the added complexity that the ratio converges with time. The coefficients from the probit and Aranda-Ordaz models are substantially different with no easy interpretation.

**Table 9.26** Comparison of estimated coefficients from a RP model with different choice of scales

| | Hazards Coeff (SE) | Odds Coeff (SE) | Probit Coeff (SE) | AO Coeff (SE) |
|---|---|---|---|---|
| ER (50 units) | 0.01 (0.06) | 0.01 (0.06) | -0.0005 (0.03) | -0.09 (0.19) |
| PgR (50 units) | -0.08 (0.05) | -0.10 (0.05) | -0.06 (0.03) | 0.35 (0.14) |
| Ki67 (10 units) | 0.17 (0.05) | 0.18 (0.05) | 0.10 (0.03) | 0.94 (0.42) |
| HER2 | 0.16 (0.20) | 0.17 (0.23) | 0.13 (0.12) | 1.07 (0.90) |
| SLC7A5 | -0.26 (0.28) | -0.30 (0.31) | -0.14 (0.17) | -0.20 (1.04) |
| CEACAM5 | 0.45 (0.22) | 0.46 (0.25) | 0.23 (0.14) | 0.70 (0.69) |
| NDRG1 | 0.47 (0.18) | 0.54 (0.20) | 0.31 (0.11) | 1.61 (0.62) |
| HTF9C | 0.30 (0.23) | 0.34 (0.26) | 0.15 (0.15) | -0.17 (0.83) |
| p53 | 0.30 (0.180 | 0.37 (0.21) | 0.22 (0.11) | 1.07 (0.61) |
| Ki67 - TV | -0.05 (0.02) | -0.05 (0.02) | -0.02 (0.01) | 0.09 (0.09) |

NOTE. Comparison of coefficients from a multivariable RP model including all 9 markers and a time-varying effect for Ki67 with different scales (hazards, odds, probit, Aranda-Ordaz (AO)). Coefficients (standard error) for risk of TTDR calculated for a 50 unit increase in ER or PgR, a 10 unit increase in Ki67 and positive compared to negative for all other markers. Values calculated for all ER-positive patients in the Edinburgh cohort with follow-up censored at 10 years. Abbreviations: TTDR, time to distant-recurrence; Coeff, coefficient; SE, standard error; TV, time-varying effect.

The difference in the choice of scales for the RP model, can best be demonstrated from plotting the exponential of the coefficients against time. This is illustrated in Figure 9.12 for a univariate RP model of Ki67. Plot A illustrates a hazards and odds model with no time-varying effects. The hazards model produces a time-constant effect whereas we can see for the odds model, there is a very subtle convergence towards a null effect (ratio of 1) over time. Plot B illustrates including a time-varying effect (covariate-log (time) interaction) of Ki67 in both the hazards and odds models resulting in a very similar estimate of the time-varying effect. Plot C illustrates the estimated effects from the probit and AO models and the time-varying hazards and odds models with plot D giving the AIC and BIC values to assess the model fit. The AO model has the lowest AIC and BIC values indicating the best fitting model, suggesting an under-estimation of the effect of Ki67 at earlier time-points (less than 4 years) with the time-varying hazards and odds models. The probit model has the highest AIC value, but the time-varying hazards and odds models have the highest BIC due to the penalisation for the extra parameter needed to fit the time-varying effect in the hazards and odds models.



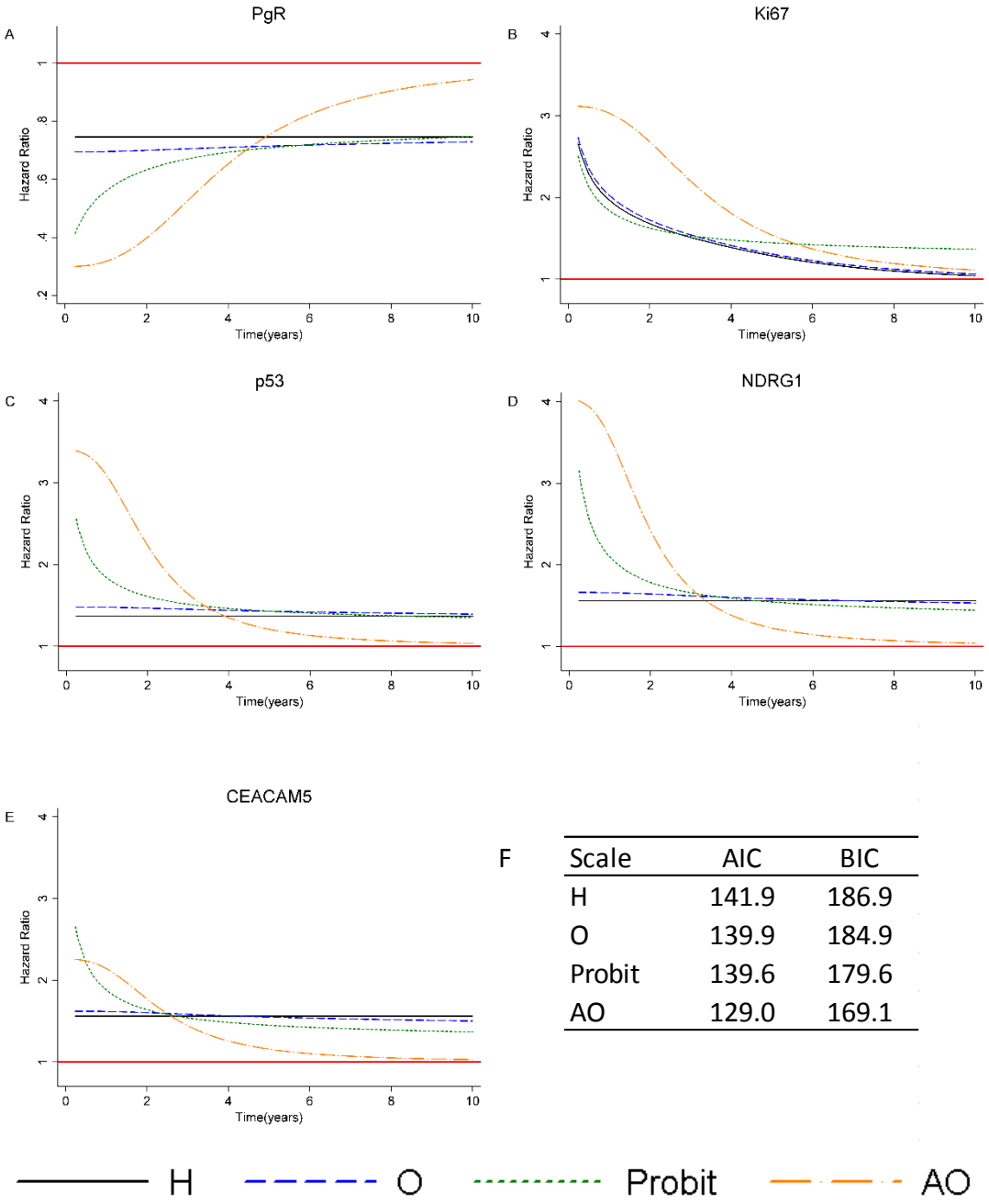

D

| Scale | AIC | BIC |
|---|---|---|
| H (TV) | 152.6 | 177.7 |
| O (TV) | 152.4 | 177.4 |
| Probit | 154.1 | 174.2 |
| AO | 150.1 | 170.1 |

**Figure 9.12** Comparison of the effect of Ki67 from a univariate RP model with different choice of scales. Plots represent (A) proportional hazards (PH) and proportional odds (PO) model, (B) proportional and time-varying (TV) hazards and odds model, (C) time-varying hazards and odds model, probit and Aranda-Ordaz (AO) models and (D) AIC and BIC values for the different models illustrated in C. Note, for legibility 1,000 has been subtracted from AIC and BIC values.

*Edinburgh BCS (10-year follow-up)*

The effects for the multivariable model identified for all ER-positive patients in the Edinburgh BCS cohort with follow-up censored at 10-years (markers identified by the hazards model: PgR, Ki67, p53, NDRG1, CEACAM5) are displayed in Figure 9.13. There are large differences in the way the AO scale models the effect of the covariates over time and suggests a time-varying effect for all of the covariates with the AO model having a substantially lower AIC and BIC values indicating improved model fit. This highlights the flexibility of the AO models, with the ability to model the covariates as time-varying effects without the addition




F

| Scale | AIC | BIC |
|---|---|---|
| H | 141.9 | 186.9 |
| O | 139.9 | 184.9 |
| Probit | 139.6 | 179.6 |
| AO | 129.0 | 169.1 |

**Figure 9.13** Comparison of estimated effects from a multivariable RP model with different scales (hazards (H), odds (O), probit and Aranda-Ordaz (AO)) for all ER-positive patients in the Edinburgh BCS cohort with follow-up censored at 10-years. The variables in the model were (A) PgR, (B) Ki67, (C) p53, (D) NDRG1 and (E) CEACAM5. A covariate-time interaction was included for Ki67 in the hazards and odds model. Model fit was assessed by the AIC and BIC displayed in (F). Note, for legibility 1,000 has been subtracted from AIC and BIC values. Red line at hazard ratio of 1 indicates a null effect.

of extra parameters to include a covariate-time interaction. It is not possible to fit the hazards or odds model with numerous time-varying effects due to issues with model convergence.

In the model building with the AO scale, a different model was identified by the MFPT approach (markers selected: PgR, Ki67, HER2, p53, NDRG1) however no difference in model fit was observed (AIC 129.1 vs. 129.0, BIC 169.1 vs. 169.1) with estimated effects and confidence intervals displayed in Figure 9.14. HER2 was selected in step 2 and was associated with follow-up censored at 4 years, however when modelled over the full follow-up HER2 was not significant in the final model. Due to the AO model having the flexibility to model the covariates as time-varying in step 1, it is not necessary to consider short and long term follow-up separately. Overall, we observe large uncertainty in the estimated effects of the markers at early time points.

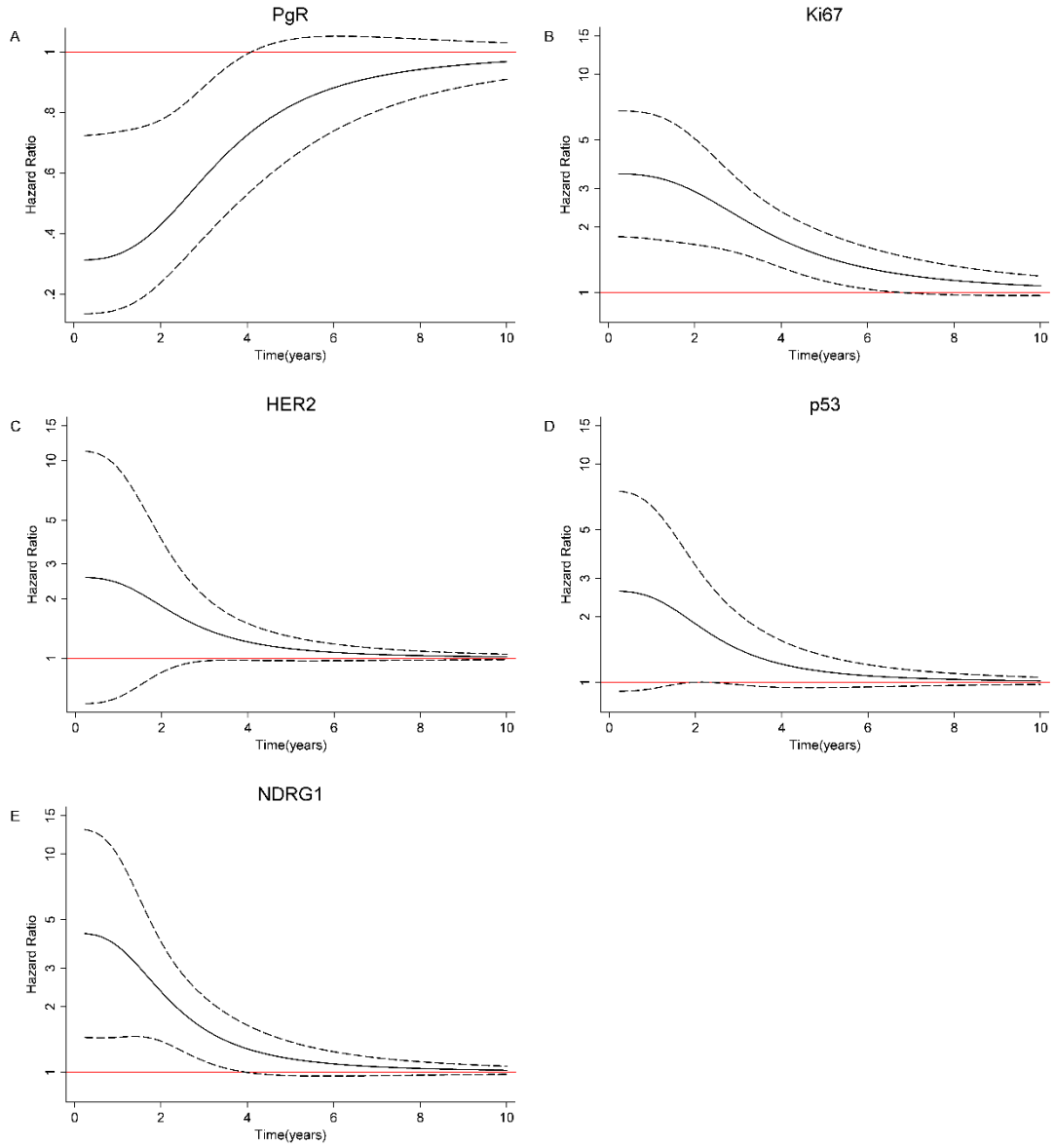



**Figure 9.14** Edinburgh BCS Cohort. Estimated hazard ratios and 95% confidence intervals (dashed lines) from a multivariable RP model with Aranda-Ordaz (AO) scale for all ER-positive patients in the Edinburgh BCS cohort with follow-up censored at 10 years. Red line at hazard ratio of 1 indicates null effect.

*TEAM*

The effects for the multivariable model identified for all ER-positive patients in the TEAM cohort (all markers except HTF9C) are displayed in Figure 9.15. In this cohort, the effects from the hazards or odds model are more similar to that from the AO model with the odds

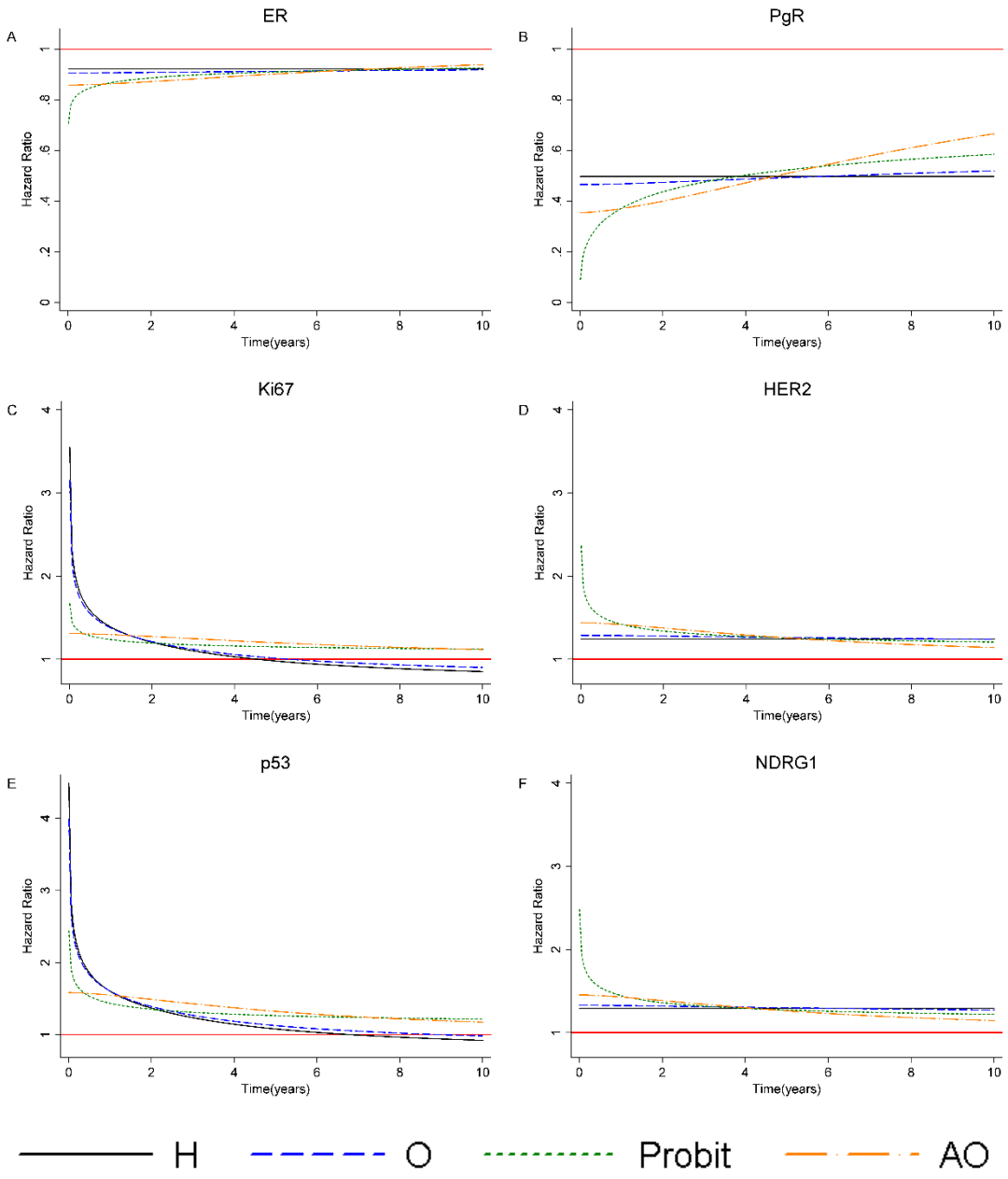
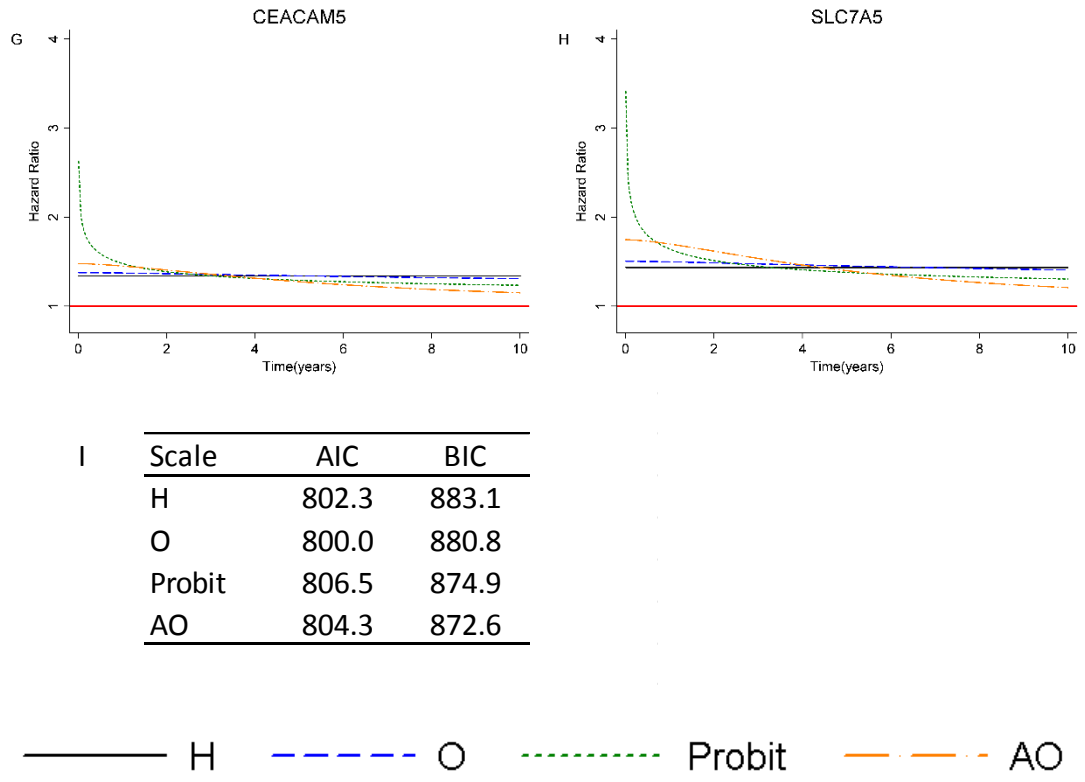



**Figure 9.15** TEAM Cohort. Comparison of estimated effects from a multivariable RP model with different scales (hazards (H), odds (O), probit and Aranda-Ordaz (AO)) for all ER-positive patients in the TEAM cohort. A covariate-time interaction was included for Ki67 and p53 in the hazards and odds model. Model fit was assessed by the AIC and BIC displayed in (I). Note, for legibility 3,000 has been subtracted from AIC and BIC values. Red line at hazard ratio of 1 indicates a null effect.




I

| Scale | AIC | BIC |
|---|---|---|
| H | 802.3 | 883.1 |
| O | 800.0 | 880.8 |
| Probit | 806.5 | 874.9 |
| AO | 804.3 | 872.6 |

**Figure 9.15 (Continued)** Comparison of estimated effects from a multivariable RP model with different scales (hazards (H), odds (O), probit and Aranda-Ordaz (AO)) for all ER-positive patients in the TEAM cohort. A covariate-time interaction was included for Ki67 and p53 in the hazards and odds model. Model fit was assessed by the AIC and BIC displayed in (I). Note, for legibility 3,000 has been subtracted from AIC and BIC values. Red line at hazard ratio of 1 indicates a null effect.

model having the lowest AIC value, but again the extra parameters to model the covariate-time interactions resulting in higher BIC values.

The time-varying effect of p53 is modelled differently from that observed in the Cox model and with the time-varying hazards (or odds) RP models, displayed in Figure 9.16. A univariate model was used to determine the difference in model fit between the AO model and time-varying hazards model which is due to p53 only. The time-varying hazards model had a marginally lower AIC value compared to the AO model (4046.9 versus 4048.1). A comparison to AIC and BIC values from the Cox model cannot be made as the RP models explicitly model the baseline hazard resulting in much lower AIC and BIC values (approx. 9000 vs. 4000 AIC

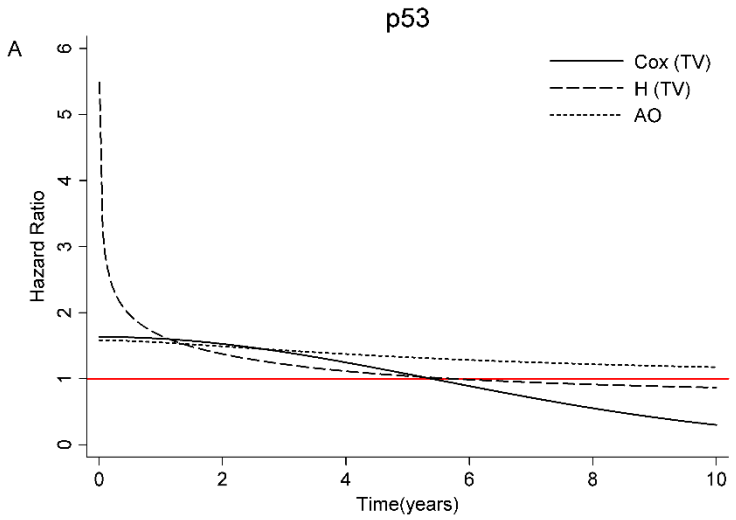



**Figure 9.16** Time-varying adjusted hazard ratio for p53 from a Cox model with a time-varying (TV) effect, a RP model with hazards (H) scale and Aranda-Ordaz (AO) scale for all ER-positive patients in the TEAM cohort.

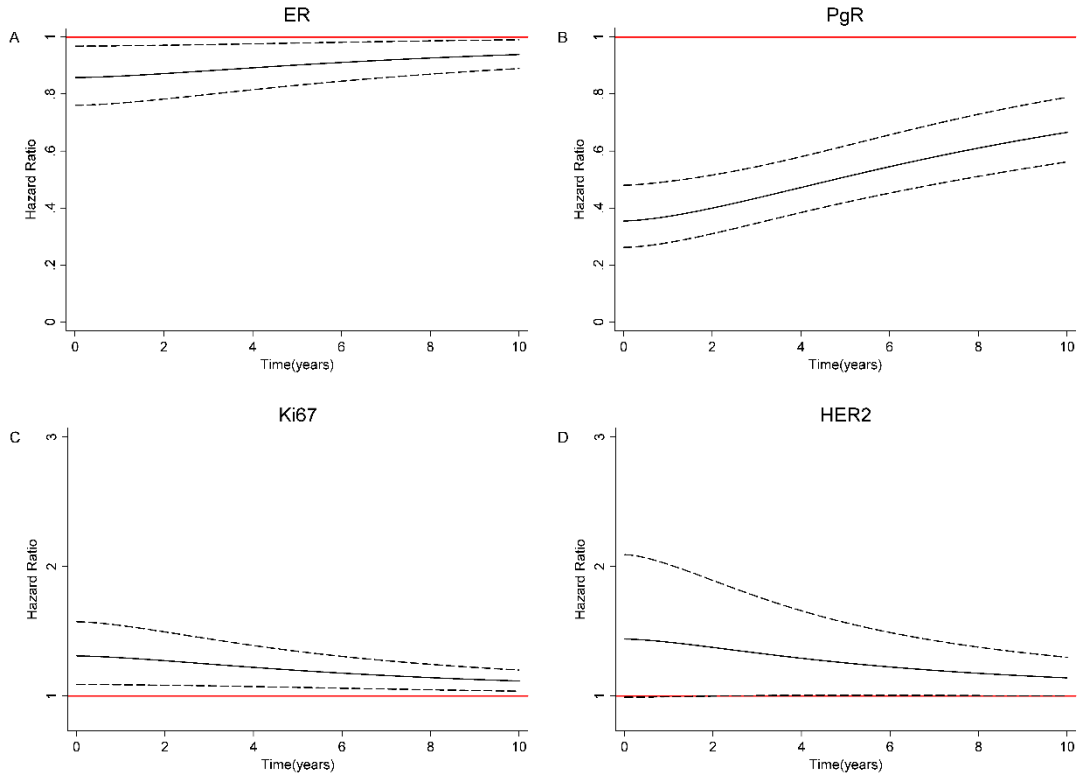



**Figure 9.17** Estimated hazard ratios and 95% confidence intervals (dashed lines) from a multivariable RP model with Aranda-Ordaz (AO) scale for all ER-positive patients in the TEAM cohort.

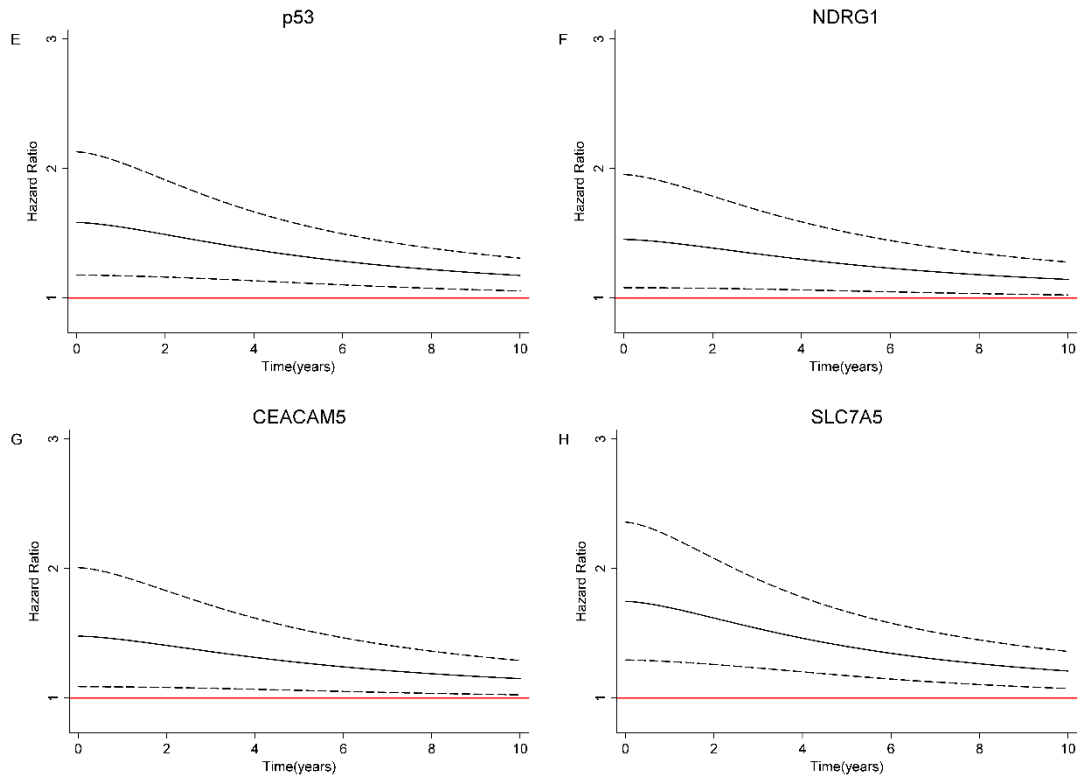
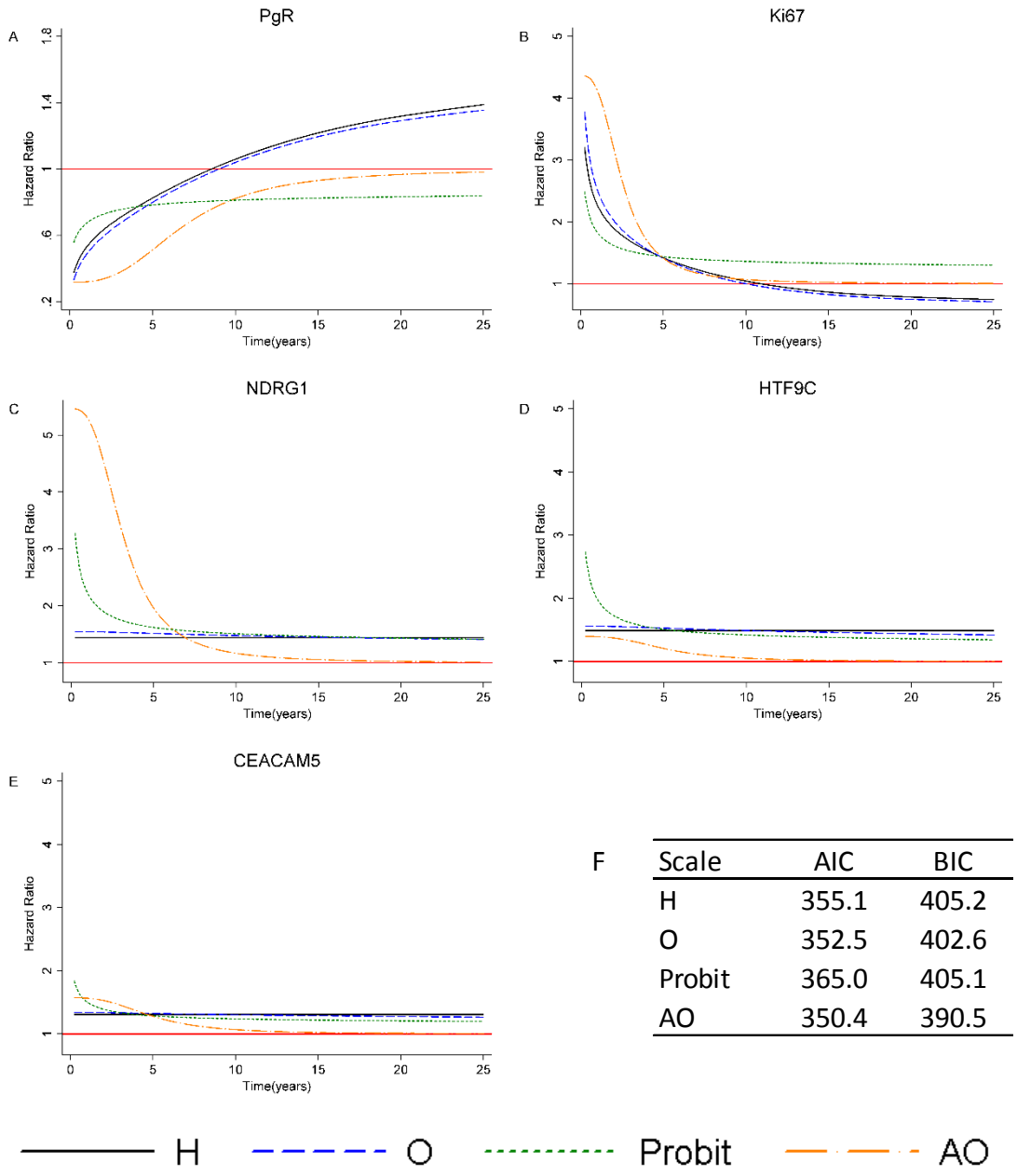



**Figure 9.17 Continued** Estimated hazard ratios and 95% confidence intervals (dashed lines) from a multivariable RP model with Aranda-Ordaz (AO) scale for all ER-positive patients in the TEAM cohort.

value for Cox and RP model respectively). The estimated effects and confidence intervals from the AO model are displayed in Figure 9.17. In the TEAM cohort, the confidence intervals are much narrower than we observed in the Edinburgh BCS cohort. All markers are showing a small decreasing effect over time.

*Edinburgh BCS (full follow-up)*

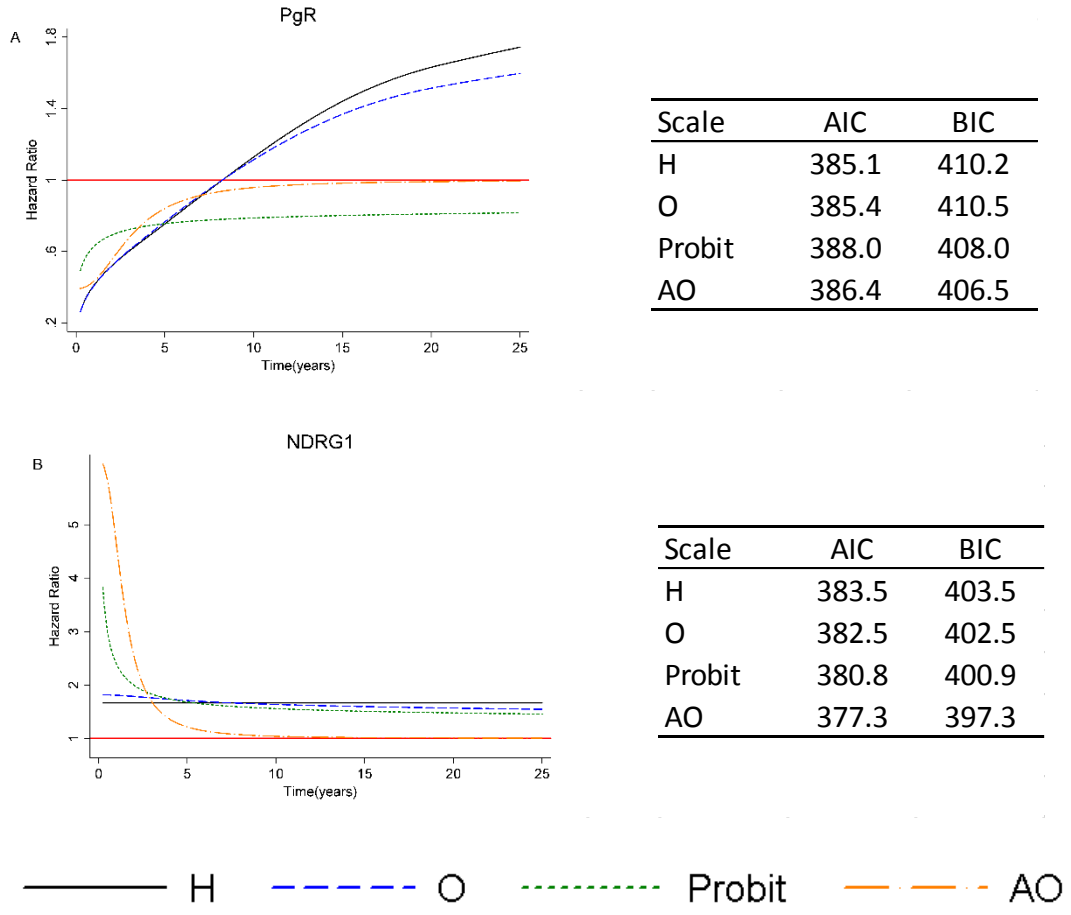
The effects for the multivariable model identified for all ER-positive patients in the Edinburgh BCS cohort (markers identified by the hazards model: PgR, Ki67, NDRG1, HTF9C, CEACAM5) with full follow-up are displayed in Figure 9.18. The probit model was the poorest fitting model with the highest AIC value, again the AO model had the lowest AIC and BIC values but the odds model had a similar AIC (1352.5 versus 1350.4). The main differences between the different scales were observed for PgR and NDRG1. In a univariate analysis




F

| Scale | AIC | BIC |
|---|---|---|
| H | 355.1 | 405.2 |
| O | 352.5 | 402.6 |
| Probit | 365.0 | 405.1 |
| AO | 350.4 | 390.5 |

**Figure 9.18** Comparison of estimated effects from a multivariable RP model with different scales (hazards (H), odds (O), probit and Aranda-Ordaz (AO)) for all ER-positive patients in the Edinburgh BCS cohort with full follow-up. A covariate-time interaction was included for Ki67 and PgR in the hazards and odds model. Model fit was assessed by the AIC and BIC displayed in (F). Note, for legibility 1,000 has been subtracted from AIC and BIC values. Red line at hazard ratio of 1 indicates a null effect.




| Scale | AIC | BIC |
|---|---|---|
| H | 385.1 | 410.2 |
| O | 385.4 | 410.5 |
| Probit | 388.0 | 408.0 |
| AO | 386.4 | 406.5 |

| Scale | AIC | BIC |
|---|---|---|
| H | 383.5 | 403.5 |
| O | 382.5 | 402.5 |
| Probit | 380.8 | 400.9 |
| AO | 377.3 | 397.3 |

**Figure 9.19** Comparison of estimated effects from a univariate RP model of (A) PgR and (B) NDRG1 with different scales (hazards (H), odds (O), probit and Aranda-Ordaz (AO)) for all ER-positive patients in the Edinburgh BCS cohort with full follow-up. A covariate-time interaction was included for PgR in the hazards and odds model. Model fit was assessed by the AIC and BIC. Note, for legibility 1,000 has been subtracted from AIC and BIC values. Red line at hazard ratio of 1 indicates a null effect.

(Figure 9.19), the time-varying hazards model for PgR had the lowest AIC value, however a very similar value was observed for the AO model (1385.1 vs. 1386.4). Whereas for NDRG1, the AO model was an improved model fit compared to the proportional hazards model (1377.3 vs. 1383.5) suggesting a time-varying effect for NDRG1. Different markers were identified by the AO model in the MFPT model building (PgR, Ki67, p53, NDRG1) with an improved model fit (AIC 1343.5 vs. 1350.4) and estimated effects and confidence intervals displayed in Figure 9.20. Wide confidence intervals were again observed and all estimated effects for the markers reached the null effect.

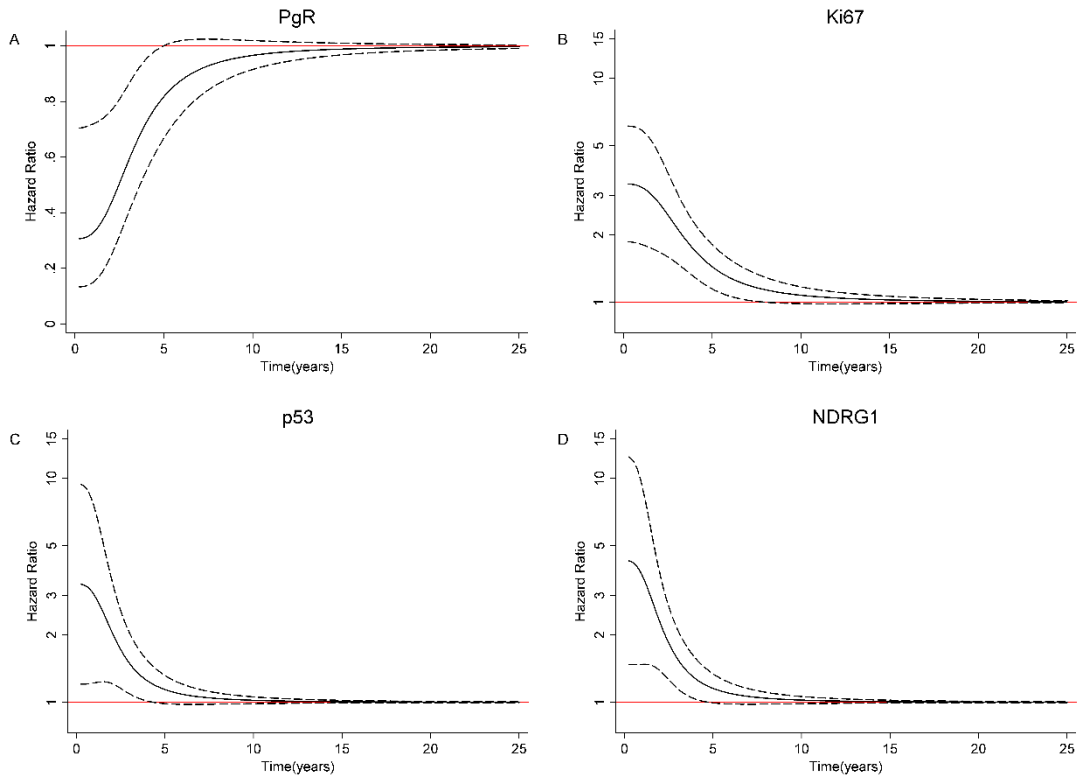



**Figure 9.20** Estimated effects and 95% confidence intervals (dashed lines) for multivariable RP model with Aranda-Ordaz (AO) scale for all ER-positive patients in the Edinburgh BCS cohort with full follow-up. Red line at hazard ratio of 1 indicates null effect.

### 9.7.4 Restricted Mean Survival Time

The Aranda-Ordaz RP models have the advantage of being flexible and can model numerous covariates as time-varying without the need for extra parameters. However, the main drawback is the difficult interpretation of the estimated coefficients. Restricted mean survival time (RMST) was introduced in chapter 8 as an alternative to estimating the hazard ratio when we had non-proportional hazards. It is an estimate of the area under the survival curve up to a pre-specified time-point (t*) and can be thought of as the 't*-life expectancy'. The RMST can also be calculated from RP models with the different choice of scales (hazards, odds, probit and AO) and could be used as an intuitive outcome measure.

The RMST estimates for t* equal to 5 and 10 years from the models identified by the MFPT approach are given in Table 9.27 for the Edinburgh BCS cohort with follow-up restricted to

10-years and Table 9.28 for the TEAM cohort. The larger differences between the modelled effects over time from the hazards and AO models in the Edinburgh BCS cohort is shown with different estimates in RMST with the two choice of scales. However, in the TEAM cohort where we observed more similar estimated effects over time, we observe similar RMST estimates. Each individual marker contributes only a small improvement in time-to-distant-recurrence with many of the estimates containing 0 in the confidence interval. The largest effects were observed for PgR, Ki67 and CEACAM5 in the Edinburgh BCS cohort with being CEACAM5 negative increasing the time-to distant-recurrence over 10 years by 0.6 (95% CI, 0.1-1.0) years compared to being CEACAM5 positive. The largest effects observed in the TEAM cohort were for PgR, NDRG1 and SLC7A5 with being in the 75$^{th}$ centile of PgR increasing the time-to distant-recurrence over 10 years by 1.2 (95% CI, 0.7-1.8) years compared to being in the 25$^{th}$ centile. The estimated effect was smaller when modelled using the AO model (1.0 years 95% CI, 0.4-1.6).

The difference in RMST when all markers were considered low risk (ER and PgR 75$^{th}$ centile, Ki67 25$^{th}$ centile and all other markers negative) and high risk (ER and PgR 25$^{th}$ centile, Ki67 75$^{th}$ centile and all other markers positive) was estimated, referred to as 'overall' in the tables. In the Edinburgh BCS cohort (all ER-positive patients) an increase in time-to distant-recurrence over 10 years of 1.8 (95% CI, 1.3-2.3) years was observed with the hazards model and 1.7 (95% CI, 0.8-2.5) years with the AO model. The improvement in time-to distant-recurrence over 10-years was much larger in the TEAM cohort, with estimated different in RMST of 5.0 (95% CI, 4.6-5.4) years with the hazards model and 3.4 (95% CI, 2.8-4.0) years with the AO model.

**Table 9.27** 5 and 10-year restricted-mean survival estimates for the Edinburgh BCS cohort (10-year follow-up) from a RP model with hazards and Aranda-Ordaz scales

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Hazards | AO | Hazards | AO |
| | Diff in RMST (95% CI) | Diff in RMST (95% CI) | Diff in RMST (95% CI) | Diff in RMST (95% CI) |
| **t* = 5 Years** | | | | |
| ER | - | - | - | |
| PgR | 0.37 (-0.08-0.82) | 0.28 (-0.28-0.84) | 0.41 (-0.08-0.89) | 0.36 (-0.26-0.99) |
| Ki67 | 0.43 (-0.03-0.89) | 0.22 (-0.33-0.77) | 0.43 (-0.06-0.93) | 0.23 (-0.39-0.85) |
| HER2 | - | 0.06 (-0.51-0.62) | - | - |
| p53 | 0.18 (-0.27-0.62) | 0.21 (-0.36-0.78) | 0.31 (-0.16-0.79) | 0.28 (-0.35-0.90) |
| NDRG1 | 0.17 (-0.28-0.61) | 0.21 (-0.36-0.78) | 0.20 (-0.28-0.68) | 0.27 (-0.36-0.90) |
| HTF9C | - | - | - | - |
| CEACAM5 | 0.45 ( 0.02-0.88) | - | 0.60 ( 0.14-1.06) | 0.28 (-0.35-0.90) |
| SLC7A5 | - | - | - | - |
| Overall | 1.46 ( 1.04-1.89) | 0.90 ( 0.34-1.46) | 1.69 ( 1.24-2.14) | 1.32 ( 0.70-1.94) |
| **t* = 10 Years** | | | | |
| ER | - | - | - | - |
| PgR | 0.51 ( 0.00-1.03) | 0.52 (-0.31-1.35) | 0.55 (-0.01-1.11) | 0.64 ( 0.26-1.53) |
| Ki67 | 0.53 ( 0.00-1.07) | 0.41 (-0.41-1.23) | 0.51 (-0.06-1.09) | 0.41 (-0.49-1.30) |
| HER2 | - | 0.11 (-0.73-0.94) | - | - |
| p53 | 0.23 (-0.27-0.73) | 0.38 (-0.45-1.22) | 0.40 (-0.13-0.92) | 0.48 (-0.42-1.38) |
| NDRG1 | 0.22 (-0.28-0.72) | 0.38 (-0.45-1.22) | 0.26 (-0.28-0.79) | 0.47 (-0.43-1.37) |
| HTF9C | - | - | - | - |
| CEACAM5 | 0.56 ( 0.08-1.04) | - | 0.72 ( 0.21-1.22) | 0.48 (-0.42-1.38) |
| SLC7A5 | - | - | - | - |
| Overall | 1.81 ( 1.30-2.31) | 1.66 ( 0.83-2.49) | 2.01 ( 1.48-2.54) | 2.29 ( 1.41-3.17) |

NOTE. Difference in RMST and 95% CI for t* equal to 5 and 10 years for a multivariable Royston-Parmer flexible parametric model with the hazards or Aranda-Ordaz (AO) scale. Difference in RMST calculated as the difference in the IQR of the continuous markers and negative compared to positive for dichotomous markers. Overall represents the difference between all markers are low (PgR 75th centile, Ki67 25th centile, all other markers negative) and all markers high (PgR 25th centile, Ki67 75th centile, all other markers positive). Abbreviations: RMST, restricted mean survival time; CI, confidence interval. Model is adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

**Table 9.28** 5 and 10-year restricted-mean survival estimates for the TEAM cohort from a RP model with hazard and Aranda-Ordaz scales

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Hazards | AO | Hazards | AO |
| | Diff in RMST (95% CI) | Diff in RMST (95% CI) | Diff in RMST (95% CI) | Diff in RMST (95% CI) |
| **t* = 5 Years** | | | | |
| ER | 0.05 (-0.28-0.38) | 0.08 (-0.29-0.45) | 0.13 (-0.29-0.54) | 0.17 (-0.28-0.61) |
| PgR | 0.34 ( 0.01-0.68) | 0.45 ( 0.07-0.82) | 0.44 ( 0.03-0.85) | 0.42 (-0.03-0.86) |
| Ki67 | 0.08 (-0.25-0.41) | 0.05 (-0.31-0.42) | 0.12 (-0.29-0.53) | 0.07 (-0.38-0.51) |
| HER2 | 0.06 (-0.27-0.40) | 0.13 (-0.24-0.51) | - | - |
| p53 | 0.26 (-0.09-0.60) | 0.22 (-0.16-0.60) | - | - |
| NDRG1 | 0.21 (-0.13-0.55) | 0.19 (-0.18-0.57) | - | - |
| HTF9C | - | - | - | - |
| CEACAM5 | 0.16 (-0.17-0.51) | 0.21 (-0.17-0.58) | - | - |
| SLC7A5 | 0.21 (-0.13-0.55) | 0.29 (-0.09-0.67) | 0.38 (-0.03-0.80) | 0.32 (-0.13-0.77) |
| Overall | 2.08 ( 1.73-2.43) | 1.69 ( 1.31-2.08) | 1.12 ( 0.71-1.53) | 0.96 ( 0.52-1.41) |
| **t* = 10 Years** | | | | |
| ER | 0.17 (-0.37-0.70) | 0.19 (-0.38-0.76) | 0.36 (-0.25-0.97) | 0.32 (-0.33-0.98) |
| PgR | 1.21 ( 0.68-1.75) | 1.01 ( 0.44-1.58) | 1.22 ( 0.61-1.83) | 0.81 ( 0.15-1.46) |
| Ki67 | 0.06 (-0.48-0.60) | 0.12 (-0.45-0.69) | 0.13 (-0.48-0.74) | 0.13 (-0.52-0.79) |
| HER2 | 0.23 (-0.31-0.77) | 0.29 (-0.28-0.86) | - | - |
| p53 | 0.62 ( 0.07-1.16) | 0.48 (-0.09-1.06) | - | - |
| NDRG1 | 0.71 ( 0.17-1.26) | 0.43 (-0.15-1.00) | - | - |
| HTF9C | - | - | - | - |
| CEACAM5 | 0.57 ( 0.03-1.12) | 0.46 (-0.12-1.03) | - | - |
| SLC7A5 | 0.71 ( 0.17-1.25) | 0.63 ( 0.05-1.20) | 1.01 (0.40-1.61) | 0.61 (-0.05-1.27) |
| Overall | 5.04 ( 4.55-5.41) | 3.39 ( 2.82-3.96) | 2.71 (2.12-3.29) | 1.84 (1.19-2.49) |

NOTE. Difference in RMST and 95% CI for t* equal to 5 and 10 years for a multivariable Royston-Parmer flexible parametric model with the hazards or Aranda-Ordaz (AO) scale. Difference in RMST calculated as the difference in the IQR of the continuous markers and negative compared to positive for dichotomous markers. Overall represents the difference between all markers are low (ER and PgR 75th centile, Ki67 25th centile, all other markers negative) and all markers high (ER and PgR 25th centile, ki67 75th centile, all other markers positive) Abbreviations: RMST, restricted mean survival time; CI, confidence interval. Model is adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

**Table 9.29** 5, 10 and 15-year restricted-mean survival estimates for the Edinburgh BCS cohort (full follow-up) from a RP model with hazards and Aranda-Ordaz scales

| | All ER-Positive | | ER-Positive Endocrine Only | |
|---|---|---|---|---|
| | Hazards | AO | Hazards | AO |
| | Diff in RMST (95% CI) | Diff in RMST (95% CI) | Diff in RMST (95% CI) | Diff in RMST (95% CI) |
| **t* = 5 Years** | | | | |
| ER | - | | - | - |
| PgR | 0.62 ( 0.15-1.08) | 0.24 (-0.32-0.80) | 0.29 (-0.21-0.78) | 0.31 (-0.32-0.94) |
| Ki67 | 0.37 (-0.10-0.84) | 0.20 (-0.35-0.75) | 0.49 (-0.02-0.99) | 0.21 (-0.41-0.83) |
| HER2 | - | - | - | - |
| p53 | - | 0.17 (-0.39-0.74) | 0.22 (-0.27-0.71) | 0.21 (-0.42-0.84) |
| NDRG1 | 0.14 (-0.32-0.60) | 0.21 (-0.35-0.78) | 0.15 (-0.34-0.64) | 0.24 (-0.39-0.87) |
| HTF9C | 0.43 (-0.02-0.88) | - | - | - |
| CEACAM5 | 0.31 (-0.14-0.76) | - | 0.40 (-0.08-0.89) | 0.19 (-0.44-0.83) |
| SLC7A5 | - | - | - | - |
| Overall | 1.70 ( 1.25-2.14) | 0.82 ( 0.26-1.38) | 1.46 ( 0.98-1.92) | 1.10 ( 0.48-1.73) |
| **t* = 10 Years** | | | | |
| ER | - | - | - | - |
| PgR | 0.77 ( 0.24-1.31) | 0.45 (-0.38-1.28) | 0.40 (-0.17-0.97) | 0.55 (-0.35-1.46) |
| Ki67 | 0.48 (-0.08-1.04) | 0.38 (-0.44-1.20) | 0.61 ( 0.01-1.20) | 0.38 (-0.52-1.28) |
| HER2 | - | - | - | - |
| p53 | - | 0.32 (-0.51-1.16) | 0.30 (-0.25-0.85) | 0.37 (-0.54-1.28) |
| NDRG1 | 0.19 (-0.33-0.72) | 0.40 (-0.44-1.23) | 0.20 (-0.36-0.76) | 0.43 (-0.48-1.34) |
| HTF9C | 0.56 ( 0.06-1.06) | - | - | - |
| CEACAM5 | 0.41 (-0.10-0.92) | - | 0.52 (-0.02-1.05) | 0.35 (-0.56-1.25) |
| SLC7A5 | - | - | - | - |
| Overall | 2.09 ( 1.56-2.62) | 1.53 ( 0.70-2.36) | 1.81 ( 1.25-2.36) | 1.97 ( 1.07-2.86) |
| **t* = 15 Years** | | | | |
| ER | - | - | - | - |
| PgR | 0.80 ( 0.24-1.36) | 0.64 (-0.39-1.68) | 0.43 (-0.17-1.02) | 0.77 (-0.35-1.88) |
| Ki67 | 0.50 (-0.09-1.08) | 0.55 (-0.48-1.57) | 0.62 ( 0.00-1.23) | 0.53 (-0.59-1.64) |
| HER2 | - | - | - | - |
| p53 | - | 0.46 (-0.57-1.50) | 0.31 (-0.25-0.88) | 0.51 (-0.60-1.63) |
| NDRG1 | 0.21 (-0.34-0.75) | 0.57 (-0.47-1.60) | 0.22 (-0.36-0.79) | 0.60 (-0.52-1.71) |
| HTF9C | 0.59 ( 0.07-1.11) | - | - | - |
| CEACAM5 | 0.44 (-0.09-0.97) | - | 0.54 (-0.01-1.09) | 0.48 (-0.63-1.60) |
| SLC7A5 | - | - | - | - |
| Overall | 2.16 ( 1.61-2.72) | 2.17 ( 1.15-3.20) | 1.86 ( 1.28-2.44) | 2.73 ( 1.63-3.83) |

NOTE. Difference in RMST and 95% CI for t* equal to 5, 10 and 15 years for a multivariable RP model with the hazards or Aranda-Ordaz (AO) scale. Difference in RMST calculated as the difference in the IQR of the continuous markers and negative compared to positive for dichotomous markers. Overall represents the difference between all markers are low (PgR 75th centile, Ki67 25th centile, all other markers negative) and all markers high (PgR 25th centile, ki67 75th centile, all other markers positive). Abbreviations: RMST, restricted mean survival time; CI, confidence interval. Model is adjusted for age, grade, tumour size, nodes positive, treatment and chemotherapy.

The RMST estimates for t* equal to 5, 10 and 15 years from the models identified by the MFPT approach are given in Table 9.29 for the Edinburgh BCS cohort with full follow-up. We observe large differences in estimates between the two choice of scales, hazards and AO. Larger effects were observed for PgR and HTF9C, with being in the 75$^{th}$ centile of PgR increasing the time-to distant recurrence over 15 years by 0.8 (95% CI 0.2-1.5) years compared to the 25$^{th}$ centile. Comparing the overall low risk to high risk, there was increase in time-to distant-recurrence over 15 years by 2.2 (95% CI, 1.6-2.7) years.

## 9.8 Discussion

This chapter explored the effects of the individual biomarkers from the IHC4 and Mammostrat residual risk panels. The MFPT approach was used to build a multivariable model with Cox models and Royston-Parmer (RP) flexible parametric models where the different choice of scales (hazards, odds, probit and AO) were explored.

The first two steps of the MFPT approach involve using the MFP algorithm to determine which variables are associated with the outcome in the short and long-term and simultaneously determine the best FP to model the functional form of the covariates. An issue arose where despite there being no statistical significant difference in the best fitting degree 2 FP compared with a linear effect on the log-hazard, considering the FP2 model resulted in variables not being identified as prognostic of outcome. It is important to consider whether the FP2 model results in over-fitting of the covariate and in the case that was observed here, a clinically implausible effect of the predictor.

The importance of considering variables associated with both long and short term when model building was illustrated. A large improvement in model discrimination was observed in the Edinburgh BCS cohort comparing variables associated over the full follow-up only and the final model including variables associated with both short and long-term outcome and identifying any time-varying effects.

The MFP and FPT algorithms used in the MFPT approach were developed to keep the overall type I error at a pre-specified level when considering which variables were prognostic of outcome and which had time-varying effects and the best fitting FPs to model these effects. However, the approach is essentially a stepwise selection method which could result in an arbitrary and over-fitted model. It is important to take into account any clinical insight when model building, and the MFPT approach allows you to have control over the process by forcing covariates into the model, selecting the significance level for the inclusion of prognostic and time-varying effects, specifying the degrees of freedom and the significance level for testing between FP models.

The MFPT approach was developed on the Cox model. A similar method was applied using RP models which have the advantage of being parametric and explicitly model the baseline hazard. The MFP algorithm used in the first two steps of the MFPT approach can be applied using RP flexible parametric models. Step 3 involves the FPT algorithm which is specific for determining time-varying effects in the Cox model, but a similar forward selection algorithm can be used to determine time-varying effects in RP models with the advantage of not having to split the time scale as in the Cox model. Fitting a RP model with the hazards scale resulted in the same models being identified as with the Cox model. A disadvantage of the RP models was illustrated where modelling on the log-cumulative hazards scale results in the time-varying effect of one covariate depending on the value of any other time-varying covariates. The difference however was small, but should be kept in mind when modelling time-varying effects on the log-cumulative hazard.

The RP model has the flexibility of allowing you to model on different scales (hazards, odds, probit and Aranda-Ordaz (AO)). The odds scale has the potential advantage that the effect of the covariates converge to 1 as time tends to infinity. The flexibility of the AO model was illustrated with the ability to model numerous covariates as time-varying effects without the extra parameters to model an interaction with time and it is not necessary to determine the best

fitting FP or the required flexibility (degrees of freedom) of the spline terms. The main drawback to the using the different scales was the difficulty interpreting the estimated coefficients. The novel measure of restricted mean survival time was illustrated as an alternative outcome measure which can be estimated for any of the scales and also with the inclusion of time-varying effects.

In the Edinburgh cohort PgR, Ki67, p53, NDRG1 and CEACAM5 were important predictors of TTDR, with PgR, Ki67 and CEACAM5 remaining prognostic of outcome when adjusting for clinical variables (age, grade, tumour size, number of positive nodes and treatment). Ki67 was identified as having a strong non-proportional effect. The overall model estimated an increase in time-to distant-recurrence over 10 years of 1.8 (95% CI, 1.3-2.3) years comparing low risk to high risk patients.

All markers except HTF9C were prognostic of outcome in the TEAM cohort and remained important predictors of outcome after adjustment for clinical variables with the exception of HER2. Ki67 and p53 were identified as having time-varying effects. The overall model estimated a large increase in time-to distant-recurrence over 10 years of 5.0 (95% CI, 4.6-5.4) years comparing low risk to high risk patients.

Differences were observed between the two cohorts in the selection of prognostic markers. This highlights the differences in the two patient populations, with the Edinburgh BCS cohort being a lower risk population than the TEAM cohort. The wide confidence intervals when modelling using RP flexible parametric models with the AO scale in the Edinburgh BCS cohort indicates large uncertainty in the estimated effects of the individual markers, and suggests this cohort may not be adequately powered to perform such a flexible analysis.

Overall, we have illustrated complex models with subtle differences depending on the cohort and patient subtype. In both cohorts, a combination of markers from the IHC4 and

Mammostrat residual risk models were prognostic of outcome, with a constantly strong time-varying effect of Ki67.

# Chapter 10: Discussion and Recommendations

## 10.1 Summary of Results

### 10.1.1 Statistical Methods for the Analysis of Non-Proportional Hazards

The Cox proportional hazards model is the most commonly used approach for analysing survival time data in medical research despite its well-known limitations. The baseline hazard function is completely unspecified which can be seen as an advantage (one avoids potential problems from specifying the wrong shape), but it has a disadvantage when the shape of the baseline hazard function may provide important information related to the time-course of an illness. Explicitly modelling the baseline hazard also allows the easy calculation of absolute risk estimates which may be of more clinical benefit than only having relative risk estimates (e.g. hazard ratios).

Another limitation of the Cox model is the proportional hazards assumption, which assumes that the hazard ratio comparing one level of a covariate to another is constant over time. This assumption is often violated when modelling prognostic factors in cancer studies. When the assumptions of the Cox PH model are not met, it is possible that subsequent analyses and risk estimates will be biased.

In this thesis, a literature review was performed to identify existing approaches for the analysis of non-proportional effects with respect to survival data with an interest in the application of these approaches in practice.

The simplest approach to deal with non-proportional hazards is splitting the time scale, known as piecewise constant effects, but the drawbacks of this approach are well established. The

categorising of the time-scale results in a reduction of power and the number of intervals chosen can have an impact on the conclusions made.

A large number of alternatives to piecewise constant effects were highlighted by the review; from fully parametric to non-parametric models, multiplicative effects of the covariates to additive effects, different methodological assumptions such as cure and frailties, as well as an expanding body of literature on Bayesian approaches. However, there appeared to be few examples of application of the different approaches in practice.

Two approaches were highlighted in the review: the multivariable fractional polynomial time (MFPT) approach by Sauerbrei and colleagues and a flexible parametric approach proposed by Royston and Parmer (Sauerbrei et al., 2007, Royston and Parmar, 2002b). Both of these approaches allow incorporating non-linear effects of continuous covariates on the log-hazard as well as non-proportional effects. They are also easily implemented in standard software which is important in encouraging the future use of these approaches.

*The MFPT Approach*

The MFPT approach is a multivariable strategy to select variables which have an effect on the outcome, determine the functional form of continuous variables and determine whether the PH assumption is sensible or if time-varying effects are present. The advantages of this approach include: the model reverts to the Cox proportional hazards model if no time-dependent effects are identified, which is the standard method of analysis for survival data and therefore most clinicians and researchers have a basic understanding of the method. It is a closed-test algorithm which keeps the overall error rate at a pre-specified level whilst simultaneously selecting the best FP to model non-linear and non-proportional effects and whether they are prognostic of outcome. The approach also highlighted the need to consider both short and long-term survival when assessing prognostic factors; considering only the full

follow-up could result in missing important predictors of short-term survival and important time-varying effects.

The main disadvantage of the MFPT approach is the need to categorise survival time when evaluating time-varying effects. The Cox model is fitted by evaluating the partial likelihood at each failure time and because the time-varying effects are a function of study time, the observations under risk are split into episodes at specified time points. Ideally the splits would be made at each event time, however with large datasets this is not always possible. Buchholz has shown that the MFPT procedure gives similar results for small categorisation intervals and propose that about 50 to 100 distinct event times can give results with sufficient precision (Buchholz, 2010). However, a relatively long computing time was observed when running the FPT algorithm with only 58 distinct event times. It was also illustrated that it is important to consider the choice of degree of flexibility for the FPs. A FP of degree 2 resulted in a clinically implausible effect of a covariate which was then excluded from the model, but considering a FP of degree 1 or a linear effect, the covariate was a significant predictor of outcome.

Overall, the MFPT approach is useful for identifying non-linear and non-proportional effects in the Cox regression model. However, the approach is essentially a stepwise selection method which could result in an arbitrary and over-fitted model. It is important to take into account any clinical insight when model building, and the MFPT approach allows you to have control over the process by forcing covariates into the model, selecting the significance level for the inclusion of prognostic and time-varying effects, specifying the degrees of freedom and the significance level for testing between FP models.

*RP Flexible Parametric Models*

The Royston-Parmer flexible parametric models involves modelling the baseline hazard and non-proportional effects through the use of regression splines. The MFPT approach was followed fitting the models using the more flexible RP models rather than the Cox model. The

MFP algorithm used in first two steps of the MFPT approach to determine which variables are prognostic of long and short-term survival and the best fitting FP to model the effect on the log-hazard can be also be used with RP flexible parametric models. Therefore the advantage of the closed test algorithm remains whilst simultaneously determining variables that have a prognostic effect on the outcome and the best FP model to be the effect. Step three in the MFPT approach involves the FPT algorithm which is specific for determining time-varying effects in the Cox model. Instead a forward selection algorithm was used to determine time-varying effects in RP models with a specified degrees of freedom. Once the time-varying effects are identified, the degree of flexibility needed to model the time-varying effect is determined by examining the model fit.

With proportional hazards, estimates from a RP model are almost identical to that from a Cox model. The same models were also identified when following the MFPT approach with either Cox models or RP models on the hazards scale. One of the advantages of the RP models is the ease of which time-dependent effects can be fitted, with no need for categorising of the survival time, a simple interaction between the covariates and spline terms is introduced. Due to the model being parametric, this enables easy prediction of absolute and relative risk estimates. The use of survival and hazard rates along with a novel measure, restricted mean survival time (RMST), was illustrated. The RMST has a more intuitive interpretation than the hazard ratio and can be thought of as a 't-year life expectancy'. For example, treatment A increases your life expectancy by 2 years compared with treatment B.

The RP models also allow modelling on different scales (hazards, odds, probit and Aranda-Ordaz (AO)). The odds scale has the advantage that the effect of the covariates converge to 1 as time tends to infinity, which may be seen as a more appropriate assumption for many prognostic factors compared to the more restrictive constant proportionality in the Cox model. The flexibility of the AO model was illustrated with the ability to model numerous covariates as time-varying effects without the extra parameters to model an interaction with time and it

is not necessary to determine the best fitting FP or the required flexibility (degrees of freedom) of the spline terms. The main drawback to the using the different scales was the difficulty interpreting the estimated coefficients. The restricted mean survival time was illustrated as an alternative outcome measure which can be estimated for any of the scales and also with the inclusion of time-varying effects. Although not demonstrated in the thesis, RP flexible parametric models can also be extended to relative survival models.

There are some drawbacks to the flexible parametric approach, as difficulties can occur with convergence when fitting these models to small samples or fitting complex models with more than one time-varying effect. Identifying time-varying effects are not in a closed test algorithm and therefore there may be an issue with multiple testing. A strict level of significance for selection of time-varying effects could help avoid this. A disadvantage of the RP models was illustrated where modelling on the log-cumulative hazards scale results in the time-varying effect of one covariate depending on the value of any other time-varying covariates. The difference however was small, but should be kept in mind when modelling time-varying effects on the log-cumulative hazard.

*Conclusion*

Overall, the RP flexible parametric models would be favoured over Cox models when investigating non-proportional effects due to the main advantage of not having to categorise survival time. It was shown the MFPT approach could be easily adapted to use RP flexible parametric models. This approach would be recommended as it highlights the need for investigating non-linear effects and considering both short- and long-term survival to ensure the identification of important predictors. It is recommended when a non-proportional effect is identified, a plot of the hazard ratio with 95% confidence intervals over continuous-time is shown along with restricted mean survival estimates. Piece-wise constant effects can also be presented as an aid to help clinical interpretation.

The focus of the thesis has been on the evaluation of non-proportional effects of prognostic factors. A recently published paper highlights the importance of this work in other areas. Royston and Parmer propose how to design and analyse a trial when hazards may be non-proportional in their paper titled 'An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect' (Royston and Parmar, 2014). They suggest there is a need to improve the power to detect a potentially more complex treatment effects than assuming proportional hazards. When determining the sample size of a trial they propose to use a 'joint test', which combines the usual log-rank or Cox test with the Grambsch-Therneau test of non-proportional hazards. When predefining the statistical analysis, they recommend the use of the flexible parametric model with 3 degrees of freedom for the baseline hazard function and 1 degrees of freedom for the time-dependent treatment effect.

It is agreed with Binquet and colleagues that there is the need for continued methodological research in this area, for more widespread use of flexible modelling in prognostic studies and encourage others to explore similar flexible analyses (Binquet et al., 2008, Binquet et al., 2009).

### 10.1.2 Evaluation of IHC4 and Mammostrat Residual Risk Panels

An aim of the thesis was to compare two prognostic IHC biomarker panels, IHC4 and Mammostrat, to determine which provides more information on the risk of recurrence in the context of additional clinical factors or whether combining both approaches would increase the information available to patients and clinicians.

*Comparison of IHC4 and Mammostrat*

The first analysis fitted the standard Cox proportional hazard model and the assumptions of linearity and proportional hazards were assessed.

IHC4 and Mammostrat risk scores were significantly associated with time to distant recurrence and added prognostic information beyond that provided by clinical factors. As a single panel IHC4 provided more information than Mammostrat in the TEAM cohort whereas in the Edinburgh BCS cohort the Mammostrat score was the stronger predictor of TTDR. However, particularly in the larger TEAM cohort and when all ER-positive patients were considered, statistically significant benefit in estimation of residual recurrence risk after treatment was observed from a combined use of both marker panels.

However, violations of the proportional hazards assumption in Cox regression analysis were observed.

Three risk groups were identified from a combined IHC4 and Mammostrat score, with those classified as high risk IHC4 and high risk Mammostrat score at an increased risk of distant recurrence compared to those classified as low risk IHC4 and low risk Mammostrat score.

*Analysis Incorporating Non-Proportional effects*

The next stage of the analysis was to identify and incorporate non-proportional effects using methods identified in the literature review.

The analyses identified a strong time-varying effect of IHC4 score. The prognostic effect of IHC4 score on TTDR decreased with increasing follow-up time. IHC4 score appeared to be prognostic of early distant recurrence only (0-5 years) when categorising time into 5-yearly intervals. Previous analysis by Sgroi and colleagues also confirmed a significant prognostic ability for IHC4 for early distant recurrence only (0-5 years) (Sgroi et al., 2013).

When evaluating the score continuously over time, IHC4 score was prognostic of outcome up to approximately 7 years in the TEAM cohort but only up to 4 years in the Edinburgh BCS cohort for all ER-positive patients. The longer follow-up in the Edinburgh BCS cohort identified a protective effect of higher IHC4 scores beyond 11 years of follow-up.

Haibe-Kains et al., stated the time-dependence of prognostic scores may be due to (i) the biology, as the biological mechanisms for the appearance of early and late relapses are suggested to be different; (ii) the statistical method, the scores being developed on cohorts without long follow-up; (iii) the quality of survival data, where one could intuitively think that the quality of survival data decreases with respect to duration of follow-up since it is difficult to follow patients during a long period resulting in a high level of censoring (Haibe-Kains et al., 2008).

When evaluating the 0-5 year and 5-10 year intervals, Mammostrat score was prognostic of early distant recurrence (0-5 years). Despite not having an overall significant effect in the 5-10 year interval there was evidence that those with high risk Mammostrat score were at increased risk of late recurrence (5-10 years) compared to those low risk for ER-positive, node-negative patients, treated with endocrine therapy only.

There was no evidence of a violation of the proportional hazards for Mammostrat score evaluating continuous-time. Mammostrat score identified a group of patients with an increased risk of distant recurrence over full follow-up in the TEAM cohort (10 years) and the Edinburgh BCS cohort (25 years).

The IHC4 score provided more prognostic information on TTDR than the Mammostrat score in the first 5 years of follow-up except for all ER-positive patients in the larger TEAM cohort where the addition of both scores provided statistically significant information.

With the inclusion of a time-varying effect of IHC4 score, the addition of both scores provided statistically significant prognostic information on TTDR for all ER-positive patients in both the Edinburgh BCS and TEAM cohorts considering full follow-up.

In the analysis of the combined IHC4/Mammostrat score, this risk score was prognostic of outcome in all patient subgroups in the TEAM cohort and identified a group of patients at an increased risk of recurrence over the 10-year follow-up. In the Edinburgh BCS cohort, the

combined score identified a group of patients with a high risk of distant recurrence in the first five years since diagnosis, especially within the first two years.

*Evaluation of Individual Markers*

The individual biomarkers from IHC4 and Mammostrat residual risk panels were analysed to determine which have a prognostic effect on outcome and which, if any, have a time-dependent effect.

A combination of markers from both IHC4 and Mammostrat score were selected as prognostic of outcome. This agrees with the results that a combined IHC4 and Mammostrat risk score provides prognostic information on time to distant recurrence.

Ki67 was identified as having a strong time-dependent effect which may be a component of the strong time-varying effect that was identified for IHC4 score.

*Conclusion*

It is important to accurately identify women with a low risk of recurrence for avoidance of chemotherapy and at low risk of late distant recurrence as some of them may be spared extended endocrine therapy whereas others may benefit from further treatment. These results suggest the possible use of a combined IHC4 and Mammostrat risk score to predict the risk of recurrence, which will need to be investigated further on other patient cohorts with long follow-up.

With many prognostic models available in breast cancer, it is important to evaluate and compare the residual risk models, and this research has added to the growing evidence base. The results of a recent trial in this area of research will be of interest, the OPTIMA trial (Bartlett et al., 2013). OPTIMA-prelim is a preliminary concordance and cost-effectiveness analysis of multiple different assays (including IHC4 and Mammostrat), in order to identify

the most cost-effective tests for a larger efficacy trial, OPTIMA-main. The main trial will be a test of whether biomarker-directed therapy is better or worse than standard care.

## 10.2 Limitations

The limitations of the literature review include the reliance on published research and identification of the studies using the method outlined in the search strategy. It is not claimed that this review includes all methodology relating to the analysis of non-proportional effects. However, a relatively large number of papers were included in the final review, giving a broad and comprehensive overview of approaches for analysing non-proportional effects.

The application of the approaches in practice was assessed using Google Scholar citations. This method has its criticisms, including the possible over-estimation of citations as they include most online documents and do not factor out self-citations.

It was decided to focus on two approaches highlighted in the literature review. It would be of interest to also evaluate some of the other approaches such as the product model introduced by Abrahamowicz and colleagues, the Cox-Aalen and some of the novel Bayesian approaches (Abrahamowicz and MacKenzie, 2007, Scheike and Zhang, 2002).

A potential criticism is that the Edinburgh BCS cohort may not be representative of current breast cancer patients as the patients were enrolled between 1981 and 1998. However, this does enable the evaluation of long-term follow-up and the Edinburgh BCS cohort is a considerably larger patient cohort than previous analyses including long-term follow-up (Drukker et al., 2014, Haibe-Kains et al., 2008)

Another issue with the Edinburgh BCS cohort was the large amount of missing data. The current recommended approaches with multiple imputation was used to overcome the biases that occur when performing complete-case analysis (Burton and Altman, 2004, Vergouwe et al., 2010).

## 10.3 Future Directions

Future directions include further examination of a combined IHC4 and Mammostrat score in a prospectively defined analysis to assess the predictive ability for chemotherapy and extended adjuvant therapy.

The Mammostrat score is a combination of 5 binary markers (negative/positive) resulting in a non-continuous distribution. An advantage of the IHC4 score is its continuous distribution and additional prognostic information may be provided if Mammostrat score was also more continuous. The markers in the Mammostrat panel were originally scored on a semi-quantitative scale and were defined as negative, weak, or strong. An analysis into whether developing a score incorporating the 3 levels of each marker would provide additional prognostic information would be of interest.

The TEAM trial is still accumulating outcome data, an evaluation of the prognostic ability of IHC4 and Mammostrat score with longer follow-up data would be beneficial in confirming the ability of the scores in predicting early and late distant recurrence. The additional follow-up data will also allow exploration of whether the time-varying effect of IHC4 score is a 'real' effect or if it is potentially due to a lack of events.

The focus of the thesis was on the effects of ICH4 and Mammostrat score. It would be of interest to also consider the prognostic clinical variables (grade, tumour size, tumour grade) in greater detail. For example, whether tumour grade provides additional prognostic information to the residual risk panels or whether the panels are a substitution for grade, providing more accurate prognoses.

The individual biomarkers were analysed to illustrate how the use of novel methods could potentially improve the modelling of the biomarkers by incorporating non-proportional effects. Differences in the effects of the individual markers were found between the two cohorts and with and without the inclusion of clinical variables. It would be of interest to look

at these differences in more detail, to gain a greater understanding of these biomarkers, in particular the effect of Ki67.

With the development of numerous prognostic models in breast cancer, it is important to assess the performance of the models with measures of discrimination and calibration. Further work is needed on the methodology to evaluate a prognostic model which incorporates a time-varying effect.

A potential issue was identified in the Edinburgh BCS cohort, where the dataset was potentially too small for the more flexible analyses. It would useful to have a general recommendation for the numbers at events and numbers at risk needed to perform more complex modelling for including time-varying effects in the Cox model and when including modelling using the flexible parametric models.

The thesis highlighted there is a need for more widespread use of flexible modelling in prognostic studies. The publication of the literature review, would provide a comprehensive overview of available methods for the analysis of non-proportional effects with survival data. To encourage others to explore similar flexible analyses on prognostic studies, tutorial papers. and practical workshops would be advantageous.

## 10.4 Conclusion

In conclusion, this thesis has expanded on the basic approach of splitting the time scale when evaluating the impact of follow-up time in the analysis of residual risk models in breast cancer. The use of two novel approaches were illustrated for the inclusion of non-proportional effects, with Royston-Parmer flexible parametric models having many advantages including modelling of the baseline hazard, estimation of absolute risk estimates and the ease of which time-dependent effects can be fitted.

The prognostic ability of IHC4 and Mammostrat were compared, which to our knowledge has not been done previously. The two scores have not only been compared, but it was evaluated whether the two scores have a combined use in providing information on the risk of recurrence, which has not been a focus in previous comparisons of models. The results suggest a combined IHC4 and Mammostrat risk score could provide information on the risk of recurrence and warrants further study.

# References


AALEN, O. O. 1980. A model for nonparametric regression analysis of counting processes. *Lecture Notes in Statistics 2, 1-25 Springer-Verlag*.

AALEN, O. O. 1989. A LINEAR-REGRESSION MODEL FOR THE ANALYSIS OF LIFE TIMES. *Statistics in Medicine,* 8**,** 907-925.

AALEN, O. O. 1993. FURTHER RESULTS ON THE NONPARAMETRIC LINEAR-REGRESSION MODEL IN SURVIVAL ANALYSIS. *Statistics in Medicine,* 12**,** 1569-1588.

ABRAHAMOWICZ, M., MACKENZIE, T. & ESDAILE, J. M. 1996. Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association,* 91**,** 1432-1439.

ABRAHAMOWICZ, M. & MACKENZIE, T. A. 2007. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine,* 26**,** 392-408.

ABRAHAMOWICZ, M., SCHOPFLOCHER, T., LEFFONDRE, K., DU BERGER, R. & KREWSKI, D. 2003. Flexible modeling of exposure-response relationship between long-term average levels of particulate air pollution and mortality in the American Cancer Society Study. *Journal of Toxicology and Environmental Health-Part A,* 66**,** 1625-1653.

AKAIKE, H. 1974. NEW LOOK AT STATISTICAL-MODEL IDENTIFICATION. *Ieee Transactions on Automatic Control,* AC19**,** 716-723.

ALBAIN, K. S., BARLOW, W. E., SHAK, S., HORTOBAGYI, G. N., LIVINGSTON, R. B., YEH, I. T., RAVDIN, P., BUGARINI, R., BOEHNER, F. L., DAVIDSON, N. E., SLEDGE, G. W., WINER, E. P., HUDIS, C., INGLE, J. N., PEREZ, E. A., PRITCHARD, K. I., SHEPHERD, L., GRALOW, J. R., YOSHIZAWA, C., ALLRED, D. C., OSBORNE, C. K., HAYES, D. F. & BREAST CANC INTERGRP, N. A. 2010. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *Lancet Oncology,* 11**,** 55-65.

ALTMAN, D. G. 2009. Prognostic Models: A Methodological Framework and Review of Models for Breast Cancer. *Cancer Investigation,* 27**,** 235-243.

ALTMAN, D. G., DESTAVOLA, B. L., LOVE, S. B. & STEPNIEWSKA, K. A. 1995. REVIEW OF SURVIVAL ANALYSES PUBLISHED IN CANCER JOURNALS. *British Journal of Cancer,* 72.

ALTMAN, D. G., VERGOUWE, Y., ROYSTON, P. & MOONS, K. G. M. 2009. Prognosis and prognostic research: validating a prognostic model. *British Medical Journal,* 338.

ANDERSON, J. A. & SENTHILSELVAN, A. 1982. A Two-Step Regression Model for Hazard Functions. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 31**,** 44-51.

ARANDA-ORDAZ, F. J. 1981. On two families of transformations to additivity for binary response data. *Biometrika,* 68**,** 357–363.

ATA, N. & SOZER, M. T. 2007. COX REGRESSION MODELS WITH NONPROPORTIONAL HAZARDS APPLIED TO LUNG CANCER SURVIVAL DATA. *Hacettepe Journal of Mathematics and Statistics,* 36**,** 157-167.

ATKINSON, A. J., COLBURN, W. A., DEGRUTTOLA, V. G., DEMETS, D. L., DOWNING, G. J., HOTH, D. F., OATES, J. A., PECK, C. C., SCHOOLEY, R. T., SPILKER, B. A., WOODCOCK, J., ZEGER, S. L. & BIOMARKERS DEFINITIONS WORKING, G. 2001. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics,* 69**,** 89-95.

AVNINDER, S., YLAYA, K. & HEWITT, S. M. 2008. Tissue microarray: A simple technology that has revolutionized research in pathology. *Journal of Postgraduate Medicine,* 54**,** 158-162.

BALDI, I., CICCONE, G., PONTI, A., ROSSO, S., ZANETTI, R. & GREGORI, D. 2006. An Application of the Cox-Aalen Model for Breast Cancer Survival. *AUSTRIAN JOURNAL OF STATISTICS,* 35**,** 77-78.

BALSLEV, I., AXELSSON, C. K., ZEDELER, K., RASMUSSEN, B. B., CARSTENSEN, B. & MOURIDSEN, H. T. 1994. THE NOTTINGHAM-PROGNOSTIC-INDEX APPLIED TO 9,149 PATIENTS FROM THE STUDIES OF THE DANISH-BREAST-CANCER-COOPERATIVE-GROUP (DBCG). *Breast Cancer Research and Treatment,* 32**,** 281-290.

BANERJEE, T., CHEN, M. H., DEY, D. K. & KIM, S. 2007. Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime Data Analysis,* 13 (2)**,** 241-260.

BARTLETT, J., CANNEY, P., CAMPBELL, A., CAMERON, D., DONOVAN, J., DUNN, J., EARL, H., FRANCIS, A., HALL, P., HARMER, V., HIGGINS, H., HILLIER, L., HULME, C., HUGHES-DAVIES, L., MAKRIS, A., MORGAN, A., MCCABE, C., PINDER, S., POOLE, C., REA, D., STALLARD, N. & STEIN, R. 2013. Selecting Breast Cancer Patients for Chemotherapy: The Opening of the UK OPTIMA Trial. *Clinical Oncology,* 25**,** 109-116.

BARTLETT, J. M. S., BLOOM, K. J., PIPER, T., LAWTON, T. J., VAN DE VELDE, C. J. H., ROSS, D. T., RING, B. Z., SEITZ, R. S., BECK, R. A., HASENBURG, A., KIEBACK, D., PUTTER, H., MARKOPOULOS, C., DIRIX, L., SEYNAEVE, C. & REA, D. 2012. Mammostrat As an Immunohistochemical Multigene Assay for Prediction of Early Relapse Risk in the Tamoxifen Versus Exemestane Adjuvant Multicenter Trial Pathology Study. *Journal of Clinical Oncology,* 30**,** 4477-4484.

BARTLETT, J. M. S., BROOKES, C. L., ROBSON, T., VAN DE VELDE, C. J. H., BILLINGHAM, L. J., CAMPBELL, F. M., GRANT, M., HASENBURG, A., HILLE, E. T. M., KAY, C., KIEBACK, D. G., PUTTER, H., MARKOPOULOS, C., KRANENBARG, E. M.-K., MALLON, E. A., DIRIX, L.,

SEYNAEVE, C. & REA, D. 2011a. Estrogen Receptor and Progesterone Receptor As Predictive Biomarkers of Response to Endocrine Therapy: A Prospectively Powered Pathology Study in the Tamoxifen and Exemestane Adjuvant Multinational Trial. *Journal of Clinical Oncology,* 29**,** 1531-1538.

BARTLETT, J. M. S., STARCZYNSKI, J., ATKEY, N., KAY, E., O'GRADY, A., GANDY, M., IBRAHIM, M., JASANI, B., ELLIS, I. O., PINDER, S. E. & WALKER, R. A. 2011b. HER2 testing in the UK: recommendations for breast and gastric in-situ hybridisation methods. *Journal of Clinical Pathology,* 64**,** 649-653.

BARTLETT, J. M. S., THOMAS, J., ROSS, D. T., SEITZ, R. S., RING, B. Z., BECK, R. A., PEDERSEN, H. C., MUNRO, A., KUNKLER, I. H., CAMPBELL, F. M., JACK, W., KERR, G. R., JOHNSTONE, L., CAMERON, D. A. & CHETTY, U. 2010. Mammostrat (R) as a tool to stratify breast cancer patients at risk of recurrence during endocrine therapy. *Breast Cancer Research,* 12.

BARTON, S., ZABAGLO, L., A'HERN, R., TURNER, N., FERGUSON, T., O'NEILL, S., HILLS, M., SMITH, I. & DOWSETT, M. 2012. Assessment of the contribution of the IHC4+C score to decision making in clinical practice in early breast cancer. *British Journal of Cancer,* 106**,** 1760-1765.

BELLERA, C. A., MACGROGAN, G., DEBLED, M., DE LARA, C. T., BROUSTE, V. & MATHOULIN-PELISSIER, S. 2010. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Medical Research Methodology,* 10**,** 20.

BERGER, U., SCHAFER, J. & ULM, K. 2003. Dynamic Cox modelling based on fractional polynomials: time-variations in gastric cancer prognosis. *Statistics in Medicine,* 22**,** 1163-1180.

BHOO-PATHY, N., YIP, C.-H., HARTMAN, M., SAXENA, N., TAIB, N. A., HO, G.-F., LOOI, L.-M., BULGIBA, A. M., VAN DER GRAAF, Y. & VERKOOIJEN, H. M. 2012. Adjuvant! Online is overoptimistic in predicting survival of Asian breast cancer patients. *European Journal of Cancer,* 48**,** 982-989.

BIAGINI, M. R., MCCORMICK, P. A., GUARDASCIONE, M., SURRENTI, C. & BURROUGHS, A. K. 1991. Prognosis in primary biliary cirrhosis. A review. *Italian Journal of Gastroenterology,* 23**,** 222-6.

BIGANZOLI, E., BORACCHI, P., MARIANI, L. & MARUBINI, E. 1998. Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statistics in Medicine,* 17 (10)**,** 1169-1186.

BINQUET, C., ABRAHAMOWICZ, M., ASTRUC, K., FAIVRE, J., BONITHON-KOPP, C. & QUANTIN, C. 2009. *Flexible statistical models provided new insights into the role of quantitative prognostic factors for mortality in gastric cancer*, Journal of Clinical Epidemiology. 62 (3) (pp 232-240), 2009. Date of Publication: March 2009.

BINQUET, C., ABRAHAMOWICZ, M., MAHBOUBI, A., JOOSTE, V., FAIVRE, J., BONITHON-KOPP, C. & QUANTIN, C. 2008. Empirical study of the dependence of the results of multivariable flexible survival analyses on model selection strategy. *Statistics in Medicine,* 27 (30)**,** 6470-6488.

BLOWS, F. M., DRIVER, K. E., SCHMIDT, M. K., BROEKS, A., VAN LEEUWEN, F. E., WESSELING, J., CHEANG, M. C., GELMON, K., NIELSEN, T. O., BLOMQVIST, C., HEIKKILA, P., HEIKKINEN, T., NEVANLINNA, H., AKSLEN, L. A., BEGIN, L. R., FOULKES, W. D., COUCH, F. J., WANG, X., CAFOUREK, V., OLSON, J. E., BAGLIETTO, L., GILES, G. G., SEVERI, G., MCLEAN, C. A., SOUTHEY, M. C., RAKHA, E., GREEN, A. R., ELLIS, I. O., SHERMAN, M. E., LISSOWSKA, J., ANDERSON, W. F., COX, A., CROSS, S. S., REED, M. W. R., PROVENZANO, E., DAWSON, S.-J., DUNNING, A. M., HUMPHREYS, M., EASTON, D. F., GARCIA-CLOSAS, M., CALDAS, C., PHAROAH, P. D. & HUNTSMAN, D. 2010. Subtyping of Breast Cancer by Immunohistochemistry to Investigate a Relationship between Subtype and Short and Long Term Survival: A Collaborative Analysis of Data for 10,159 Cases from 12 Studies. *Plos Medicine,* 7.

BOGAERTS, J., CARDOSO, F., BUYSE, M., BRAGA, S., LOI, S., HARRISON, J. A., BINES, J., MOOK, S., DECKER, N., RAVDIN, P., THERASSE, P., RUTGERS, E., VAN'T VEER, L. J., PICCART, M. & CONSORTIUM, T. 2006. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nature Clinical Practice Oncology,* 3**,** 540-551.

BRADBURN, M. J., CLARK, T. G., LOVE, S. B. & ALTMAN, D. G. 2003. Survival Analysis Part II: Multivariate data analysis - an introduction to concepts and methods. *British Journal of Cancer,* 89.

BRESLOW, N. 1974. COVARIANCE ANALYSIS OF CENSORED SURVIVAL DATA. *Biometrics,* 30**,** 89-99.

BROWN, D., KAUERMANN, G. & FORD, I. 2007. A partial likelihood approach to smooth estimation of dynamic covariate effects using penalised splines. *Biometrical Journal,* 49**,** 441-52.

BUCHHOLZ, A. 2010. *Assessment of time–varying long–term effects of therapies and prognostic factors.* Doctor of Natural Sciences, Technical University of Dortmund.

BUCHHOLZ, A. & SAUERBREI, W. 2011. Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data. *Biometrical Journal,* 53**,** 308-331.

BUENO-DE-MESQUITA, J. M., LINN, S. C., KEIJZER, R., WESSELING, J., NUYTEN, D. S. A., VAN KRIMPEN, C., MEIJERS, C., DE GRAAF, P. W., BOS, M. M. E. M., HART, A. A. M., RUTGERS, E. J. T., PETERSE, J. L., HALFWERK, H., DE GROOT, R., PRONK, A., FLOORE, A. N., GLAS, A. M., VAN'T VEER, L. J. & VAN DE VIJVER, M. J. 2009. Validation of 70-gene prognosis signature in node-negative breast cancer. *Breast Cancer Research and Treatment,* 117**,** 483-495.

BURTON, A. & ALTMAN, D. G. 2004. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer,* 91**,** 4-8.

BUYSE, M., LOI, S., VAN'T VEER, L., VIALE, G., DELORENZI, M., GLAS, A. M., D'ASSIGNIES, M. S., BERGH, J., LIDEREAU, R., ELLIS, P., HARRIS, A., BOGAERTS, J., THERASSE, P., FLOORE, A., AMAKRANE, M., PIETTE, F., RUTGERS, E., SOTIRIOU, C., CARDOSO, F., PICCART, M. J. & CONSORTIUM, T. 2006. Validation and clinical utility of a 70-gene

prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute,* 98**,** 1183-1192.

CAI, Z. W. & SUN, Y. Q. 2003. Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics,* 30**,** 93-111.

CALIFF, R. M., SHAH, S. H. & NEWBY, L. K. 2010. Biomarker Bonanza? *Journal of the American College of Cardiology,* 55**,** 1197-1199.

CAMPBELL, H. E., TAYLOR, M. A., HARRIS, A. L. & GRAY, A. M. 2009. An investigation into the performance of the Adjuvant! Online prognostic programme in early breast cancer for a cohort of patients in the United Kingdom. *British Journal of Cancer,* 101**,** 1074-1084.

CANCER RESEARCH UK. 2012. *Breast Cancer - UK Incidence Statistics* [Online]. Available: http://info.cancerresearchuk.org/cancerstats/types/breast/incidence/ [Accessed 12 Jun 2012].

CANCER RESEARCH UK. 2015. *Cancer Research UK* [Online]. Available: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer [Accessed 19/08/2015 2015].

CHANDLER, I., HOULSTON, R. & LANDBERG, G. 2011. A Practical Guide to Constructing and Using Tissue Microarrays. *In:* DILLNER, J. (ed.) *Methods in Biobanking.*

CHENG, L., SWARTZ, M. D., ZHAO, H., KAPADIA, A. S., LAI, D., ROWAN, P. J., BUCHHOLZ, T. A. & GIORDANO, S. H. 2012. Hazard of Recurrence among Women after Primary Breast Cancer Treatment-A 10-Year Follow-up Using Data from SEER-Medicare. *Cancer Epidemiology Biomarkers & Prevention,* 21**,** 800-809.

CHIA, S. K., BRAMWELL, V. H., TU, D., SHEPHERD, L. E., JIANG, S., VICKERY, T., MARDIS, E., LEUNG, S., UNG, K., PRITCHARD, K. I., PARKER, J. S., BERNARD, P. S., PEROU, C. M., ELLIS, M. J. & NIELSEN, T. O. 2012. A 50-Gene Intrinsic Subtype Classifier for Prognosis and Prediction of Benefit from Adjuvant Tamoxifen. *Clinical Cancer Research,* 18**,** 4465-4472.

CHOODARI-OSKOOEI, B., ROYSTON, P. & PARMAR, M. K. B. 2012. A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine,* 31**,** 2627-2643.

CLARK, G. M. 1994. DO WE REALLY NEED PROGNOSTIC FACTORS FOR BREAST-CANCER. *Breast Cancer Research and Treatment,* 30**,** 117-126.

COLZANI, E., LILJEGREN, A., JOHANSSON, A. L. V., ADOLFSSON, J., HELLBORG, H., HALL, P. F. L. & CZENE, K. 2011. Prognosis of Patients With Breast Cancer: Causes of Death and Effects of Time Since Diagnosis, Age, and Tumor Characteristics. *Journal of Clinical Oncology,* 29.

CORADINI, D., DAIDONE, M. G., BORACCHI, P., BIGANZOLI, E., ORIANA, S., BRESCIANI, G., PELLIZZARO, C., TOMASIC, G., DI FRONZO, G. & MARUBINI, E. 2000. Time-dependent relevance of steroid receptors in breast cancer. *Journal of Clinical Oncology,* 18**,** 2702-2709.

CORTESE, G., SCHEIKE, T. H. & MARTINUSSEN, T. 2010. Flexible survival regression modelling. *Statistical Methods in Medical Research,* 19**,** 5-28.

COSMATOS, D. & CHOW, S.-C. (eds.) 2008. *Translational Medicine Strategies and Statistical Methods*: Chapman & Hall/CRC Biostatistics Series.

COSTA, M. J. & SHAW, J. E. H. 2009. Parametrization and penalties in spline models with an application to survival analysis. *Computational Statistics & Data Analysis,* 53**,** 657-670.

COX, D. R. 1972. REGRESSION MODELS AND LIFE-TABLES. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* 34**,** 187-&.

COX, D. R. 1975. PARTIAL LIKELIHOOD. *Biometrika,* 62**,** 269-276.

CUZICK, J., DOWSETT, M., PINEDA, S., WALE, C., SALTER, J., QUINN, E., ZABAGLO, L., MALLON, E., GREEN, A. R., ELLIS, I. O., HOWELL, A., BUZDAR, A. U. & FORBES, J. F. 2011. Prognostic Value of a Combined Estrogen Receptor, Progesterone Receptor, Ki-67, and Human Epidermal Growth Factor Receptor 2 Immunohistochemical Score and Comparison With the Genomic Health Recurrence Score in Early Breast Cancer. *Journal of Clinical Oncology,* 29**,** 4273-4278.

DABROWSKA, D. M. 1997. Smoothed Cox regression. *Annals of Statistics,* 25**,** 1510-1540.

DAVIES, C., PAN, H., GODWIN, J., GRAY, R., ARRIAGADA, R., RAINA, V., ABRAHAM, M., MEDEIROS ALENCAR, V. H., BADRAN, A., BONFILL, X., BRADBURY, J., CLARKE, M., COLLINS, R., DAVIS, S. R., DELMESTRI, A., FORBES, J. F., HADDAD, P., HOU, M.-F., INBAR, M., KHALED, H., KIELANOWSKA, J., KWAN, W.-H., MATHEW, B. S., MITTRA, I., MUELLER, B., NICOLUCCI, A., PERALTA, O., PERNAS, F., PETRUZELKA, L., PIENKOWSKI, T., RADHIKA, R., RAJAN, B., RUBACH, M. T., TORT, S., URRUTIA, G., VALENTINI, M., WANG, Y., PETO, R. & ADJUVANT TAMOXIFEN LONGER, A. 2013. Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. *Lancet,* 381**,** 805-816.

DE BOOR, C. 1978. *A Practical Guide to Splines*, Springer.

DE IORIO, M., JOHNSON, W. O., MULLER, P. & ROSNER, G. L. 2009. Bayesian Nonparametric Nonproportional Hazards Survival Modeling. *Biometrics,* 65 (3)**,** 762-771.

DEARNALEY, D. P., SYDES, M. R., GRAHAM, J. D., AIRD, E. G., BOTTOMLEY, D., COWAN, R. A., HUDDART, R. A., JOSE, C. C., MATTHEWS, J. H. L., MILLAR, J., MOORE, A. R., MORGAN, R. C., RUSSELL, J. M., SCRASE, C. D., STEPHENS, R. J., SYNDIKUS, I., PARMAR, M. K. B. & COLLABORATORS, R. T. 2007. Escalated-dose versus standard-dose conformal radiotherapy in prostate cancer: first results from the MRC RT01 randomised controlled trial. *Lancet Oncology,* 8.

DESMEDT, C., PIETTE, F., LOI, S., WANG, Y., D'ASSIGNIES, M. S., BERGH, J., LIDEREAU, R., ELLIS, P., HARRIS, A. L., KLIJN, J. G. M., FOEKENS, J. A., CARDOSO, F., PICCART, M. J., BUYSE, M., SOTIRIOU, C. & CONSORTIUM, T. 2007. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical Cancer Research,* 13**,** 3207-3214.

DOWSETT, M., CUZICK, J., WALE, C., FORBES, J., MALLON, E. A., SALTER, J., QUINN, E., DUNBIER, A., BAUM, M., BUZDAR, A., HOWELL, A., BUGARINI, R., BAEHNER, F. L. & SHAK, S. 2010. Prediction of Risk of Distant Recurrence Using the 21-Gene Recurrence Score in Node-Negative and Node-Positive Postmenopausal Patients With Breast Cancer Treated With Anastrozole or Tamoxifen: A TransATAC Study. *Journal of Clinical Oncology,* 28.

DOWSETT, M., CUZICK, J., WALES, C., FORBES, J., MALLON, L., SALTER, J., QUINN, E., BUGARINI, R., BAEHNER, F. L., SHAK, S. & GRP, A. T. 2009. Risk of distant recurrence using oncotype DX in postmenopausal primary breast cancer patients treated with anastrozole or tamoxifen: a TransATAC study. *Cancer Research,* 69**,** 75S-76S.

DOWSETT, M., NIELSEN, T. O., A'HERN, R., BARTLETT, J., COOMBES, R. C., CUZICK, J., ELLIS, M., HENRY, N. L., HUGH, J. C., LIVELY, T., MCSHANE, L., PAIK, S., PENAULT-LLORCA, F., PRUDKIN, L., REGAN, M., SALTER, J., SOTIRIOU, C., SMITH, I. E., VIALE, G., ZUJEWSKI, J. A. & HAYES, D. F. 2011. Assessment of Ki67 in Breast Cancer: Recommendations from the International Ki67 in Breast Cancer Working Group. *Journal of the National Cancer Institute,* 103**,** 1656-1664.

DOWSETT, M., SESTAK, I., LOPEZ-KNOWLES, E., SIDHU, K., DUNBIER, A. K., COWENS, J. W., FERREE, S., STORHOFF, J., SCHAPER, C. & CUZICK, J. 2013. Comparison of PAM50 Risk of Recurrence Score With Oncotype DX and IHC4 for Predicting Risk of Distant Recurrence After Endocrine Therapy. *Journal of Clinical Oncology,* 31**,** 2783-+.

DRUKKER, C. A., BUENO-DE-MESQUITA, J. M., RETEL, V. P., VAN HARTEN, W. H., VAN TINTEREN, H., WESSELING, J., ROUMEN, R. M. H., KNAUER, M., VAN 'T VEER, L. J., SONKE, G. S., RUTGERS, E. J. T., VAN DE VIJVER, M. J. & LINN, S. C. 2013. A prospective evaluation of a breast cancer prognosis signature in the observational RASTER study. *International Journal of Cancer,* 133**,** 929-936.

DRUKKER, C. A., VAN TINTEREN, H., SCHMIDT, M. K., RUTGERS, E. J. T., BERNARDS, R., VAN DE VIJVER, M. J. & VAN'T VEER, L. J. 2014. Long-term impact of the 70-gene signature on breast cancer outcome. *Breast Cancer Research and Treatment,* 143**,** 587-592.

DUBSKY, P., BRASE, J. C., JAKESZ, R., RUDAS, M., SINGER, C. F., GREIL, R., DIETZE, O., LUISSER, I., KLUG, E., SEDIVY, R., BACHNER, M., MAYR, D., SCHMIDT, M., GEHRMANN, M. C., PETRY, C., WEBER, K. E., FISCH, K., KRONENWETT, R., GNANT, M., FILIPITS, M. & AUSTRIAN BREAST COLORECTAL CANC, S. 2013. The EndoPredict score provides prognostic information on late distant metastases in ER+/HER2-breast cancer patients. *British Journal of Cancer,* 109**,** 2959-2964.

DUNSON, D. B. & HERRING, A. H. 2005. Bayesian model selection and averaging in additive and proportional hazards models. *Lifetime Data Analysis,* 11**,** 213-232.

DURAIYAN, J., GOVINDARAJAN, R., KALIYAPPAN, K. & PALANISAMY, M. 2012. Applications of immunohistochemistry. *Journal of pharmacy & bioallied sciences,* 4**,** S307-9.

EARLY BREAST CANCER TRIALISTS' COLLABORATIVE, G. 2005. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet,* 365**,** 1687-717.

ELSTON, C. W., ELLIS, I. O. & PINDER, S. E. 1999. Pathological prognostic factors in breast cancer. *Critical Reviews in Oncology Hematology,* 31**,** 209-223.

EUROPEAN MEDICINES AGENCY 2011. Reflection paper on methodological issues associated with pharmacogenomic biomarkers in relation to clinical development and patient selection.

FARATIAN, D., MUNRO, A., TWELVES, C. & BARTLETT, J. M. S. 2009. Membranous and cytoplasmic staining of Ki67 is associated with HER2 and ER status in invasive breast carcinoma. *Histopathology,* 54**,** 254-257.

FAREWELL, V. T. 1982. THE USE OF MIXTURE-MODELS FOR THE ANALYSIS OF SURVIVAL-DATA WITH LONG-TERM SURVIVORS. *Biometrics,* 38.

FISHER, B., REDMOND, C., FISHER, E. R. & CAPLAN, R. 1988. RELATIVE WORTH OF ESTROGEN OR PROGESTERONE-RECEPTOR AND PATHOLOGIC CHARACTERISTICS OF DIFFERENTIATION AS INDICATORS OF PROGNOSIS IN NODE NEGATIVE BREAST-CANCER PATIENTS - FINDINGS FROM NATIONAL SURGICAL ADJUVANT BREAST AND BOWEL PROJECT PROTOCOL B-06. *Journal of Clinical Oncology,* 6**,** 1076-1087.

GAGNON, B., ABRAHAMOWICZ, M., XIAO, Y., BEAUCHAMP, M. E., MACDONALD, N., KASYMJANOVA, G., KREISMAN, H. & SMALL, D. 2010. Flexible modeling improves assessment of prognostic value of C-reactive protein in advanced non-small cell lung cancer. *British Journal of Cancer,* 102.

GALEA, M. H., BLAMEY, R. W., ELSTON, C. E. & ELLIS, I. O. 1992. THE NOTTINGHAM PROGNOSTIC INDEX IN PRIMARY BREAST-CANCER. *Breast Cancer Research and Treatment,* 22**,** 207-219.

GAMERMAN, D. 1991. DYNAMIC BAYESIAN MODELS FOR SURVIVAL-DATA. *Applied Statistics-Journal of the Royal Statistical Society Series C,* 40**,** 63-79.

GOLDHIRSCH, A., GLICK, J. H., GELBER, R. D., COATES, A. S., THURLIMANN, B., SENN, H., ALBAIN, K. S., BERGH, J., CASTIGLIONE-GERTSCH, M., COSTA, A., CUZICK, J., DAVIDSON, N., FORBES, J. F., GOSS, P., HARRIS, J., HOWELL, A., INGLE, J. N., JAKESZ, R., JASSEM, J., KAUFMANN, M., MARTIN, M., MAURIAC, L., MORROW, M., MOURIDSEN, H. T., NAMER, M., PICCART-GEBHART, M. J., POSSINGER, K., PRITCHARD, K., RUTGERS, E. J. T., VIALE, G., WALLGREN, A. & WOOD, W. C. 2005. Meeting highlights: International Expert Consensus on the Primary Therapy of Early Breast Cancer 2005. *Annals of Oncology,* 16**,** 1569-1583.

GORE, S. M., POCOCK, S. J. & KERR, G. R. 1984. REGRESSION-MODELS AND NON-PROPORTIONAL HAZARDS IN THE ANALYSIS OF BREAST-CANCER SURVIVAL. *Applied Statistics-Journal of the Royal Statistical Society Series C,* 33**,** 176-195.

GOSS, P. E., INGLE, J. N., MARTINO, S., ROBERT, N. J., MUSS, H. B., PICCART, M. J., CASTIGLIONE, M., TU, D., SHEPHERD, L. E., PRITCHARD, K. I., LIVINGSTON, R. B., DAVIDSON, N. E., NORTON, L., PEREZ, E. A., ABRAMS, J. S., THERASSE, P., PALMER, M. J. & PATER, J. L. 2003. A randomized trial of letrozole in postmenopausal women after five years of tamoxifen therapy for early-stage breast cancer. *New England Journal of Medicine,* 349.

GRAMBSCH, P. M. & THERNEAU, T. M. 1994. PROPORTIONAL HAZARDS TESTS AND DIAGNOSTICS BASED ON WEIGHTED RESIDUALS. *Biometrika,* 81**,** 515-526.

GRAY, R. G., REA, D., HANDLEY, K., BOWDEN, S. J., PERRY, P., EARL, H. M., POOLE, C. J., BATES, T., CHETIYAWARDANA, S., DEWAR, J. A., FERNANDO, I. N., GRIEVE, R., NICOLL, J., RAYTER, Z., ROBINSON, A., SALMAN, A., YARNOLD, J., BATHERS, S., MARSHALL, A., LEE, M. & A, T. C. G. 2013. aTTom: Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years in 6,953 women with early breast cancer. *Journal of Clinical Oncology,* 31.

GRAY, R. J. 1992. FLEXIBLE METHODS FOR ANALYZING SURVIVAL-DATA USING SPLINES, WITH APPLICATIONS TO BREAST-CANCER PROGNOSIS. *Journal of the American Statistical Association,* 87**,** 942-951.

GUERRA, I., ALGORTA, J., PELAYO, R. D. D. & FARINA, J. 2003. Immunohistochemical prognostic index for breast cancer in young women. *Journal of Clinical Pathology-Molecular Pathology,* 56**,** 323-327.

GUSTAFSON, P. 1998. Flexible Bayesian modelling for survival data. *Lifetime Data Analysis,* 4 (3)**,** 281-299.

GUSTAFSON, P., AESCHLIMAN, D. & LEVY, A. R. 2003. A simple approach to fitting Bayesian survival models. *Lifetime Data Analysis,* 9**,** 5-19.

HAIBE-KAINS, B., DESMEDT, C., PIETTE, F., BUYSE, M., CARDOSO, F., VAN'T VEER, L., PICCART, M., BONTEMPI, G. & SOTIRIOU, C. 2008. Comparison of prognostic gene expression signatures for breast cancer. *Bmc Genomics,* 9.

HAJAGE, D., DE RYCKE, Y., BOLLET, M., SAVIGNONI, A., CALY, M., PIERGA, J.-Y., HORLINGS, H. M., VAN DE VIJVER, M. J., VINCENT-SALOMON, A., SIGAL-ZAFRANI, B., SENECHAL, C., ASSELAIN, B., SASTRE, X. & REYAL, F. 2011. External Validation of Adjuvant! Online Breast Cancer Prognosis Tool. Prioritising Recommendations for Improvement. *Plos One,* 6.

HANEUSE, S., RUDSER, K. D. & GILLEN, D. L. 2008. The separation of timescales in Bayesian survival modeling of the time-varying effect of a time-dependent exposure. *Biostatistics,* 9**,** 400-410.

HARRELL, F. E. 1986. 'The PHGLM procedure' SAS supplemental Library User's Guide, Version 5 Edition, SAS Institute, Cary, N.C.

HARRIS, L., FRITSCHE, H., MENNEL, R., NORTON, L., RAVDIN, P., TAUBE, S., SOMERFIELD, M. R., HAYES, D. F. & BAST, R. C., JR. 2007. American society of clinical oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of Clinical Oncology,* 25**,** 5287-5312.

HASTIE, T. & TIBSHIRANI, R. 1993. VARYING-COEFFICIENT MODELS. *Journal of the Royal Statistical Society Series B-Methodological,* 55**,** 757-796.

HAYBITTLE, J. L., BLAMEY, R. W., ELSTON, C. W., JOHNSON, J., DOYLE, P. J., CAMPBELL, F. C., NICHOLSON, R. I. & GRIFFITHS, K. 1982. A PROGNOSTIC INDEX IN PRIMARY BREAST-CANCER. *British Journal of Cancer,* 45**,** 361-366.

HE, J., MCGEE, D. L. & NIU, X. 2010. Application of the Bayesian dynamic survival model in medicine. *Statistics in Medicine,* 29**,** 347-360.

HENNERFEIND, A., BREZGER, A. & FAHRMEIR, L. 2006. Geoadditive survival models. *Journal of the American Statistical Association,* 101**,** 1065-1075.

HERTPICCIOTTO, I. & ROCKHILL, B. 1997. Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics,* 53**,** 1151-1156.

HESS, K. R. 1994. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine,* 13**,** 1045-62.

HIELSCHER, T., ZUCKNICK, M., WERFT, W. & BENNER, A. 2010. On the prognostic value of survival models with application to gene expression signatures. *Statistics in Medicine,* 29**,** 818-829.

HILSENBECK, S. G., RAVDIN, P. M., DE MOOR, C. A., CHAMNESS, G. C., OSBORNE, C. K. & CLARK, G. M. 1998. Time-dependence of hazard ratios for prognostic factors in primary breast cancer. *Breast Cancer Research and Treatment,* 52**,** 227-237.

HOPPIN, J. A., TOLBERT, P. E., TAYLOR, J. A., SCHROEDER, J. C. & HOLLY, E. A. 2002. Potential for selection bias with tumor tissue retrieval in molecular epidemiology studies. *Annals of Epidemiology,* 12.

HUFFER, F. W. & MCKEAGUE, I. W. 1991. WEIGHTED LEAST-SQUARES ESTIMATION FOR AALEN ADDITIVE RISK MODEL. *Journal of the American Statistical Association,* 86**,** 114-129.

JATOI, I., ANDERSON, W. F., JEONG, J.-H. & REDMOND, C. K. 2011. Breast Cancer Adjuvant Therapy: Time to Consider Its Time-Dependent Effects. *Journal of Clinical Oncology,* 29**,** 2301-2304.

JAWHAR, N. M. T. 2009. Tissue Microarray: A rapidly evolving diagnostic and research tool. *Annals of Saudi Medicine,* 29**,** 123-127.

KALBFLEISCH, J. D. & PRENTICE, R. L. 1981. ESTIMATION OF THE AVERAGE HAZARD RATIO. *Biometrika,* 68**,** 105-112.

KAPLAN, E. L. & MEIER, P. 1958. NONPARAMETRIC-ESTIMATION FROM INCOMPLETE OBSERVATIONS. *Journal of the American Statistical Association,* 53**,** 457-481.

KIM, Y., PARK, J. K. & KIM, G. 2011. Bayesian analysis for monotone hazard ratio. *Lifetime Data Analysis,* 17 (2)**,** 302-320.

KIRKEGAARD, T., EDWARDS, J., TOVEY, S., MCGLYNN, L. M., KRISHNA, S. N., MUKHERJEE, R., TAM, L., MUNRO, A. F., DUNNE, B. & BARTLETT, J. M. S. 2006. Observer variation in immunohistochemical analysis of protein expression, time for a change? *Histopathology,* 48**,** 787-794.

KNEIB, T. & FAHRMEIR, L. 2007. A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics,* 34**,** 207-228.

LAMBERT, P. & EILERS, P. H. C. 2005. Bayesian proportional hazards model with time-varying regression coefficients: A penalized Poisson regression approach. *Statistics in Medicine,* 24**,** 3977-3989.

LEHR, S. & SCHEMPER, M. 2007. Parsimonious analysis of time-dependent effects in the Cox model. *Statistics in Medicine,* 26**,** 2686-2698.

LIN, D. Y. & YING, Z. L. 1994. SEMIPARAMETRIC ANALYSIS OF THE ADDITIVE RISK MODEL. *Biometrika,* 81**,** 61-71.

MACKENZIE, G. 1996. Regression models for survival data: The generalized time-dependent logistic family. *Statistician,* 45**,** 21-34.

MAMOUNAS, E. P., TANG, G., FISHER, B., PAIK, S., SHAK, S., COSTANTINO, J. P., WATSON, D., GEYER, C. E., JR., WICKERHAM, D. L. & WOLMARK, N. 2010. Association Between the 21-Gene Recurrence Score Assay and Risk of Locoregional Recurrence in Node-Negative, Estrogen Receptor-Positive Breast Cancer: Results From NSABP B-14 and NSABP B-20. *Journal of Clinical Oncology,* 28**,** 1677-1683.

MARTINUSSEN, T. & SCHEIKE, T. H. 2002. A flexible additive multiplicative hazard model. *Biometrika,* 89**,** 283-298.

MARTINUSSEN, T., SCHEIKE, T. H. & SKOVGAARD, I. M. 2002. Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scandinavian Journal of Statistics,* 29**,** 57-74.

MATHOULIN-PELISSIER, S., GOURGOU-BOURGADE, S., BONNETAIN, F. & KRAMAR, A. 2008. Survival end point reporting in randomized cancer clinical trials: A review of major journals. *Journal of Clinical Oncology,* 26.

MCGEECHAN, K., MACASKILL, P., IRWIG, L., LIEW, G. & WONG, T. Y. 2008. Assessing New Biomarkers and Predictive Models for Use in Clinical Practice A Clinician's Guide. *Archives of Internal Medicine,* 168.

MCKEAGUE 1988. A COUNTING PROCESS APPROACH TO THE REGRESSION-ANALYSIS OF GROUPED SURVIVAL-DATA. *Stochastic Processes and Their Applications,* 28**,** 221-239.

MCKEAGUE, I. W. & SASIENI, P. D. 1994. A PARTLY PARAMETRIC ADDITIVE RISK MODEL. *Biometrika,* 81**,** 501-514.

MCKEAGUE, I. W. & TIGHIOUART, M. 2000. Bayesian estimators for conditional hazard functions. *Biometrics,* 56**,** 1007-1015.

MOOK, S., SCHMIDT, M. K., RUTGERS, E. J., VAN DE VELDE, A. O., VISSER, O., RUTGERS, S. M., ARMSTRONG, N., VAN'T VEER, L. J. & RAVDIN, P. M. 2009. Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study. *Lancet Oncology,* 10**,** 1070-1076.

MOONS, K. G. M., ALTMAN, D. G., VERGOUWE, Y. & ROYSTON, P. 2009. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *British Medical Journal,* 338.

MOREAU, T., OQUIGLEY, J. & LELLOUCH, J. 1986. ON D. SCHOENFELD APPROACH FOR TESTING THE PROPORTIONAL HAZARDS ASSUMPTION. *Biometrika,* 73**,** 513-515.

MOREAU, T., OQUIGLEY, J. & MESBAH, M. 1985. A GLOBAL GOODNESS-OF-FIT STATISTIC FOR THE PROPORTIONAL HAZARDS MODEL. *Applied Statistics-Journal of the Royal Statistical Society Series C,* 34**,** 212-218.

MULLIGAN, A. M., PINNADUWAGE, D., BULL, S. B., O'MALLEY, F. P. & ANDRULIS, I. L. 2008. Prognostic effect of basal-like breast cancers is time dependent: Evidence from tissue microarray studies on a lymph node-negative cohort. *Clinical Cancer Research,* 14**,** 4168-4174.

MURPHY, S. A. & SEN, P. K. 1991. TIME-DEPENDENT COEFFICIENTS IN A COX-TYPE REGRESSION-MODEL. *Stochastic Processes and Their Applications,* 39**,** 153-180.

NGANDU, N. H. 1997. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Statistics in Medicine,* 16**,** 611-626.

NICE 2013. NICE diagnostics guidance [DG10]. London: National Institute for Health and Care Excellence.

NICE. 2015. *National Institute for Clinical Excellence Guideline. Breast cancer (early and locally advanced) (CG80).* [Online]. Available: http://www.nice.org.uk/CG80 [Accessed 20 August 2015].

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 Receptor-Positive Breast Cancer. *Clinical Cancer Research,* 16**,** 5222-5232.

NYGARD, J. F., SAUER, T., NYGARD, M., SKARE, G. B. & THORESEN, S. O. 2004. CIN 2/3 and cervical cancer in an organised screening programme after an unsatisfactory or a normal Pap smear: a seven-year prospective study of the Norwegian population-based screening programme. *Journal of Medical Screening,* 11.

OLIVOTTO, I. A., BAJDIK, C. D., RAVDIN, P. M., SPEERS, C. H., COLDMAN, A. J., NORRIS, B. D., DAVIS, G. J., CHIA, S. K. & GELMON, K. A. 2005. Population-based validation of the prognostic model ADJUVANT! for early breast cancer. *Journal of Clinical Oncology,* 23**,** 2716-2725.

OSULLIVAN, F. 1988. NONPARAMETRIC-ESTIMATION OF RELATIVE RISK USING SPLINES AND CROSS-VALIDATION. *Siam Journal on Scientific and Statistical Computing,* 9**,** 531-542.

PAIK, S., SHAK, S., TANG, G., KIM, C., BAKER, J., CRONIN, M., BAEHNER, F. L., WALKER, M. G., WATSON, D., PARK, T., HILLER, W., FISHER, E. R., WICKERHAM, D. L., BRYANT, J. & WOLMARK, N. 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine,* 351**,** 2817-2826.

PAIK, S., TANG, G., SHAK, S., KIM, C., BAKER, J., KIM, W., CRONIN, M., BAEHNER, F. L., WATSON, D., BRYANT, J., COSTANTINO, J. P., GEYER, C. E., JR., WICKERHAM, D. L. & WOLMARK, N. 2006. Gene expression and benefit of chemotherapy in women with

node-negative, estrogen receptor-positive breast cancer. *Journal of Clinical Oncology,* 24**,** 3726-3734.

PARKER, J. S., MULLINS, M., CHEANG, M. C. U., LEUNG, S., VODUC, D., VICKERY, T., DAVIES, S., FAURON, C., HE, X., HU, Z., QUACKENBUSH, J. F., STIJLEMAN, I. J., PALAZZO, J., MARRON, J. S., NOBEL, A. B., MARDIS, E., NIELSEN, T. O., ELLIS, M. J., PEROU, C. M. & BERNARD, P. S. 2009. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology,* 27**,** 1160-1167.

PENG, Y. W. 2003. Fitting semiparametric cure models. *Computational Statistics & Data Analysis,* 41**,** 481-490.

PENG, Y. W., DEAR, K. B. G. & DENHAM, J. W. 1998. A generalized F mixture model for cure rate estimation. *Statistics in Medicine,* 17.

PENNELL, M. L., WHITMORE, G. A. & TING LEE, M. L. 2010. Bayesian random-effects threshold regression with application to survival data with nonproportional hazards. *Biostatistics,* 11 (1)**,** 111-126.

PERPEROGLOU, A., KERAMOPOULLOS, A. & VAN HOUWELINGEN, H. C. 2007. Approaches in modelling long-term survival: an application to breast cancer. *Statistics in Medicine,* 26**,** 2666-85.

PERPEROGLOU, A., LE CESSIE, S. & VAN HOUWELINGEN, H. C. 2006a. Reduced-rank hazard regression for modelling non-proportional hazards. *Statistics in Medicine,* 25**,** 2831-2845.

PERPEROGLOU, A., VAN HOUWELINGEN, H. C. & HENDERSON, R. 2006b. A relaxation of the gamma frailty (Burr) model. *Statistics in Medicine,* 25**,** 4253-66.

PETTITT, A. N. & DAUD, I. B. 1990. Investigating Time Dependence in Cox's Proportional Hazards Model. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 39**,** 313-329.

PULTZ, B. D. A., CORDERO DA LUZ, F. A., DE FARIA, P. R., LIMA OLIVEIRA, A. P., DE ARAUJO, R. A. & BARBOSA SILVA, M. J. 2014. Far Beyond the Usual Biomarkers in Breast Cancer: A Review. *Journal of Cancer,* 5**,** 559-571.

QUANTIN, C., ABRAHAMOWICZ, M., MOREAU, T., BARTLETT, G., MACKENZIE, T., TAZI, M. A., LALONDE, L. & FAIVRE, J. 1999. *Variation over time of the effects of prognostic factors in a population-based study of colon cancer: Comparison of statistical models*, American Journal of Epidemiology. 150 (11) (pp 1188-1200), 1999. Date of Publication: 01 Dec 1999.

RAKHA, E., SORIA, D., LEMETRE, C., GREEN, A., POWE, D., NOLAN, C., GARIBALDI, J., BALL, G. & ELLIS, I. 2012. Nottingham Prognostic Index Plus: a modern clinical decision making tool in breast cancer.

RAMOS-VARA, J. A. & MILLER, M. A. 2014. When Tissue Antigens and Antibodies Get Along: Revisiting the Technical Aspects of Immunohistochemistry-The Red, Brown, and Blue Technique. *Veterinary Pathology,* 51**,** 42-87.

RAVDIN, P. M., GREEN, S., DORR, T. M., MCGUIRE, W. L., FABIAN, C., PUGH, R. P., CARTER, R. D., RIVKIN, S. E., BORST, J. R., BELT, R. J., METCH, B. & OSBORNE, C. K. 1992. PROGNOSTIC-SIGNIFICANCE OF PROGESTERONE-RECEPTOR LEVELS IN ESTROGEN RECEPTOR-POSITIVE PATIENTS WITH METASTATIC BREAST-CANCER TREATED WITH TAMOXIFEN - RESULTS OF A PROSPECTIVE SOUTHWEST-ONCOLOGY-GROUP STUDY. *Journal of Clinical Oncology,* 10**,** 1284-1291.

RAVDIN, P. M., SIMINOFF, L. A., DAVIS, G. J., MERCER, M. B., HEWLETT, J., GERSON, N. & PARKER, H. L. 2001. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of Clinical Oncology,* 19**,** 980-991.

RICHANI, K., ROMERO, R., KIM, Y. M., CUSHENBERRY, E., SOTO, E., HAN, Y. M., ESPINOZA, J. & KIM, C. J. 2006. Tissue microarray: An effective high-throughput method to study the placenta for clinical and research purposes. *Journal of Maternal-Fetal & Neonatal Medicine,* 19**,** 509-515.

RING, B. Z., SEITZ, R. S., BECK, R., SHASTEEN, W. J., TARR, S. M., CHEANG, M. C. U., YODER, B. J., BUDD, G. T., NIELSEN, T. O., HICKS, D. G., ESTOPINAL, N. C. & ROSS, D. T. 2006. Novel prognostic immunohistochemical biomarker panel for estrogen receptor-positive breast cancer. *Journal of Clinical Oncology,* 24**,** 3039-3047.

RODRIGUEZ, G. 2005. *Non-parametric estimation in survival models* [Online]. Available: http://data.princeton.edu/pop509/NonParametricSurvival.pdf [Accessed].

ROMOND, E. H., PEREZ, E. A., BRYANT, J., SUMAN, V. J., GEYER, C. E., DAVIDSON, N. E., TAN-CHIU, E., MARTINO, S., PAIK, S., KAUFMAN, P. A., SWAIN, S. M., PISANSKY, T. M., FEHRENBACHER, L., KUTTEH, L. A., VOGEL, V. G., VISSCHER, D. W., YOTHERS, G., JENKINS, R. B., BROWN, A. M., DAKHIL, S. R., MAMOUNAS, E. P., LINGLE, W. L., KLEIN, P. M., INGLE, J. N. & WOLMARK, N. 2005. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *New England Journal of Medicine,* 353**,** 1673-1684.

ROSS, D. T., KIM, C.-Y., TANG, G., BOHN, O. L., BECK, R. A., RING, B. Z., SEITZ, R. S., PAIK, S., COSTANTINO, J. P. & WOLMARK, N. 2008. Chemosensitivity and Stratification by a Five Monoclonal Antibody Immunohistochemistry Test in the NSABP B14 and B20 Trials. *Clinical Cancer Research,* 14**,** 6602-6609.

ROYSTON, P. & ALTMAN, D. G. 1994. REGRESSION USING FRACTIONAL POLYNOMIALS OF CONTINUOUS COVARIATES - PARSIMONIOUS PARAMETRIC MODELING. *Applied Statistics-Journal of the Royal Statistical Society Series C,* 43**,** 429-467.

ROYSTON, P. & ALTMAN, D. G. 2013. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology,* 13.

ROYSTON, P. & LAMBERT, P. C. (eds.) 2011. *Flexible parametric survival analysis using stata: beyond the Cox model.*: Stata Press.

ROYSTON, P. & PARMAR, M. K. B. 2002a. 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**,** 2175-97.

ROYSTON, P. & PARMAR, M. K. B. 2002b. 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**,** 2175-2197.

ROYSTON, P. & PARMAR, M. K. B. 2011. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine,* 30**,** 2409-2421.

ROYSTON, P. & PARMAR, M. K. B. 2014. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials,* 15.

ROYSTON, P., PARMAR, M. K. B. & ALTMAN, D. G. 2010. External validation and updating of a prognostic survival model.

ROYSTON, P. & SAUERBREI, W. 2004. A new measure of prognostic separation in survival data. *Statistics in Medicine,* 23**,** 723-748.

ROYSTON, P. & WHITE, I. R. 2011. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software,* 45**,** 1-20.

RUBIN, D. 1987. *Multiple imputation for nonresponse in surveys*, Wiley.

RYAN, T. P. & WOODALL, W. H. 2005. The most-cited statistical papers. *Journal of Applied Statistics,* 32**,** 461-474.

SARGENT, D. J. 1997. A flexible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Analysis,* 3 (1)**,** 13-25.

SARGENT, D. J., CONLEY, B. A., ALLEGRO, C. & COLLETTE, L. 2005. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology,* 23**,** 2020-2027.

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**,** 71-94.

SAUERBREI, W., ROYSTON, P. & LOOK, M. 2007. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal,* 49**,** 453-473.

SCHEIKE, T. H. & ZHANG, M. J. 2002. An additive-multiplicative Cox-Aalen regression model. *Scandinavian Journal of Statistics,* 29**,** 75-88.

SCHEMPER, M. 1992. COX ANALYSIS OF SURVIVAL-DATA WITH NONPROPORTIONAL HAZARD FUNCTIONS. *Statistician,* 41**,** 455-465.

SCHEMPER, M., WAKOUNIG, S. & HEINZE, G. 2009. The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine,* 28**,** 2473-89.

SCHOENFELD, D. 1982. PARTIAL RESIDUALS FOR THE PROPORTIONAL HAZARDS REGRESSION-MODEL. *Biometrika,* 69**,** 239-241.

SCHWARZ, G. 1978. ESTIMATING DIMENSION OF A MODEL. *Annals of Statistics,* 6**,** 461-464.

SEER. 2010. *Surveillance, Epidemiology, and End Results (SEER) Program: SEER*Stat Database: Incidence-SEER 13 Regs Research Data, November 2009 Submission (1992-2007) Katrinia/Rita Population Adjustment—Linked To County Attributes. Total U.S., 1969-2007 Counties, National Cancer Institute, Division of Cancer Control and Population Sciences, Surveillance Research Program, Cancer Statistics Branch, released April 2010, based on November 2009 submission.* [Online]. Available: http://seer.cancer.gov/. [Accessed].

SESTAK, I., DOWSETT, M., ZABAGLO, L., LOPEZ-KNOWLES, E., FERREE, S., COWENS, J. W. & CUZICK, J. 2013. Factors Predicting Late Recurrence for Estrogen Receptor-Positive Breast Cancer. *Jnci-Journal of the National Cancer Institute,* 105**,** 1504-1511.

SGROI, D. C., SESTAK, I., CUZICK, J., ZHANG, Y., SCHNABEL, C. A., SCHROEDER, B., ERLANDER, M. G., DUNBIER, A., SIDHU, K., LOPEZ-KNOWLES, E., GOSS, P. E. & DOWSETT, M. 2013. Prediction of late distant recurrence in patients with oestrogen-receptor-positive breast cancer: a prospective comparison of the breast-cancer index (BCI) assay, 21-gene recurrence score, and IHC4 in the TransATAC study population. *Lancet Oncology,* 14**,** 1067-1076.

SKACEL, M., SKILTON, B., PETTAY, J. D. & TUBBS, R. R. 2002. Tissue microarrays: A powerful tool for high-throughput analysis of clinical specimens - A review of the method with validation data. *Applied Immunohistochemistry & Molecular Morphology,* 10**,** 1-6.

SLAMON, D. J., CLARK, G. M., WONG, S. G., LEVIN, W. J., ULLRICH, A. & MCGUIRE, W. L. 1987. HUMAN-BREAST CANCER - CORRELATION OF RELAPSE AND SURVIVAL WITH AMPLIFICATION OF THE HER-2 NEU ONCOGENE. *Science,* 235**,** 177-182.

STABLEIN, D. M., CARTER, W. H., JR. & NOVAK, J. W. 1981. Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials,* 2**,** 149-59.

STATACORP 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.

STERNE, J. A. C., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. & CARPENTER, J. R. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal,* 339.

SUNDQUIST, M., THORSTENSON, S., BRUDIN, L., NORDENSKJOLD, B. & GRP, S. E. S. B. C. S. 1999. Applying the Nottingham Prognostic Index to a Swedish breast cancer population. *Breast Cancer Research and Treatment,* 53**,** 1-8.

TANG, G., SHAK, S., PAIK, S., ANDERSON, S. J., COSTANTINO, J. P., GEYER, C. E., JR., MAMOUNAS, E. P., WICKERHAM, D. L. & WOLMARK, N. 2011. Comparison of the prognostic and predictive utilities of the 21-gene Recurrence Score assay and Adjuvant! for women with node-negative, ER-positive breast cancer: results from NSABP B-14 and NSABP B-20. *Breast Cancer Research and Treatment,* 127**,** 133-142.

THERNEAU, T. M., GRAMBSCH, P. M. & FLEMING, T. R. 1990. MARTINGALE-BASED RESIDUALS FOR SURVIVAL MODELS. *Biometrika,* 77**,** 147-160.

THOMAS, J. S. J., KERR, G. R., JACK, W. J. L., CAMPBELL, F., MCKAY, L., PEDERSEN, H.-C., KUNKLER, I. H., CAMERON, D. A., CHETTY, U. & BARTLETT, J. M. S. 2009. Histological

grading of invasive breast carcinoma - a simplification of existing methods in a large conservation series with long-term follow-up. *Histopathology,* 55**,** 724-731.

TIAN, L., ZUCKER, D. & WEI, L. J. 2005. On the cox model with time-varying regression coefficients. *Journal of the American Statistical Association,* 100**,** 172-183.

TODD, J. H., DOWLE, C., WILLIAMS, M. R., ELSTON, C. W., ELLIS, I. O., HINTON, C. P., BLAMEY, R. W. & HAYBITTLE, J. L. 1987. CONFIRMATION OF A PROGNOSTIC INDEX IN PRIMARY BREAST-CANCER. *British Journal of Cancer,* 56**,** 489-492.

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**,** 2763-2780.

VAN'T VEER, L. J., DAI, H. Y., VAN DE VIJVER, M. J., HE, Y. D. D., HART, A. A. M., MAO, M., PETERSE, H. L., VAN DER KOOY, K., MARTON, M. J., WITTEVEEN, A. T., SCHREIBER, G. J., KERKHOVEN, R. M., ROBERTS, C., LINSLEY, P. S., BERNARDS, R. & FRIEND, S. H. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature,* 415**,** 530-536.

VAN DE VELDE, C. J. H., REA, D., SEYNAEVE, C., PUTTER, H., HASENBURG, A., VANNETZEL, J.-M., PARIDAENS, R., MARKOPOULOS, C., HOZUMI, Y., HILLE, E. T. M., KIEBACK, D. G., ASMAR, L., SMEETS, J., NORTIER, J. W. R., HADJI, P., BARTLETT, J. M. S. & JONES, S. E. 2011. Adjuvant tamoxifen and exemestane in early breast cancer (TEAM): a randomised phase 3 trial. *Lancet,* 377**,** 321-331.

VAN DE VIJVER, M. J., HE, Y. D., VAN 'T VEER, L. J., DAI, H., HART, A. A. M., VOSKUIL, D. W., SCHREIBER, G. J., PETERSE, J. L., ROBERTS, C., MARTON, M. J., PARRISH, M., ATSMA, D., WITTEVEEN, A., GLAS, A., DELAHAYE, L., VAN DER VELDE, T., BARTELINK, H., RODENHUIS, S., RUTGERS, E. T., FRIEND, S. H. & BERNARDS, R. 2002. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine,* 347**,** 1999-2009.

VERGOUWE, Y., ROYSTON, P., MOONS, K. G. M. & ALTMAN, D. G. 2010. Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of Clinical Epidemiology,* 63**,** 205-214.

VERWEIJ, P. J. & VAN HOUWELINGEN, H. C. 1995. Time-dependent effects of fixed covariates in Cox regression. *Biometrics,* 51**,** 1550-6.

WALKER, R. A., BARTLETT, J. M. S., DOWSETT, M., ELLIS, I. O., HANBY, A. M., JASANI, B., MILLER, K. & PINDER, S. E. 2008. HER2 testing in the UK: further update to recommendations. *Journal of Clinical Pathology,* 61**,** 818-824.

WEI, G. & SCHAUBEL, D. E. 2008. Estimating cumulative treatment effects in the presence of nonproportional hazards. *Biometrics,* 64**,** 724-32.

WEIGEL, M. T. & DOWSETT, M. 2010. Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocrine-Related Cancer,* 17**,** R245-R262.

WHITE, I. R. & ROYSTON, P. 2009. Imputing missing covariate values for the Cox model. *Statistics in Medicine,* 28**,** 1982-1998.

WHITE, I. R., ROYSTON, P. & WOOD, A. M. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine,* 30**,** 377-399.

XU, R. H. & ADAK, S. 2002. Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics,* 58**,** 305-315.

YEO, B. J., DOWSETT, M., SMITH, I. E. & ZABAGLO, L. 2014. THE CLINICAL IMPACT OF USING THE IHC4 SCORE: OUR MDT EXPERIENCE IN A PROSPECTIVE SERIES OF POSTMENOPAUSAL WOMEN WITH ER POSITIVE EARLY BREAST CANCER. *Annals of Oncology,* 25**,** 2-2.

ZHANG, Y., SCHNABEL, C. A., SCHROEDER, B. E., JEREVALL, P.-L., JANKOWITZ, R. C., FORNANDER, T., STAL, O., BRUFSKY, A. M., SGROI, D. & ERLANDER, M. G. 2013. Breast Cancer Index Identifies Early-Stage Estrogen Receptor-Positive Breast Cancer Patients at Risk for Early- and Late-Distant Recurrence. *Clinical Cancer Research,* 19**,** 4196-4205.

ZUCKER, D. M. & KARR, A. F. 1990. NONPARAMETRIC SURVIVAL ANALYSIS WITH TIME-DEPENDENT COVARIATE EFFECTS - A PENALIZED PARTIAL LIKELIHOOD APPROACH. *Annals of Statistics,* 18**,** 329-353.

ZUJEWSKI, J. A. & KAMIN, L. 2008. Trial Assessing Individualized Options for Treatment for breast cancer: the TAILORx trial. *Future Oncology,* 4**,** 603-610.