Dynamic Predictive Joint Models to Characterise Localised Prostate Cancer Prognosis after Radiotherapy

March 2023

Harry George Parr

The Division of Clinical Studies

Clinical Trials & Statistics Unit

The Institute of Cancer Research

University of London, United Kingdom

A thesis submitted to The University of London for the degree of

Doctor of Philosophy

Declaration

I, Harry George Parr, confirm that the work presented in this thesis has been performed by me unless otherwise stated in the relevant sections.

Harry George Parr

Abstract

This thesis aims to develop and validate dynamic predictive joint models (JMs) to characterise the prognosis of patients with localised prostate cancer who are treated with moderately hypofractionated radiotherapy with neoadjuvant and concurrent hormone therapy.

Current clinical prediction models rely on baseline features, e.g. tumour severity and age at diagnosis, which do not adequately predict cancer recurrence. This thesis proposes using routinely collected longitudinal prostate-specific antigen (PSA) measurements, in addition to baseline prognostic factors, to obtain more accurate and dynamically updated predictions.

This thesis uses CHHiP, the largest known moderately hypofractionated phase-III trial for localised prostate cancer, to develop a mixed-effects submodel for longitudinal PSAs and a relative risk submodel for time-to-recurrence. The dynamics of PSA trajectories, including concentration and rate-of-change, are considered. Predictions are compared across patient subgroups with contrasting prognostic factors, and PSA thresholds are explored to correlate with prognosis. The performance of the JM is validated using bias-corrected bootstraps and on external cohorts to assess its utility and generalisability. The model is extended to account for the competing risk of deaths unrelated to prostate cancer.

This study finds that patients who developed recurrence generally had higher baseline and overall PSA values during follow-up and experienced an exponentially rising PSA in the two years before recurrence. Most baseline risk factors were significant in both submodels, and PSA value and rate-of-change were predictive of future recurrence. PSA thresholds $\lessapprox$0.23ng/mL after treatment correlated with good prognosis. The model's predictive performance was good across differing external cohorts and prediction times.

Overall, this thesis demonstrates that dynamically updated PSA information can improve prognostication, which can be used to guide follow-up and treatment management options. It provides evidence for the potential use of JMs in clinical practice, for instance, instigating PSA-driven imaging in high-risk patients and recommending fewer PSA collections for low-risk patients.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, firstly a sincere thank you to my primary supervisor Prof Emma Hall whose diligence, knowledge, pragmatism, and commitment to improving patients' lives knows no bounds. It's been an absolute privilege to carry out this work.

I also wish to sincerely thank Dr Nuria Porta, my associate supervisor who's been a huge support (and grief counsellor) from the very start, and whom this PhD wouldn't have been possible without her. Thank you for your tremendous support, time, and guidance throughout my PhD journey.

I want to sincerely thank and pay tribute to my secondary and clinical supervisors, Prof Judith Bliss & Dr Alison Tree; outside of my supervisory team: 'The Prof' David Dearnaley, for all their expertise, knowledge, patience (and patients!), and for all the fruitful discussions, which have been invaluable in shaping this research and academic development; for the lasting gift that CHHiP keeps on giving well over two decades after first recruiting. I am hugely grateful to the patients and all those who participated and facilitated the CHHiP trial. I also wish to acknowledge and thank the trial teams for RT01 (Prof Matt Sydes & Dr Victoria Yorke-Edwards), and RADAR (Prof Jim Denham & Allison Steigler) who provided their trial data to enable valuable validation of the developed models in this thesis.

I am continually in awe of the multidisciplinary unit that is the ICR-CTSU, for its dedication in benefiting patients and their families, with their ethos of shorter and kinder treatments. A few people I'd like to mention from the unit in particular are Clare Griffin – for answering my many many queries on CHHiP. Prof Christina Yap, Drs Maggie Cheang, Jan Rekowski & George Seed for all the inspired discussions throughout the years, and involvement with the De Bono Lab. Also, to Holly Tovey – who is incredibly knowledgeable, helpful and a friend to me from the start, Vicki Hinder – my joiner buddy, the entire CHHiP trial team, and a special thank you to Tara Thorneycroft who has been immensely helpful to me over the years, including booking various conference trips and keeping my memberships valid!

I also wish to thank Prof Jeff Bamber as my senior tutor for his counsel, not only academically, but pastorally too. I of course wish to thank The Institute of Cancer Research for choosing to fund this PhD.

I want to thank some of my oldest and closest friends who read extracts and provided invaluable feedback. Particularly to Craig Carruthers who read an earlier draft of the entire thesis for his critical review and advice on the prose and sentence structure. Drs Adele Crapnell and Beshara Sheehan for their clinical input and scientific insight, as well as morale support over the years.

Finally, to my parents, Anni & Nick for all their support not only during the PhD but throughout my upbringing, who very sadly separated recently; my gran, and finally my partner Alex whose love and encouragement knows no bounds. I hope to have made you proud.

Publications & Talks

Publications

A Personalised Clinical Dynamic Prediction Model to Characterise Prognosis for Patients with Localised Prostate Cancer: analysis of the CHHiP Phase III Trial.

Harry Parr, Nuria Porta, Alison C Tree, David Dearnaley, Emma Hall, *International Journal of Radiation Oncology - Biology - Physics* (2023), https://doi.org/10.1016/j.ijrobp.2023.02.022

Joint models for dynamic prediction in localised prostate cancer: a literature review.

Harry Parr, Emma Hall, Nuria Porta. *BMC Medical Research Methodology* volume 22, Article number: 245 (2022) https://doi.org/10.1186/s12874-022-01709-3

Talks

Aug 2022 – Externally Validating Clinical Dynamic Prediction Joint Models for Localised Prostate Cancer, The International Society for Clinical Biostatistics (ISCB), Newcastle, UK

May 2022 - Developing and Externally Validating Clinical Dynamic Prediction Joint Models for Localised Prostate Cancer, Society for Clinical Trials (SCT), San Diego, US.

Jun 2021 - Developing & Validating Dynamic Prediction Joint Models to Predict Prognosis in Prostate Cancer Patients' Radiotherapy, PhD Student Presentations, ICR Conference.

May 2021 - Developing and Validating Clinical Dynamic Prediction Joint Models to Predict Prognosis in Prostate Cancer Patients, Society for Clinical Trials (SCT, virtual).

Oct 2020 - Proposing and Developing Clinical Dynamic Prediction Models to Predict Prognosis for Prostate Cancer Patients, Early Career Researchers' Health Data Science Symposium, Plymouth/Virtually (invited).

Aug 2020 - Dynamic Prediction Modelling with Applications in Localised Prostate Cancer to Characterise Prognosis: Extending to a Competing Risk Framework, The International Society for Clinical Biostatistics (ISCB), Kraków, Poland/Virtually.

Feb 2020 - Developing Clinical Dynamic Prediction Models to Characterise the Prognosis of Post-radiotherapy Localised Prostate Cancer Patients, ICR Student Conference, London, UK

Aug 2019 - Mixed-effects Models to Characterise Prostate-specific Antigen (PSA) Dynamics Post-Radiotherapy in the CHHiP Randomised Clinical Trial, Young Statisticians' Meeting (YSM), Leeds, UK.

Contents

Declaration **2**

Abstract **3**

Acknowledgements **4**

Publications & Talks **6**

Publications 6

Talks 7

Contents **8**

List of Figures **12**

List of Tables **17**

List of Abbreviations **20**

1 Chapter 1 – Introduction **22**

1.1 Diagnosis & treatment of prostate cancer 22

1.2 Motivation 26

1.3 The CHHiP Trial 29

1.4 Developing a clinical prediction model 30

1.5 Thesis purpose and objectives 32

1.6 Thesis description & outline of subsequent chapters 33

2 Chapter 2 - Joint Modelling Methodology **35**

2.1. Publications relating to this chapter 35

2.2. Introduction 35

2.2.1. Notation 36

2.3. Joint modelling specification and estimation 37

2.3.1. Shared-parameter joint model 38

2.3.2. Joint latent class model 42
2.4. Model comparisons 43
2.5. Dynamic predictions 44
2.6. Evaluating predictive performance 46
2.7. Discussion 49
3 Chapter 3 – Literature Review 50
3.1 Publications related to this chapter 50
3.2 Introduction 50
3.3 Literature Search 50
3.4 Shared-parameter joint models to predict recurrence in localised prostate cancer 59
3.4.1 Model specification 59
3.4.2 Estimation, prediction and validation 60
3.5 Latent class joint models to predict recurrence in localised prostate cancer 62
3.5.1 Comparison between latent-class and shared-parameter joint models 63
3.6 Extensions to the shared-parameter joint model 63
3.6.1 Joint-Cure models 63
3.6.2 Competing risks joint models 65
3.6.3 Multi-state joint models 66
3.7 Discussion 67
4 Chapter 4 – Development of a Personalised Clinical Dynamic Predictive Joint Model to Characterise Prognosis for Patients with Localised Prostate Cancer Patients: Analysis of the CHHiP Phase III Trial 74
4.1 Publications and presentations related to this chapter 74
4.2 Introduction 74
4.3 Methods & Materials 75
4.3.1 Study design & procedure 75

4.3.2 Outcomes 76
4.3.3 Specification of the joint model 76
4.3.4 Estimation 79
4.3.5 Dynamic predictions 80
4.3.6 Assessing predictive performance and risk thresholds 80
4.4 Results 81
4.4.1 Dataset for model building 81
4.4.2 Modelling of PSA trajectories 82
4.4.3 Joint modelling time-to-recurrence 87
4.4.4 Dynamic predictions 89
4.4.5 Assessing predictive performance 92
4.4.6 PSA risk thresholds 93
4.5 Discussion & conclusions 95
Chapter 5 – External Validation of the Clinical Dynamic Predictive Joint Model for Recurrence in Localised and Locally Advanced Prostate Cancer 104
5.1 Introduction 104
5.2 Methods & Materials 105
5.2.1 External cohorts 105
5.2.2 Outcomes 106
5.2.3 Statistical analysis 107
5.2.4 Multiple imputation for missing Gleason levels 108
5.3 Results 109
5.3.1 Baseline characteristics comparison 109
5.3.2 Follow-up characteristics comparison 111
5.3.3 Predictive performance comparison 113
5.3.4 Gleason imputation sensitivity analysis for RT01 120

5.4 Discussion 121

Chapter 6 – Competing Risks Joint Models 127

6.1 Introduction 127

6.2 Methodology 128

6.2.1 Competing risks joint model 129

6.2.2 Assessing predictive performance 129

6.3 Results 131

6.3.1 Cumulative incidences 131

6.3.2 Competing risk joint model parameter estimates 132

6.3.3 Predictive performance 135

6.4 Discussion 136

Chapter 7 – Concluding remarks 141

7.1. Summary 141

7.2. Discussion and future work 143

7.2.1. Methodology used 143

7.2.2. Specification of the model to predict prostate cancer recurrence 147

7.2.3. Implementation of prediction models in clinical practice 152

7.3 Conclusions 158

8 Bibliography 159

Appendices 183

Appendix A (chapter 4) 183

Appendix B (chapter 5) 188

Appendix C (chapter 6) 206

List of Figures

Figure 1-1 localised prostate cancer risk grouping and corresponding treatment management option, in accordance with NICE guidance. NCCN = National Comprehensive Cancer Network; NICE = National Institute for Health and Care Excellence. 25
Figure 1-2 trial schema of CHHiP; ctrl = control, PSA = prostate-specific antigen. 30
Figure 2-1 a graphical representation of how the random effects elicit conditional independence of the two outcomes in the shared-parameter joint model specification..... 41
Figure 3-1 a flowchart for identifying studies of the literature review. 51
Figure 4-1 top (a): aggregated PSAs and boxplots by year and outcome since starting treatment, some non-recurrent PSA flares are observed. Patient numbers still at risk are presented below the plot. Bottom (b): smoothed reverse-year PSA trajectory plot, stratified by outcome, natural cubic spline smoothers shown. In the nonrecurring patients, a few PSAs >5 ng/mL are recorded; these PSAs were considered bounces/flares and therefore did not achieve the protocol's definition of biochemical failure. 83
Figure 4-2 the predicted effect plots of PSA, stratified by outcome (solid – recurrence, dashed – censored) and each baseline subgroup over time. The top-left panel depicts the overall average PSA trajectories for each outcome. The natural cubic spline smoother is depicted... 85
Figure 4-3 traceplots of the four α log-hazard ratio parameter chains, for the value and slope association structure for PSA. 88
Figure 4-4 dynamic predictions of two patients: A & B, over five panels (V—Z). Patients A & B are ages 63 and 64 respectively and both received the same treatments, with contrasting prognostic factors. The left-hand side of each plot shows their modelled PSA values over time and the right-hand side shows their cumulative risk of recurrence at particular landmarks by ten years after initiating treatment. The 95% credible intervals are shown (shaded). 92
Figure 4-5 scatter plots of PSA predicting prognosis/recurrence risk by 8 years (horizon), each panel represents landmarks 0 – 5 years. Each grey dot indicates a patient's PSA (nearest to that landmark time) and risk at each landmark time. PSAs ≤ 3ng/mL are considered after t=1. The blue line indicates regression fit with the corresponding equation and R^2 labelled in each panel, with 95% confidence intervals. The wider grey bands indicate 95% prediction

intervals. At the intercept (or less) indicates the predicted recurrence risk for a nil PSA; for the regression lines at t=3,4,5, each PSA threshold is labelled that predicts a <5% risk. 94
Figure 5-1 comparison of the trial treatments and follow-up schemas for CHHiP, RT01 & RADAR. 106
Figure 5-2 Kaplan-Meier curves comparing recurrence free survival in CHHiP (development cohorts), with RADAR & RT01 (validation cohorts). 111
Figure 5-3 averaged logged-PSA trajectories for the three studies with RADAR stratified by hormone duration: (6 or 18 months), over follow-up, separated by outcome, top – no recurrence, bottom – recurrence. Lowess smoothers are depicted. 112
Figure 5-4 comparing the internal (CHHiP – red, 50 bootstrapped samples) and external validation cohort performance metrics (RADAR – blue stratified by hormone treatment duration: 6 & 18 months; RT01 – green). The AUC (top panel) assess discrimination, i.e., to distinguish between patients who do and do not have recurrence of their cancer, based on their accrued PSA. A higher AUC values indicate improved discrimination. The bottom panel assesses overall prognostic performance through the Brier score loss function (lower values are better). These are based on patient follow-up landmarks from zero to seven years, to predict recurrence by a horizon time of eight years, πi8t = 0, …,7. AUC = area under the receiver operating characteristic curve 113
Figure 5-5 comparing the predictive performance for u = 8t = 0, …,7 between RADAR (6-month hormonal therapy schedule) and CHHiP, for the subgroup of T-stage=2 and Gleason score = 7 patients in both cohorts. AUC= area under the receiver operating characteristic curve, ICI=integrated calibration index. 115
Figure 5-6 comparing prediction windows of the recalibrated index of the two external cohorts: RADAR & RT01. 116
Figure 5-7 visually assessing calibration-in-the-large via graphical calibration plots of the RADAR cohort, before (1st & 3rd rows, orange) and after recalibration (2nd & 4th rows, green), for a fixed horizon of eight years at landmark zero to three years (top panel) and four to seven years (bottom panel). 118
Figure 5-8 visually assessing calibration-in-the-large via graphical calibration plots of the RT01 cohort, before (1st & 3rd rows, orange) and after recalibration (2nd & 4th rows, green), for

a fixed horizon of eight years at landmark zero to three years (top panel) and four to seven years (bottom panel). 119
Figure 5-9 predictions assessed using AUC, Brier, and recalibrated ICI, of RT01 Gleason scoring imputed levels via MICE, comparing to Gleason scores of 3+4=7 or 4+3=7. AUC = area under the receiver operating characteristic curve; ICI = integrated calibration index; MICE = multiple imputation by chained equations. 120
Figure 6-1 a graphical representation of a competing risk model with two causes (K=2): recurrence of prostate cancer (k=1), or death unrelated to prostate cancer (k=2). 128
Figure 6-2 compares the 1-KM and cumulative incidence function estimators of each outcome (recurrence or death unrelated to prostate cancer). 132
Figure 6-3 compares the IPCW and M-B methods applied to both the StdJM and CRJM in assessing each of its apparent validation metrics for the AUC (top) and the Brier scores (bottom) at each landmark time to predict up to a horizon time of eight years. AUC = area under the receiver operating characteristic curve; IPCW = inverse probability of censored weighting; M-B = model-based; CRJM = competing risk joint model; StdJM = standard joint model (chapter 4). 135

Supplementary Figure A1 Quantile-Quantile plots of the residuals from the longitudinal joint model (top panel) and random effects (bottom panel). 1844
Supplementary Figure B1 Kaplan-Meier (similar to **Figure 5-2**) with a breakdown comparing outcomes of the RADAR 6- and 18-month hormone schedules. 1899
Supplementary Figure B2 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RADAR cohort, before recalibration, for a fixed prediction window of two years. 193
Supplementary Figure B3 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RADAR cohort, after recalibration, for a fixed prediction window of two years 194
Supplementary Figure B4 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RT01 cohort, before recalibration, for a fixed prediction window of two years. 1955

Supplementary Figure B5 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RT01 cohort, after recalibration, for a fixed prediction window of two years. 1966
Supplementary Figure B6 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RADAR cohort, before recalibration, for a fixed prediction window of five years. 1977
Supplementary Figure B7 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RADAR cohort, after recalibration, for a fixed prediction window of five years. 1988
Supplementary Figure B8 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RT01 cohort, before recalibration, for a fixed prediction window of five years. 199
Supplementary Figure B9 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RT01 cohort, after recalibration, for a fixed prediction window of five years. 200
Supplementary Figure B10 workflow to calculate the required minimum external sample size to precisely estimate the calibration slope at a particular time point. Replicated from [190]. 202
Supplementary Figure B11 histogram of the linear predictor of the Cox submodel for CHHiP. 2033
Supplementary Figure B12 calibration plot of the Cox survival submodel for CHHiP at a horizon time of 8 years. 2044
Supplementary Figure B13 simulated calibration curves for a sample of 20,000 patients and a target standard error of 0.1. The grey curves show the estimated and simulated calibration, and the black 45-degree line indicates perfect calibration. 2055
Supplementary Figure C1 competing risk dynamic predictions for two patients from the competing risk joint model. Blue on the left-hand side indicates the PSA readings (dots), predicted PSAs (curve), on the right-hand side the cumulative incidence functions of each competing cause: recurrence (green) and death unrelated to prostate cancer (red), with the corresponding shaded 95% credible intervals. The vertical dotted line indicates the landmark

time for each patient and their accrued PSAs up to that time point, to then make predictions for the prediction window with horizon time up to ten years. 210

List of Tables

Table 1-1 risk stratification by clinical risk factors: Clinical T-stage, Gleason score, and presenting PSA. Locally advance prostate cancer includes high-risk localised patients, as defined by the National Comprehensive Cancer Network (NCCN). 24
Table 1-2 risk stratification by clinical risk factors: Clinical T-stage, Gleason score, and presenting PSA. Locally non-metastatic prostate cancer includes high-risk localised patients, as defined by Cambridge Prognostic Group (CPG) classification. 26
Table 3-1 a summary table of the articles where joint modelling was applied to localised prostate cancer to predict recurrence, in chronological order. Abbreviations include: JLCM = joint latent class model, SPJM = shared-parameter joint model, PSA= prostate-specific antigen, EBRT = external-beam radiotherapy, HT = hormone therapy, Gy = Gray, EPOCE = expected prognostic observed cross-entropy, CVPOL = cross-validated prognostic observed log-likelihood, IBS = integrated Brier score. 53
Table 4-1 baseline characteristics of CHHiP patients (N=3071) considered in model development, stratified by outcome. LHRHa – Luteinizing-Hormone-Releasing-Hormone analogue + possible anti-androgen; [1] n (%); Median (IQR). 82
Table 4-2 knot selection procedure fitted with maximum likelihood mixed-effect models for each internal knot for the natural cubic splines in the fixed and random effects, whilst adjusting for baseline covariates (treatment received, T-stage, Gleason, age). The LRT & p-value is comparing to the row above it. df=degrees of freedom, AIC=Akaike's information criterion, BIC=Bayesian information criterion, MSE=mean squared error, CV=cross validation, LRT=likelihood-ratio test. 84
Table 4-3 fixed effect model parameters from the joint mixed-effect submodel. HT = hormone therapy. 86
Table 4-4 the estimated symmetric variance-covariance matrix, D, of the random effect from the fitted joint model. The diagonal elements in bold indicate the standard deviations of the random effects 86
Table 4-5 comparing CHHiP hazard ratios from the standard Cox submodel and joint model. Age was median-centred (minusing 69 from all ages), * indicates 95% Bayesian credible intervals and R from the joint model. 87

Table 4-6 measuring the strength of association parameters between the two outcomes, log-hazard ratio (α) per unit increase in log(PSA) and its slope, with 95% credible intervals (CIs). 88
Table 4-7 optimism-corrected model metrics from landmark times t=0—7 predicting at a horizon time of 8 years. Discrimination – AUC (area under the curve); calibration – ICI (integrated calibration index); overall predictive performance – (Brier score). Mean, [95% bootstrapped CIs] refers to the bootstrapped replications of the posterior means. The ICI & Brier are loss functions (where lower is better), with higher AUC measures indicating better discrimination. Ns are patients remaining at risk at the development landmark. 93
Table 5-1 baseline characteristics of the development (CHHiP) and external validation cohorts (RADAR & RT01) total N=4956; * indicates imputation via MICE. MICE = multiple imputation by chained equations. 110
Table 6-1 parameter estimates of the time-to-event outcomes, comparing the competing risk joint model specification (left) to the standard joint model (right, developed in chapter 4); ref = reference level, HT = hormone therapy, LHRHa = Luteinizing-hormone-releasing-hormone analogue, PCa = prostate cancer. 134
Table 7-1 medical device classification guidance, taken from MDCG 2019-11 Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR, 2019 [261]. 153

Supplementary Table A1 TRIPOD checklist for Chapter 4. *Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D; V. TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis. 183
Supplementary Table A2 development apparent/internal prediction accuracy metrics at varying landmark times either with a fixed horizon time of t=8 years (first panel), or a fixed prediction window Dt=2 or Dt=5 years (second and third pane respectively). Time-dependent metrics: AUROC, Brier, ICI. Ns indicate the number of patients still at risk by the landmark time t. 185

Supplementary Table B1 TRIPOD checklist for Chapter 5. *Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D; V. TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis. 188

Supplementary Table B2 assessing overall calibration using the integrated calibration index (ICI) of the external cohorts before and after recalibration. The calibration metrics at landmarks $t = 0, \ldots, 7$ to predict by a fixed horizon time of 8 years (first panel), and with varying time horizons, i.e., fixed prediction intervals of two and five years are presented in the second and third panels respectively (continued next page). The negative percentage difference indicates improvement in ICI after recalibration. 1900

List of Abbreviations

Abbreviation	Description
ADT	androgen deprivation therapy
(X)AI	(explainable) artificial intelligence
AIC	Akaike information criterion
AUC	area under the curve
AUROC	area under the receiver operating characteristic curve
BIC	Bayesian information criterion
(I)BS	integrated Brier score
CDPJM	clinical dynamic predictive joint model
CHHiP	Conventional or Hypofractionated High dose intensity modulated radiotherapy in Prostate cancer
CI	confidence/credible interval
CIF	cumulative incidence function
CNN	convolution neural networks
CPG	Cambridge Prognostic Group
CPM	Clinical Prediction Model
CPO	conditional predictive ordinate
CRFs	case-report forms
CRPC	castrate resistant prostate cancer
CT	computed tomography
CVPOL	cross-validation of the prognostic observed log-likelihood
DIC	Deviance information criterion
EBRT	external beam radiotherapy
EHR	electronic health record
EPOCE	expected prognostic observed cross-entropy
f	fraction
GBM	gradient boosting machine
GGG/GS	Gleason grade grouping / Gleason score
Gy	Gray
HT	hormone therapy
ICI	integrated calibration index
ICR-CTSU	The Institute of Cancer Research - Clinical Trials & Statistics Unit
IHC	immunohistochemistry
IMRT	intensity-modulated radiotherapy
IPCW	inverse probability of censoring weighting
ISUP	International Society of Urological Pathology
JLCM	joint latent class model
LHRHa	Luteinizing Hormone-Releasing Hormone analogue
LLM	large language model
LP	linear predictor
LPML	log pseudo-marginal likelihood
MCMC	Markov chain Monte Carlo

Abbreviation	Description
MHRA	The Medicines and Healthcare products Regulatory Agency
MICE	multiple imputation by chained equations
ML	machine learning
MRI	magnetic resonance imaging
NCCN	National Comprehensive Cancer Network
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
ONS	Office for National Statistics
PCa	Prostate Cancer
PE	prediction error
PET	positron emission tomography
PI	prediction interval
PRACTICAL	Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome
PROBAST	Prediction model Risk Of Bias ASsessment Tool
PSA	prostate-specific antigen
PSMA	prostate-specific membrane antigen
PSUR	periodic safety update report
QMS	quality management system
RADAR	Randomised Androgen Deprivation And Radiotherapy
RAPPER	Radiogenomics: Assessment of Polymorphisms for Predicting the Effects of Radiotherapy
RCT	randomised control trial
RDSM	recurrent deep survival machine
RNN	recurrent neural networks
PROFIT	Prostate Fractionated Irradiation Trial
RSF	random survival forest
RTOG	Radiation Therapy Oncology Group
SBRT	stereotactic body radiotherapy
SNPs	single nucleotide polymorphisms
SPJM	shared-parameter joint model
ST	salvage therapy
TNM	Tumour, Node, Metastasis (staging)
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis
TROG	The Trans-Tasman Radiation Oncology Group
TRUS	transrectal ultrasound
UKCA	UK Conformity Assessed
wAIC	widely-applicable/Watanabe-Akaike information criterion
WAPE	weighted average absolute prediction error

Chapter 1 – Introduction

1.1 Diagnosis & treatment of prostate cancer

Prostate cancer is highly prevalent, affecting almost 1.5 million people in 2020 and is the second most common cancer diagnosis in men worldwide [1]. Prostate cancer became the most commonly diagnosed cancer, overtaking breast cancer, in the UK in 2018, with almost 50,000 registrations [2]. This was the first year where the incidence of new prostate cancers was higher than that of breast cancers, in-part due to the so-called Fry-Turnbull effect, whereby extensive media coverage of the diagnosis of two well-known UK television personalities prompted a large spike in referrals [3]. Prostate cancer survival has vastly improved in the UK within the last 50 years, with a ten-year survival rate increasing from 25% in 1971-72 to ~80% by 2013-2017. However there are still over 12,000 prostate cancer related UK deaths per year [4].

Diagnosis of prostate cancer is characterised by the stage, grade, and how widespread it is: the TNM (tumour, node, metastasis) staging. When prostate cancer is suspected, blood samples are taken for assessment of prostate-specific antigen (PSA) and the patient undergoes a digital rectal examination. PSA is a serine protease protein biomarker secreted by the prostate [5,6]. A multiparametric magnetic resonance imaging (MRI) is used for further investigation for patients with suspected localised prostate cancer, reported using a five-point Likert scale (1—5); if the MRI Likert score is ≥ 3, or the PSA density of the prostate is high (>0.12 ng/mL) a biopsy is recommended. A biopsy is required to identify the Gleason grade (a histopathological grading of the aggressiveness of the cancer), and type of prostate cancer. Biopsies were historically guided by transrectal ultrasound (TRUS); recently a trans-perineal approach is more common as it reduces the risk of rectal bleeding and infection [7]. In the UK (excluding Scotland) of all diagnoses, 10% of staging is unknown; of the known stages 51—61% are localised stages I or II (where the disease is confined to the prostate and has yet to spread to the nodes or other organs in the body), and the remainder are III or IV. Around 13% of patients present with lymph node spread and 20% of men have metastatic disease [8–11]. The spread of cancer from its primary site can be determined via imaging, using positron emission tomography (PET), MRI, computed tomography (CT) and/or bone scans.

There are several treatment options available to patients with localised prostate cancer, including radiotherapy, brachytherapy, and radical prostatectomy. In the UK, intensity-modulated radiotherapy (IMRT) is most commonly used [9]. In recent years, hypofractionated radiotherapy has been investigated, delivering fewer but larger doses of radiotherapy to the patient. Several phase-III randomised control trials (RCTs) have investigated the efficacy of hypofractionation treatment and have concluded that it is non-inferior to conventionally fractionated radiotherapy [12–14]. This resulted in changing clinical practice, whereby moderate hypofractionation (60Gy/20f), delivered with curative intent, became the gold standard of care globally. This is advantageous in reducing patient exposures and healthcare costs, with more optimal resource allocation, compared to historically longer conventional radiotherapy delivered over more fractions [15].

To further improve outcomes and disease control, neoadjuvant or adjuvant hormonal therapy is usually given in conjunction with radiotherapy [16]. A typical androgen suppression therapy is a luteinizing hormone-releasing hormone analogue (LHRHa); which is given first-line in order to reduce the serum levels of testosterone, to prevent further tumour growth, then radiotherapy is delivered to ablate tumoural tissue [13,17].

When diagnosed with prostate cancer, patients usually present with high concentrations of PSA. Routine follow-ups are carried out with patients during and after their treatment, with repeated PSA concentrations taken and recorded. During neoadjuvant or adjuvant treatment, PSA quickly drops to near-zero levels, known as the nadir (the lowest observed PSA concentration). When treatment ceases, PSA levels slowly increase as testosterone recovers and ideally return to a healthy plateau. A continued post-treatment increase in PSA suggests a regrowth of prostate cancer cells and increased risk of prostate cancer recurrence. PSA itself is used to determine biochemical failure; this is defined as a PSA concentration greater than the nadir plus 2ng/mL, which is a primary event of interest that reflects recurrence/relapse of the disease [18]. In this thesis, I focus on understanding clinical pathways, PSA dynamics and relevant events occurring in *localised* prostate cancer. PSA dynamics in metastatic prostate cancer vary significantly; the treatment pathway is considerably different with an intent on prolongment of life.

There are known patient and tumour risk factors that affect prognosis of prostate cancer. These include the age of the patient, PSA levels at diagnosis (or presenting PSA), tumour stage (T-stage, as per the TNM scoring system) and Gleason score/grade grouping [19]. These clinical risk factors are used to categorise patients into the NCCN (National Comprehensive Cancer Network) of low, intermediate, or high risk groups [20], as shown in **Table 1-1**. Prognosis of localised prostate cancer (T1–T2N0M0) is generally good after treatment; historically 5-year disease-free survival rates around 76% [21]. More recently these rates have increased to 80-90% [12,13,22].

Table 1-1 risk stratification by clinical risk factors: Clinical T-stage, Gleason score, and presenting PSA. Locally advance prostate cancer includes high-risk localised patients, as defined by the National Comprehensive Cancer Network (NCCN).

NCCN risk level	Clinical T-stage	Gleason score	Presenting PSA	*Condition to be met*
Low risk	T1-T2a	≤ 6	< 10ng/mL	*All three*
Intermediate risk	T2b-c	7	10-20ng/mL	*Any**
High risk	T3a	≥ 8	>20ng/mL	*Any*
	** i.e., exclusion by absence of any low- and high-risk features*			

The treatment pathway for localised prostate cancer, according to the UK National Institute for Health and Care Excellence (NICE), is summarised in **Figure 1-1** [23].

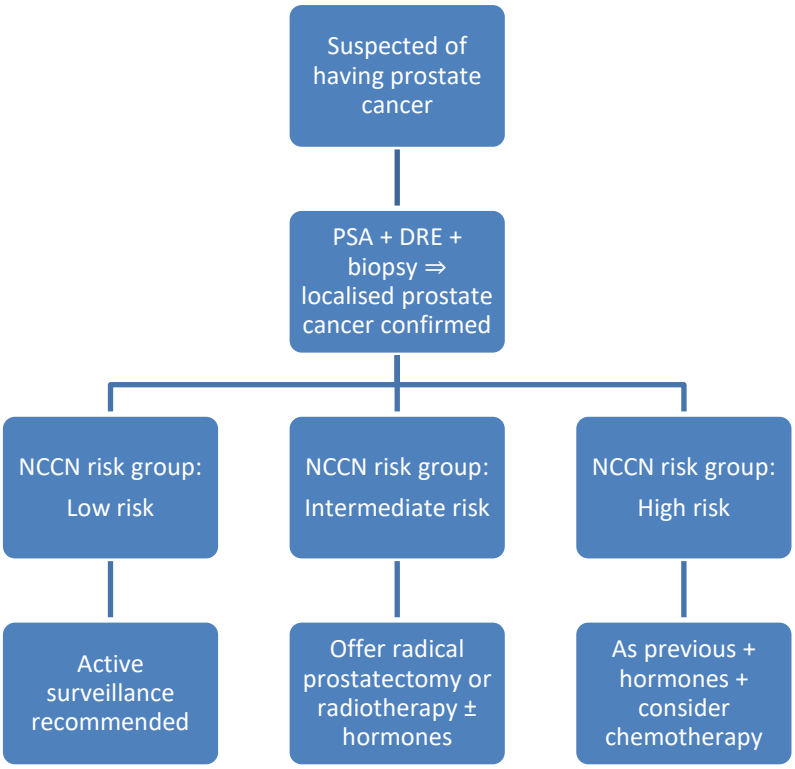


Figure 1-1 localised prostate cancer risk grouping and corresponding treatment management option, in accordance with NICE guidance. NCCN = National Comprehensive Cancer Network; NICE = National Institute for Health and Care Excellence.

There has been more recent work in reclassifying the risk level of patients using the five-tiered Cambridge Prognostic Group (CPG) classification, which is a superior predictor of prostate cancer-specific mortality [24]. NICE performed a review in December 2021 and updated their recommendations on prostate cancer risk stratification to CPG (**Table 1-2**).

Table 1-2 risk stratification by clinical risk factors: Clinical T-stage, Gleason score, and presenting PSA. Locally non-metastatic prostate cancer includes high-risk localised patients, as defined by Cambridge Prognostic Group (CPG) classification.

CPG Risk level	Clinical T-stage	Gleason score	Presenting PSA	*Condition to be met*
1	T1-T2	≤ 6	< 10ng/mL	*All three*
2	T1-T2	$3 + 4$	10-20ng/mL	*Gleason OR both PSA AND T-stage*
3	T1-T2	$3 + 4$ *or* $4 + 3$	10-20ng/mL	*All*
4	T3	8	$>$ 20ng/mL	*Any*
5	T4	9-10	—	*Any*

Though it is interesting that there have been recent advances and improvements in risk stratification, they are still defined by the underlying baseline prognostic factors and categorising PSA.

1.2 Motivation

There have been many advances in the treatment of localised prostate cancer over the last few decades, in particular the use of non-invasive modalities such as IMRT, dose-escalation of radiotherapy together with the use of neoadjuvant / concurrent hormonal therapy. There have been major advances in the understanding of the radiobiology of prostate cancer that lends itself well to exposure of hypofractionated radiotherapy, which is the current standard of care. Despite these improvements, patients' cancer can still return and so the aim is to be able to better predict which patients will experience recurrence and when.

After undergoing radiation therapy, the recurrence of the disease is monitored using PSA blood tests. These tests are repeated over time to check for any increases in PSA levels, which may indicate the return of the disease and require further evaluation. Dynamic prognostic tools that consider the full post-treatment PSA changes and other relevant disease information would be beneficial in improving the monitoring of those patients following radiotherapy, to help make informed clinical decisions.

Clinical prediction models (CPMs) are predictive tools that have become a cornerstone in healthcare and clinical practice. They allow clinicians to stratify risks of concern within a systematic framework. These tools enable personalised predictions of risk given a set of known inputs, such as clinical prognostic factors or predictors. CPMs are typically developed using routinely collected healthcare data, historical cohorts, or ongoing clinical trials via follow-up case report forms (CRFs). There are three broad applications of CPMs: screening programs (mammography) [25], diagnosis [26], and therapeutic prognosis (Predict- prostate and breast) [27,28].

Historically, CPMs for prostate cancer recurrence have been developed using clinical baseline prognostic factors, such as T-stage, Gleason score and presenting PSA. There are many existing CPMs (>100) in circulation for prostate cancer under various treatments [29,30]. In general, they use only baseline prognostic factors and therapy; very few of them use all the available longitudinal information, for example the repeated PSA concentrations over time. It is known that these clinical risk factors generally do not adequately predict recurrence on their own [31].

Individual patients have multiple outcomes that are measured and typically analysed individually. However, incorporating the different outcomes and relevant clinical and pathological markers collected during follow-up through joint statistical modelling can result in more accurate predictions of a patient's prognosis. The extended period of follow-up enhances our understanding of each patient's underlying biological mechanisms that can help to predict a possible recurrence of their disease. There is a need to better use all prognostic information from patients, particularly when patients have typical first-line hypofractionated regimes, as there is a clear lack of prognostic tools developed under these treatment modalities.

Clinicians, and patients alike, are interested in their own bespoke prognosis, and this in turn will facilitate personalised clinical management. An important aim is to distinguish between those patients who are likely to do well and those potentially at risk of recurrence, and to quantify that risk using well calibrated models. For example, "what is their recurrence-free probability within the next five years?", or "if they are yet to develop recurrence, what is their prognosis in the next two- or five years?". Is it good, (i.e., a high recurrence-free probability)

with relatively good precision/small prediction intervals? Does this patient exhibit a good chance of prolonged event-free survival, or are they likely to have recurrence within some prediction window of interest? By addressing these aims, clinicians could personalise treatment and follow-up and appropriately divert resources to provide bespoke management plans to patients who need it most. For example, whether PSA levels surpass an unacceptable risk threshold, occurring within a clinically relevant prediction window of interest. This in turn can direct the frequency of follow-ups, further appointments/additional PSA tests taken, PSA-led imaging, and then initiating appropriate salvage therapies. Conversely, if the patient is deemed very low-risk, fewer clinical visits (e.g., reduced PSA frequencies) can be recommended, thereby reducing patient burden [32]. To increase efficiency and statistical power [33], it is crucial to utilise all available patient information.

Prospective studies with extensive follow-up and high-quality data from clinical trials present a unique opportunity to undertake further work to examine various patterns of disease prognosis. In this thesis, using the data from the CHHiP clinical trial [13], I propose a clinical dynamic predictive joint model (CDPJM) for prostate cancer recurrence which, in addition to the clinical baseline prognostic factors, combines with longitudinal repeated PSA measurements obtained through follow-up over many years.

There are many existing predictive tools that have been developed to predict these outcomes, however these are developed with older treatment deliveries receiving conventional radiotherapy. The novelty of this thesis is that the model includes patient cohorts that include hypofractionation. There are no known prediction tools that have been developed under the curative hypofractionated radiotherapy with hormone treatment pathway; this thesis addresses this area of unmet need.

1.3 The CHHiP Trial

The CHHiP (**C**onventional or **H**ypofractionated **H**igh dose intensity modulated radiotherapy **i**n **P**rostate cancer, ISRCTN97182923) trial is the largest known phase-III RCT in localised prostate cancer and investigates whether hypofractionated radiotherapy is non-inferior in comparison with conventional radiotherapy. Recruitment of 3,216 patients occurred between October 2002 and June 2011. These patients predominately had NCCN intermediate localised prostate cancer (T1b–T3aN0M0). They were randomised 1:1:1 to three IMRT fractionation schedules: the standard conventional fractionation of 74 Gray (Gy) given over 37 fractions (f), over 7.4 weeks (2Gy per day for 37 weekdays); two experimental hypofractionated schedules of 57Gy/19f given over 3.8 weeks, and 60Gy/20f over four weeks where both these hypofractionation arms delivered 3Gy/f. A visual schema of the trial is depicted in **Figure 1-2**. Full study details, including details of ethics approval, have previously been published; the five-year primary analysis was reported in June 2016 [13], with a subsequent 10-year updated publication planned. This update was recently presented at *ASCO Genitourinary Cancers Symposium*, February 2023 [34]. The study showed that the 60Gy/20f fractionation schedule / treatment arm was non-inferior to the conventional 74Gy/37f program. The CHHiP RCT successfully provided evidence 60Gy/20f should be the new standard of care for external beam radiotherapy (EBRT) radical treatment of localised prostate cancer.

Due to the practice-changing evidence of CHHiP, moderate hypofractionation is now mandated by NHS England standardly used within clinics in the UK, with 96% of eligible intermediate-risk patients receiving hypofractionation [35,36]. This thesis will focus on EBRT modalities of localised prostate cancer; metastatic progression and corresponding treatment is beyond the scope of this thesis. Although a third of CHHiP patients were randomised to the conventional non-hypofractionated radiotherapy (control) arm, I have included all patients in the development of these models, the non-inferiority established in the main CHHiP results and to enable use of all clinical data to make the model more robust with an adequate sample size.

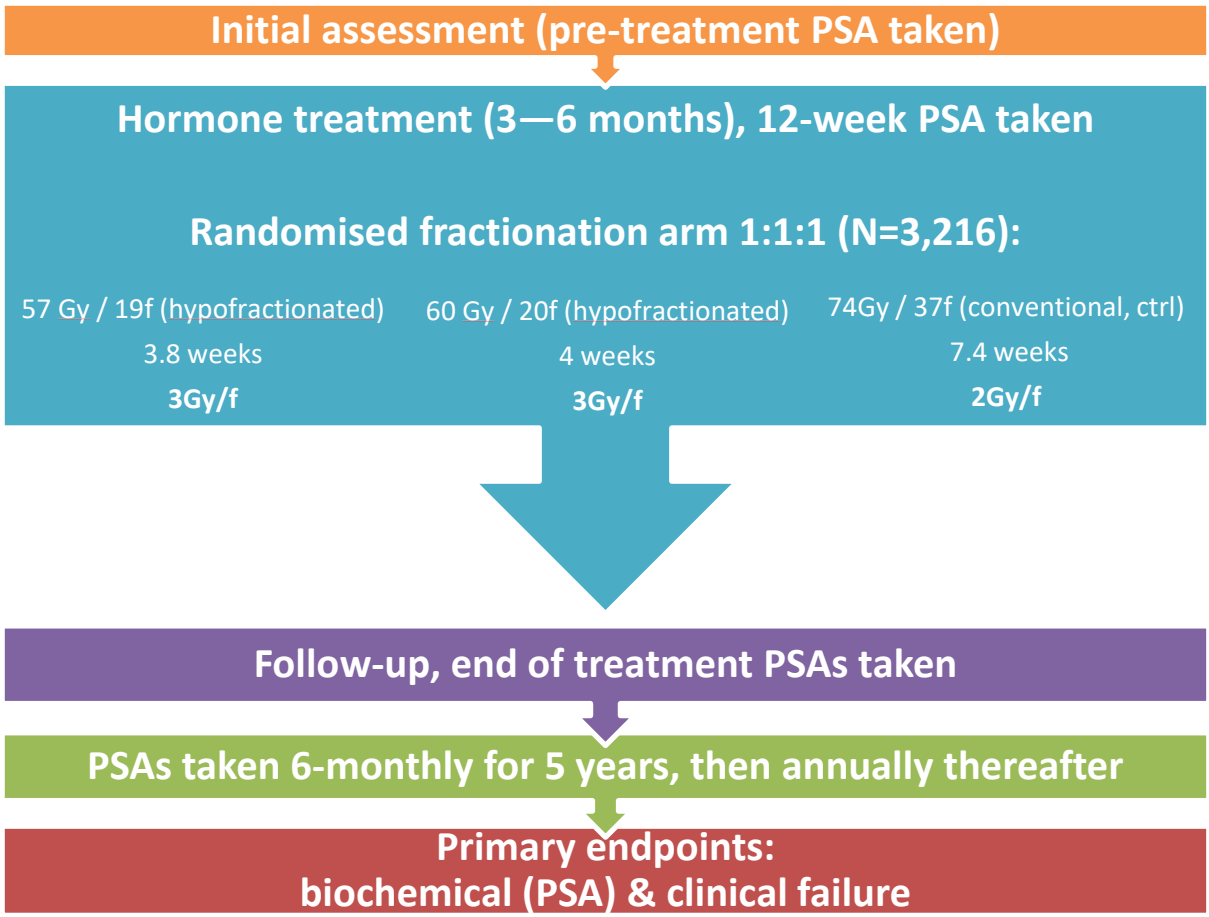


Figure 1-2 trial schema of CHHiP; ctrl = control, PSA = prostate-specific antigen.

1.4 Developing a clinical prediction model

Following PROGRESS III [37] best practice, there are mainly three stages to CPM research: model development with internal validation; external validation, and investigating impact on clinical utility. It is relatively easy to develop a CPM, however very few make it through to clinical practice, and most are forgotten [38,39].

Developing a CPM involves several considerations, including:

1) Defining the target population of interest: it is important to clearly define the target population for the prediction model, including the inclusion and exclusion criteria. In the case of this thesis, it is patients with localised prostate cancer treated with hormones and radiotherapy.

2) Identifying the predictors: these independent variables should be selected based on their clinical relevance and the availability of data. Here, these predictors will be tumour aggressiveness, age, treatment received, and the observed PSA readings.

3) Selecting the outcome: the outcome should be clearly defined and relevant to the target population. This will be recurrence of cancer.

4) Determining the sample size: it should be sufficient to accurately estimate the model parameters and to provide adequate power for statistical testing; the sample size of the studies have already been calculated. In the subsequent chapters, I will address the sample size needed for development of the prediction tool, and its validation in external cohorts.

5) Collecting the data: the data should be collected in a consistent and reliable manner. For CHHiP, this is collected through CRFs provided by hospital trusts and given to the CHHiP trial team within the ICR-CTSU to facilitate data checking and integrity.

6) Cleaning and preparing the data: the data should be cleaned and prepared for analysis, including the handling of missing data, sense-checking and querying any unusual observations.

7) Selecting the model: the appropriate model should be selected based on the nature of the outcome and the predictors.

8) Assessing model performance: this should be assessed using appropriate evaluation metrics, such as discrimination and calibration, together with resampling methods to assess possible over-optimism, and the use of external cohorts to assess generalisability.

The TRIPOD (**T**ransparent **R**eporting of a multivariable prediction model for **I**ndividual **P**rognosis **O**r **D**iagnosis) framework provides recommendations for the reporting of developing and evaluating the CPM in a comprehensive and transparent manner [40]. The TRIPOD checklist will be used to guide the development of the proposed clinical dynamic predictive tool.

1.5 Thesis purpose and objectives

The overall aim of this thesis is to propose, develop and validate a statistical clinical dynamic prediction model to predict cancer recurrence in localised prostate cancer patients treated with hormone therapy and radiotherapy, including hypofractionation. To achieve this, I will use well curated data from clinical trials (including CHHiP for developing the model) to fit a joint model that simultaneously models the longitudinal process (repeated PSA measurements over time) and a time-to-event outcome (time to recurrence).

The specific research aims and objectives of this thesis are the following:

1) To review and use the appropriate modelling framework and corresponding methodology currently in existence.
2) Using this literature to inform development of a prediction tool, using all known information on the patient, given hypofractionated radiotherapy.
3) To assess whether the joint model accurately reproduces similar treatment effects of CHHiP in the hypofractionation schedule, compared to the standard Cox model.
4) To quantify the association between PSA trajectories and time-to-recurrence of prostate cancer.
5) To produce reliable and clinically relevant dynamic predictions of prognosis, given updated PSA data for the patient.
6) To determine how well the developed predictive tool can work in practice; validate and assess the generalisability of the model in other unseen patients of localised / locally advanced prostate cancer populations, and where alternative treatment pathways may be used.

7) To extend the joint model to consider the presence of competing events, for instance a death unrelated to prostate cancer, and ascertain whether they lead to more accurate predictions for the main event of interest, recurrence.
8) Assessing whether the tool can be crafted to have clinical utility, considering the regulatory requirements involved in publishing said tool.

How these aims will be addressed in the subsequent chapters are discussed in turn in section **1.6**.

1.6 Thesis description & outline of subsequent chapters

Chapter 2 will set out the proposed infrastructure for the joint model. It will lay the foundation of the underlying mathematical methodology in the literature and the basis for the subsequent results chapters. This includes the modelling of both the stochastic longitudinal PSA biomarker and the time-to-recurrence of cancer processes. It will set about the subsequent notation used, the estimation procedure of each joint modelling framework, introduce modelling comparisons and the underlying mathematics to elicit the important dynamic predictions. Finally, the evaluation of those dynamic predictions will be specified.

Chapter 3 synthesises a review of the literature in already existing studies that have applied joint modelling to prostate cancer over the last two decades. Given changes in clinical practice and the historical nature of the existing clinical prediction models, it is necessary to 'take stock' of what is already published. This will lay the foundation and modelling rationale for the subsequent application chapters. This chapter is related to the publication by the author of this thesis [41].

Chapter 4 will feature the application of joint modelling applied to the CHHiP trial [13]. The aim is to develop an individualised CDPJM, to predict possible recurrence of prostate cancer, in particular predicting the primary endpoint of biochemical and clinical failure. How the baseline prognostic factors impact PSA trajectories over time and quantify changes in PSA, (e.g. how an increase in PSA value and its rate of change impacts the risk of recurrence). Furthermore, dynamic predictions are extracted, comparing patients with opposing baseline prognostic factors (low- *vs* high-risk) with the same treatment schedules. I quantify the predictive performance of the joint model, with internally validated bias-corrected metrics on

calibration and discrimination. I also investigate PSA thresholds / cut-offs over various follow-up landmark times that are indicative of good prognosis. This chapter is related to the recent publication by the author of this thesis [42].

Chapter 5 will investigate the proposed dynamic predictive joint model (developed in **Chapter 4**) in external settings, in particular to externally validate the CDPJM in unseen patients not used to develop the model. External validation is the gold standard when it comes to evaluating the predictive performance of any developed prediction tool [37,40,43–45]. The two external RCT cohorts are RADAR and RT01 [46,47], with their own disease stages and treatment regimens, which differ slightly from the CHHiP trial. Further details on these external cohorts will be presented in the chapter. The predictive performance of the CDPJM is assessed in these external cohorts, together with the generalisability and clinical utility of the model in these extended unseen populations.

Chapter 6 considers extending the original CDPJM of **Chapter 4** to consider competing endpoints, in particular the competing risk of death unrelated to prostate cancer. It is important to consider competing risks in developing a clinical prediction model because the competing event can affect the accuracy of the model and overpredict the risks of the primary endpoint of prostate cancer recurrence. Competing risks refer to the fact that a patient may die from a cause other than the disease that the model is to predict. For example, the developed model is trying to predict the likelihood of prostate cancer recurrence, therefore the death of a patient from a heart attack (or another cause of death completely unrelated to the original disease) would be considered a competing event. If these competing risks are not considered appropriately, the model can underestimate the true risk of the endpoint of interest. This can lead to incorrect predictions and potentially harmful clinical decisions. On the other hand, if competing risks are properly accounted for, the model can provide more accurate predictions, which can help clinicians make more informed decisions about patient care.

Finally, **Chapter 7** will summarise and discuss the findings of the thesis, draw conclusions, and make recommendations for future work. Some further work, supplementary figures and tables can be found in the **Appendices,** corresponding to **Chapters 4, 5, & 6.**

Chapter 2 - Joint Modelling Methodology

2.1. Publications relating to this chapter

Joint models for dynamic prediction in localised prostate cancer: a literature review. **Harry Parr**, Emma Hall, Nuria Porta. *BMC Medical Research Methodology* volume 22, Article number: 245 (2022) https://doi.org/10.1186/s12874-022-01709-3

2.2. Introduction

Clinical prediction models are developed from patient and tumour features at diagnosis, as well as information on treatment received (such as fractionation dose) to predict future prognosis. To date, there are many CPMs to guide management decisions for localised prostate cancer, e.g., visualised in nomograms and online calculators [26,27,29,30,48–52]. These CPMs only consider information available at the time of diagnosis and/or at start/end of a treatment course, and PSA values collected after that timepoint are rarely considered, and then often only for the definition of biochemical failure. However, it is of interest to both patient and clinician to examine the association of the biomarker of interest over time to prognosis. Knowing the patient is alive and recurrence-free at a new post-treatment visit, with an updated PSA value, is informative. Including this new information into a prediction model can elicit dynamic predictions that enable updated prognosis of patients' outcome(s).

One approach to model the time-dependent PSA biomarker and assess its impact on recurrence is the extended Cox, or Andersen-Gill model [53,54]. This model is built using a counting-process methodology that divides data into time intervals (start-stop] for each PSA measurement recorded over time. This formulation is set to handle exogenous time-dependent covariates, which is not an appropriate assumption for the PSA biomarker [55,56]. Exogenous covariates are entirely predictable processes that are fully specified, and measured without error, and do not change when the endpoint / event is observed. They are assumed to remain constant in between visits and only change when observed during follow-up, which is a very unrealistic assumption for biomarkers. For example, a patient's age at the start of a study is an exogenous time-varying covariate because the age of a patient does not change due to the event of interest (such as disease recurrence). Conversely, endogenous time-dependent covariates, are variables that are influenced by the event of interest. For example,

a biomarker such as PSA is an endogenous time-dependent covariate because the value of PSA is affected by the event of interest (biochemical failure). The value of an endogenous time-dependent covariate can be impacted by measurement error and biological variability, making it challenging to model accurately. More comprehensive details of this can be found in [57].

A further extension is to use landmark modelling [55,58–62]: dynamic predictions are obtained by fitting time-dependent Cox models to the patient subsample still at risk at several prediction, or landmark times of interest, together with the value of the longitudinal biomarker at that time. Landmark models are straightforward to fit with standard software, but no measurement error for the time-varying biomarker is considered nor is the entire longitudinal history of the biomarker utilised (due to using the last observation carried forward) [60]. To improve predictions, a two-stage approach to landmarking (also known as mixed model landmarking [55,63]) can be considered to model measurement error and incorporate the full biomarker history. However, uncertainties in the mixed-effect model estimates are not carried through to the survival submodel, resulting in overexact estimates [64].

Joint models permit dynamic prediction in localised prostate cancer by considering two time-dependent processes simultaneously: the repeated longitudinal PSA biomarker over time (modelled using a mixed-effects submodel), and the time to an event of interest (modelled using a relative-risk, or Cox submodel). The event of interest can be biochemical failure, recurrence of disease (either locally in the prostate gland, in the regional lymph nodes, or distant organ metastases), clinical failure (need to recommence hormone therapy), death, or a composite of all these events.

In this chapter, I will review the underlying methodology for developing a joint model, extracting dynamic predictions, and evaluating the predictive performance of the model.

2.2.1. Notation

In the treatment of localised prostate cancer, the endpoint I use is time to recurrence (recurrence-free survival); which is a composite of biochemical and/or clinical failure. The definition of biochemical failure after patients are treated with hormones and/or radiotherapy is where a PSA threshold is greater than the nadir PSA (lowest recorded PSA concentration at

any time after commencement of treatment, i.e., hormone therapy and radiotherapy) plus 2 ng/mL, ($\mathrm{PSA} > \mathrm{PSA}_{\mathrm{nadir}} + 2\mathrm{ng/mL}$). This is known as the Phoenix definition of biochemical failure and requires clinical confirmation [18]. Clinical failure is defined as a recommencement of hormonal therapy, local recurrence, lymph nodal or pelvic recurrence, or distant metastases; these are composite endpoints of clinical failure and can (rarely) preclude biochemical failure.

Here I introduce some formal mathematical notation. For the i^{th} patient, let T_i be the observed failure time, the minimum of $T_i^* \,\&\, C_i$; $T_i = \min(T_i^*, C_i)$, i.e., the first of the two outcomes: the true event time is denoted T_i^*, i.e., the duration (in years) from the initial pre-treatment PSA time ($t = 0$) to recurrence, and C_i is the censoring time. An indicator variable $\delta_i = I(T_i^* \leq C_i)$ is unity if the event of interest is observed for that patient, or zero otherwise. Using the observed event indicators and times, one wishes to make inferences on the possible true time-to-recurrence T_i^*.

As visits and follow-up typically range from biannually to annually, interval-censoring may take place, i.e., time to recurrence is not observed exactly, however it is known it may have taken place sometime between two consecutive visits for a patient i, $t_1 < T_i^* \leq t_2$. However, I will assume the event occurred exactly at the first visit when there is knowledge of it (i.e., at t_2), and assume the data is only subject to right censoring, as defined above.

For the analysis of the repeated biomarker measurements, I define $y_i(t)$ to be the longitudinal covariate at time t and $\boldsymbol{y}_i = \mathrm{y_i}(\mathrm{t_{ij}}); \mathrm{i} = 1, \ldots, \mathrm{N};\ \mathrm{j} = 1, \ldots, \mathrm{n}_i\}$, the longitudinal response vector of the continuous biomarker measurements for the i^{th} patient and j^{th} biomarker reading taken at time t_{ij}. There are N patients with n_i longitudinal measurements per patient. Let $\mathcal{Y}_i(t) = \{y_i(u), 0 \leq u < t\}$ be the biomarker history up to time point t.

2.3. Joint modelling specification and estimation

In developing the joint model, there are two components that need to be accounted for: the longitudinal and the time-to-event processes. From these two submodel components, one can develop a *dynamically* updated clinical predictive joint model.

For the longitudinal process, i.e., the repeatedly-measured biomarker PSA; typically a mixed-effect submodel is used [65]. It is 'mixed' in the sense that it contains both fixed and random

effects. The former describes the average longitudinal trajectory of the patient population of interest, averaged across all patients. This submodel is employed as the assumed underlying / latent patient-specific random effects account for the correlated within-patient repeated measures; these also account for the biological variation and possible unbalanced panel data (i.e., one patient has more longitudinal PSA information than another). These are unique to each patient; they allow one to quantify the deviation away from the overall fixed effects of the population. Hence the namesake *mixed*, combining these two sources of effects. Given these random effects, each patient is assumed to have a 'random intercept' i.e., have their own unique presenting pre-treatment PSA and random slopes i.e., have their own subject-specific mean PSA response over time, modelled with a specified parametric form.

The time-to-event process is modelled using a relative risk submodel, whereby a Cox submodel is usually used. This is estimated using the usual survival analysis framework with corresponding log-hazard ratios to quantify the association of the baseline covariates to the outcome of interest: recurrence. There are two additional components of the relative risk model; unlike the standard Cox model, the baseline hazard is typically specified with a flexible parametric function, usually with penalised basis splines, or utilising a piecewise constant model.

There are two joint modelling frameworks one can use: shared parameter joint models, (SPJMs), or joint latent class models (JLCMs). In both frameworks, the random effects account for the variability of the PSA biomarker. For SPJMs, those same random effects are assumed to account for the association between the longitudinal process and its impact on the relative risk component, via the inclusion of the functional form to quantify its association. Whereas in the JLCM, the random effects are used only to account for the correlated repeated biomarker measures, whilst the latent classes account for the dependency of the two outcomes [66].

Further details on each framework are presented below in sections **2.3.1** & **2.3.2** with a further discussion on the pragmatic differences found in **Chapter 3**.

2.3.1. Shared-parameter joint model

The longitudinal process $\boldsymbol{y}_i$ is assumed to follow a mixed-effects model, defined by a linear combination of possibly time-dependent main and random effects $Y_i(t_{ij}) = m_i(t) + \boldsymbol{\epsilon}_i(t_{ij}) =$

$\boldsymbol{\beta} X_i(t_{ij}) + \boldsymbol{b}_i Z_i(t_{ij}) + \boldsymbol{\epsilon}_i(t_{ij})$. The vector $\boldsymbol{\beta}$ are coefficients for the main and time-effect covariates of the design matrix X_i, and the corresponding random effects $\boldsymbol{b}_i$ for the Z_i design matrix. The measurement errors $\boldsymbol{\epsilon}_i(t_{ij}) = \{\epsilon_i(t_{i1}), \dots, \epsilon_i(t_{in_i})\}^T$ are time-dependent and assumed to follow $\boldsymbol{\epsilon}_i(t_{ij}) \sim N(0, \sigma_e^2)$, or t-distributed with several degrees of freedom, with the fatter tails used to accommodate possible outliers. The random effects $\boldsymbol{b}_i$ (independent of $\boldsymbol{\epsilon}_i(t_{ij})$), are usually assumed to follow a multivariate normal distribution, with an unknown square covariance matrix structure D, $\boldsymbol{b}_i \sim \text{MVN}(\boldsymbol{0}, D)$.

A relative risk, or proportional hazards model, is used for the parameterisation of the survival submodel:

$$h_i(t|\boldsymbol{M}_i(t), \boldsymbol{w}_i) = \lim_{\Delta t \to 0} \frac{\Pr\{t \le T_i^* < t + \Delta t \,|T_i^* \ge t, \boldsymbol{M}_i(t), \boldsymbol{w}_i\}}{\Delta t}$$
$$= h_0(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + f(\boldsymbol{M}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha})\}.$$

Where $\boldsymbol{M}_i(t) = \{m_i(s), 0 \le s < t\}$ denotes the true (unobserved) and entire longitudinal biomarker history up to time point t, with $m_i(t)$ indicating the true value at t (i.e., the mixed effect model not contaminated with measurement error). The baseline hazard is denoted $h_0(t)$, with covariates in the hazard submodel being $\boldsymbol{w}_i$, with $\boldsymbol{\gamma}^T$ corresponding to the log-hazard ratio coefficients. An example of the parameterisation of the functional form $f(\boldsymbol{M}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha})$ can be a linear combination of value and gradient of the longitudinal biomarker, $f(\boldsymbol{M}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}) = \alpha_1 m_i(t) + \alpha_2 \frac{\mathrm{d}m_i(t)}{\mathrm{d}t}$. The corresponding $\boldsymbol{\alpha}$ parameters are interpreted as the log-hazard ratios that quantify the intensity of association between the two outcomes. Other functional forms of f exist, such as the (weighted) cumulative effect (1), or random effects association (2),

$$f = \alpha \int_0^t \omega(t - s) \times m_i(s)\, ds \tag{1}$$

$$f = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{b}_i \tag{2}$$

The former, (1), quantifies the risk of recurrence from the area under the biomarker trajectory and can allocate greater weights to more recent biomarker observations (e.g. using a standard normal density function for ω). The latter, (2), parameterisation uses only the random effects as a linear predictor, this requires no numerical integration which is computationally

advantageous. Using a simple random intercept and slopes structure is most interpretable, whereby patient deviations from the population average is expressed [67]. General extensions to $f(\boldsymbol{M}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha})$ exist, incorporating multiple longitudinal outcomes (e.g. testosterone in addition to PSA) with more elaborate structures; however these are challenging to interpret [56,68,69].

In the usual Cox survival framework, it is typical to leave the baseline hazard function unspecified, then estimate the regression coefficients of the relative risk model by maximising its partial likelihood function. A full parametric specification of the baseline hazard function, $h_0(t)$, is recommended (e.g. using constant-piecewise, or regression splines models, with an adequate number of knots for flexibly modelling the underlying baseline risk). Leaving $h_0(t)$ unspecified can lead to underestimating the precision of the parameter estimates [70]. In particular to the joint modelling framework, penalised basis-splines are often employed to estimate $h_0(t)$ and can be expressed as:

$$h_0(t) = \exp\left(\psi_{h_{0,0}} + \sum_{q=1}^{Q} \psi h_{o,q}\, B_q(t, \boldsymbol{v})\right)$$

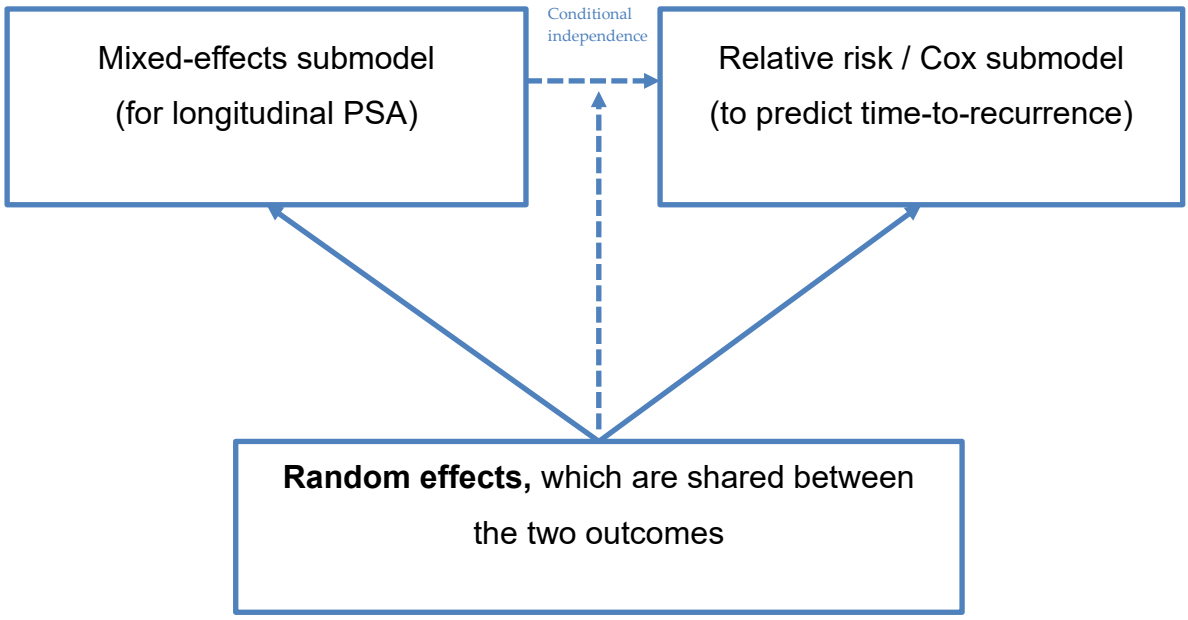
i.e., a linear combination of B-splines $B_q(t, \boldsymbol{v})$ with q^{th} basis function and a vector of spline coefficients ψ. Alternatively, a constant piecewise model can be considered taking the form,

$$h_0(t) = \sum_{q=1}^{Q} \zeta_q\, I\left(v_{q-1} < t \leq v_q\right).$$

Where $v_0 < v_1 < \cdots < v_Q$ are the splits in the follow-up time scale, with v_Q being greater than the maximum observed time, ζ_q is the value of the hazard indicated within the interval $\left(v_{q-1}, v_q\right]$.

Given these time-independent random effects $\boldsymbol{b}_i$, both outcomes Y_i and T_i become independent from one other, known as conditional independence [56]. This is depicted in **Figure 2-1**.

Figure 2-1 a graphical representation of how the random effects elicit conditional independence of the two outcomes in the shared-parameter joint model specification.



I define the likelihood functions of the mixed-effect model as $p_1(y_i|\boldsymbol{b}_i;\boldsymbol{\theta})$, the time-to-event relative risk model as $p_2(T_i,\delta_i \mid \boldsymbol{b}_i;\boldsymbol{\theta})$ and its joint distribution $p_{12}(T_i,\delta_i,y_i \mid \boldsymbol{b}_i;\boldsymbol{\theta})$ given the random effects under the conditional independence assumption and its estimated parameters.

$$p_1(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}) = \prod_j p_{\bar{1}}\big(\boldsymbol{y}_i(t_{ij})\big|\boldsymbol{b}_i;\boldsymbol{\theta}\big) \tag{3}$$

$$p_{12}(T_i,\delta_i,\boldsymbol{y}_i \mid \boldsymbol{b}_i;\boldsymbol{\theta}) = p_1(y_i|\boldsymbol{b}_i;\boldsymbol{\theta}) \times p_2(T_i,\delta_i \mid \boldsymbol{b}_i;\boldsymbol{\theta}). \tag{4}$$

Estimation of the parameters of the joint model can be undertaken by maximising the joint likelihood function (equation 4) under the frequentist framework, or by Markov chain Monte Carlo (MCMC) algorithms within the Bayesian framework. The latter is chosen in this thesis as it is computationally more efficient, model comparison is more straightforward, and asymptotic approximations are not required.

The Bayesian framework estimates the model parameters by sampling from the posterior distribution, using the sampling algorithms MCMC [71]. Together with equation 3, the full posterior via conditional independence is,

$$f(\boldsymbol{\theta},\boldsymbol{b}) \propto \prod_{i=1}^{n}\prod_{j=1}^{n_i} f\big(y_{ij}\,|\boldsymbol{b}_i,\boldsymbol{\theta}\big)\, f(T_i,\delta_i|\boldsymbol{b}_i,\boldsymbol{\theta}) f(\boldsymbol{b}_i,\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

The priors used are typically standard and non-informative to attain stabilised MCMC convergence, which give similar parameter estimates to the maximum likelihood [71–73]. One way to ensure the parameters have converged is using the diagnostic convergence measure: potential scale reduction factor, or $\hat{R}$. It compares the between- and within-chain estimates for model parameters; when the chains are well-mixed, $\hat{R} \to 1$ [74].

Excellent overviews of shared-parameter joint modelling can be found by Rizopoulos, and Papageorgiou et al. [56,67].

2.3.2. Joint latent class model

Alternatively, the joint latent class modelling framework assumes the existence of latent classes, which capture the association between the longitudinal biomarker trajectory and the relative risk of the endpoint of interest. Following similar notation as above, I define the JLCM by the mixed-effect and relative risk submodels for each latent class $\boldsymbol{c}_i \in \{1,..,G\}^{\mathrm{T}}$:

$$\left(Y_i(t_{ij})\middle|\boldsymbol{c}_i = g\right) = \boldsymbol{\beta}_g X_i^T(t_{ij}^*) + \boldsymbol{b}_{ig} Z_i^T(t_{ij}) + \epsilon(t_{ij});\ \epsilon(t_{ij}) \sim N(0, \sigma_e^2), \boldsymbol{b}_{ig} \sim \mathrm{MVN}(\boldsymbol{\mu}_g, D)$$

$$h_i(t\,|\boldsymbol{c}_i = g) = h_{0g}(t)\exp\left(\boldsymbol{\gamma}_g^T \boldsymbol{w}_i\right),$$

where assignment to latent class g is given by a multinomial submodel,

$$\pi_{ig} = \Pr(\boldsymbol{c}_i = g \mid X_i) = \frac{\exp\left(\boldsymbol{\lambda}_g^T X_i\right)}{\Sigma_{j=1}^{G}\exp\left(\boldsymbol{\lambda}_j^T X_i\right)}.$$

With X_i a fixed baseline design matrix associated with classification and corresponding coefficients $\boldsymbol{\lambda}_g^T = (\lambda_0^T = 0, \lambda_1^T, \ldots, \lambda_G^T = 0)$. Given the latent class $\boldsymbol{c}_i$, conditional independence between the longitudinal and time-to-event outcomes is assumed.

JLCMs are typically estimated by maximum likelihood using the Marquardt algorithm, given a fixed number of latent classes G. There can be issues in using Bayesian estimation, known as 'label switching', due to the symmetric nature of the likelihood and model parameters, where a different set of parameters will provide the equivalent likelihood [75]. This can be identified by large $\hat{R}$ and with non-overlapping traceplots [76]. This issue does not apply in the frequentist framework [66].

Assuming the conditional independence assumption, the log-likelihood of the observed data can be given by,

$$\sum_{i=1}^{N} \log\left(\sum_{g=1}^{G} \pi_{ig}\, f\left(Y_i(t_{ij})\middle|\boldsymbol{c}_i = g\right) h_i(T_i|\boldsymbol{c}_i = g;\boldsymbol{\theta}_G)^{\delta_i}\, S_i(T_i|c_i = g;\boldsymbol{\theta}_G)\right).$$

Where $\boldsymbol{\theta}_G$ is the full vector of JLCM parameters with G classes; π_{ig} is the above class-membership probability; $f\left(Y_i(t_{ij})\middle|\boldsymbol{c}_i = g\right)$ is the probability density function of the longitudinal mixed-effect model, given G classes; the instantaneous risk function is $h_i(T_i|\boldsymbol{c}_i = g;\boldsymbol{\theta}_G)$, and $S_i(T_i|c_i = g;\boldsymbol{\theta}_G)$ is the class-specific survival function.

The JLCM has some advantages compared to the SPJM: it does not need to specify a suitable functional form to link the two processes, and thus the conditional independence assumption in the JLCMs results in less onerous computations. However, as the number of latent classes are not known a priori, it is another component to be estimated, and as these are not observed, the conditional independence assumption is nontrivial to evaluate. Jacqmin-Gadda and colleagues proposed a trivariate score test to evaluate this assumption; they showed that their score test was uniformly most powerful and simpler than all others they considered [77].

2.4. Model comparisons

The below model comparison methods can be applied to both joint modelling frameworks. In the frequentist joint modelling framework (both SPJMs & JLCMs), model comparisons are typically made using Akaike information criterion (AIC) and Bayesian information criterion (BIC). Both criteria are a function of model complexity and goodness-of-fit (usually the maximised likelihood value *vs* the number of parameters); the latter is used typically to determine the optimum number of latent classes.

Model comparisons in the Bayesian framework (typically SPJMs) have some differences, on the Deviance information criterion (DIC), which is analogous to the AIC/BIC, i.e., the goodness-of-fit offset by model complexity. The DIC is calculated by $\text{deviance}(\bar{\theta}) + 2P_D$ [78], i.e., the sum of the deviance function (goodness-of-fit) estimated using the posterior estimates of the parameters, and twice the effective number of parameters $\left(P_D = \overline{D(\theta)} - \text{deviance}(\bar{\theta})\right)$, where $\overline{D(\theta)}$ is the posterior mean of the deviance, i.e., the average of the variances that are

calculated using the estimated parameters at each iteration of the MCMC sampler, minus the deviance evaluated at the posterior mean of the parameters.

$$\text{DIC} = \frac{2}{S}\sum_{s=1}^{S}\sum_{i=1}^{N}\log\{f\left(T_i^*, \delta_i, \boldsymbol{y}_i|\theta^{(s)}\right)\} - \sum_{i=1}^{N}\log\{f(T_i^*, \delta_i, \boldsymbol{y}_i|\bar{\theta})\},$$

where $\theta^{(s)}$ is the parameter sample at the s^{th} $(s = 1, \dots, S)$ iteration of the (Gibbs) sampler and $\bar{\theta}$, are the means of the posterior samples. As with the AIC/BIC, a lower DIC indicates better fit when comparing other models [79].

An alternative measure used in the Bayesian literature is the conditional predictive ordinate (CPO) [80]. The CPO is a model diagnostic metric, as it explains how the i^{th} patient's data / covariates agree or diverge from the model, with larger values of the CPO are indicative of a better fit. For the i^{th} patient observation, the CPO-statistic is defined as,

$$\begin{aligned}\text{CPO}_i &= f\left(T_i^*, \delta_i, \boldsymbol{y}_i|D^{(-i)}\right)\\ &= \int f(T_i^*, \delta_i, \boldsymbol{y}_i|\theta, D_i)\,\pi(\theta|D^{(-i)})\,d\theta,\end{aligned}$$

where D_i is the i^{th} patient's covariate data, and $D^{(-i)}$ are the data for all patients other than i^{th}. A solution for the integral is analytically intractable, particularly in higher dimensions, so MCMC approaches are used to estimate the integral; as proposed by Zhang and colleagues [81]. A related model-comparison measure is the log pseudo-marginal likelihood (LPML), which is defined as $\text{LPML} = \sum_{i=1}^{n}\log(\text{CPO}_i)$ [72]. Alternatively, there is the widely-applicable/Watanabe-Akaike information criterion (wAIC) proposed in [82]. The wAIC is seen as a more stable estimator of the DIC, where the DIC can be seen as an approximation to the wAIC . The LPML and wAIC are asymptotically equivalent estimators of external validation prediction errors [83].

2.5. Dynamic predictions

The crux of developing joint models is to elicit dynamic predictions of the probability of the event of interest occurring, which can be obtained from both modelling frameworks. These models can provide clinicians with bespoke and tailored dynamic predictions for an individual patient, to help to convey the involved risk of recurrence and guide clinical decision-making management decisions. For each individual patient and at specific time

points during their follow-up, one wants to consider all the available information, i.e., the accrued PSA biomarker information, in addition to baseline prognostic factors.

The probability for a patient to be free of the event of interest by a future time $u > t$ can be estimated, i.e., $\pi_i^*(u|t)$, the conditional probability of being event-free at time u. Given the information available up to time $t > 0$, the patient is still at risk of the event at time t ($T^* > t$), the baseline covariates / fixed effects design matrix X_i, the biomarker longitudinal values observed up to time t, $\boldsymbol{y}_i(t)$, and the parameters $\boldsymbol{\theta}$ estimated from the joint model. These predictions can be *dynamically* updated when new biomarker information becomes available at $t' > t$ [55,84–86].

For the SPJMs, these can be obtained by integrating over the random effects:

$$\begin{aligned}\pi_i^*(u|t) &= \Pr(T_i^* \geq u \,|T_i^* > t, X_i, \boldsymbol{y}_i(t), T_i, \delta_i, \boldsymbol{w}_i; \boldsymbol{\theta}) \\ &= \int \Pr(T_i^* \geq u \,|T_i^* > t, \boldsymbol{y}_i(t), \boldsymbol{b}_i;\ \boldsymbol{\theta}) \times p(\boldsymbol{b}_i|T_i^* > t, \boldsymbol{y}_i(t); \boldsymbol{\theta})\, d\boldsymbol{b}_i \\ &= \int \Pr(T_i^* \geq u \,|T_i^* > t, \boldsymbol{b}_i, \boldsymbol{\theta}) \times p(\boldsymbol{b}_i \mid T_i^* > t, \boldsymbol{y}_i(t); \boldsymbol{\theta})\, d\boldsymbol{b}_i \\ &= \int \frac{S_i(u|\boldsymbol{b}_i; \boldsymbol{\theta})}{S_i(t|\boldsymbol{b}_i; \boldsymbol{\theta})}\, p(\boldsymbol{b}_i|T_i^* > t; \boldsymbol{y}_i(t); \boldsymbol{\theta})\, d\boldsymbol{b}_i\,,\end{aligned}$$

where $S_i(\ldots)$ is the conditional survival function given the random effects for the i^{th} patient. That is, for a specific i^{th} individual, it is estimated based on the posterior predictive distribution with the conditional independence assumption from equation (1).

Similarly, for the JLCM, the predicted probabilities are given by summing over the latent classes (instead of integrating over the distribution of the random effects):

$$\begin{aligned}\pi_i^*(u|t) = \sum_{g=1}^{G} &\Pr(T_i^* \geq u|T_i^* > t, c_i = g, X_i, \boldsymbol{y}_i(t), T_i, \delta_i, \boldsymbol{w}_i; \boldsymbol{\theta}) \\ &\times \Pr(c_i = g|T_i^* > t\,, X_i, \boldsymbol{y}_i(t), T_i, \delta_i, \boldsymbol{w}_i; \boldsymbol{\theta})\end{aligned}$$

In either joint modelling framework, $\pi_i^*(u|t)$ is difficult to compute analytically, therefore MCMC methods are implemented. MCMC extracts the predicted probabilities of the posterior distribution of $\pi_i^*(u|t)$ and corresponding credible intervals from the Monte Carlo sample percentiles of interest [66,85]. These can be extracted using the following MCMC sampling scheme:

1. Draw $\boldsymbol{\theta}^*$ from MCMC sampling of the posterior $p(\boldsymbol{\theta}|\boldsymbol{D}_n)$;
2. draw $\boldsymbol{b}_i^*$ from $p(\boldsymbol{b}_i \mid T_i^* > t, \boldsymbol{y}_i(t); \boldsymbol{\theta})$;
3. compute $\pi_i(u|t, \boldsymbol{b}_i^*; \boldsymbol{\theta}^*) = \frac{s_i(u|\boldsymbol{b}_i; \boldsymbol{\theta})}{s_i(t|\boldsymbol{b}_i; \boldsymbol{\theta})}$

The above three steps are repeated many M times to obtain an estimate of $\pi_i^*(u|t)$,

$$\widehat{\pi^*}_i(u|t) = \frac{\sum_{m=1}^{M} \pi_i^{(m)}(u|t)}{M}.$$

From the above estimator of π_i^*, the 95% credible intervals can be obtained using the sampled Monte Carlo percentiles at 2.5% and 97.5%.

Conversely, one can calculate the probability of recurrence, the cumulative incidence (one minus recurrence-free event) $\pi_i(u|t) = 1 - \pi_i^*(u|t) = \Pr(T_i^* < u | T_i^* > t)$. The focus in the subsequent chapters will be on the cumulative incidence parameterisation.

2.6. Evaluating predictive performance

Measuring predictive ability is crucial to assess the proposed model(s) performance in producing accurate predictions, the end goal for any prediction model. Two aspects of modelling performance can be assessed: calibration (how well the model predicts the observed data) and discrimination (how well the model can distinguish between those patients who do and do not have an event). Previous work has developed these assessments using inverse probability of censored weighting (IPCW) from Kaplan-Meier-based estimators [87,88]. In this chapter and thesis, the focus remains on model-based weighting that captures informative censoring (e.g. removal from a study before formal biochemical failure has been achieved and therefore not recording the event directly). The model-based estimators correct for censoring by modelling the censoring distribution directly [55].

Given the full likelihood of the joint model, consider the longitudinal PSAs $\boldsymbol{y}_i$, the true time to recurrence T_i^*, the PSA schedule T_i^{PSA}, and an indicator of removal from follow-up without having observed recurrence $r_i^* = (\delta_i = 0)$. It is assumed that T_i^{PSA} & r_i^* can only depend on the observed PSA values $\boldsymbol{y}_i$; the full joint likelihood is,

$$f_{12}\left(\boldsymbol{y}_i, T_i^*, T_i^{\text{PSA}}, r_i^* \mid \boldsymbol{\theta}, \boldsymbol{\phi}\right) = f_1(\boldsymbol{y}_i, T_i^* \mid \boldsymbol{\theta}) \times f_2\left(T_i^{\text{PSA}}, r_i^* \mid \boldsymbol{y}_i, \boldsymbol{\phi}\right).$$

Where $\boldsymbol{\theta}$ is the vector of the complete joint model parameters and $\boldsymbol{\phi}$ is another vector of parameters for the T_i^{PSA} & r_i^* processes only. This decomposition reveals that although the processes T_i^{PSA} & r_i^* can be established from $\boldsymbol{y}_i$, the f_2 term can be ignored and focus the inference on f_1, as f_2 does not carry any information for $\boldsymbol{\theta}$, when the interest lies in the inference of the $\boldsymbol{\theta}$ parameters. This shows the joint model is valid in the presence of informative censoring, hence valid predictive performance metrics, presented below.

Discrimination is typically assessed by considering the time-dependent AUROC (area under the receiver operating characteristic curve) [85,87,89,90]. Within a particular chosen prediction window, AUROC (or simply AUC) values of 0.5 indicate random chance assignment and values closer to unity indicate better model discrimination. One wishes to predict whether patients are likely to continue being recurrence-free at some point in the future (time $u = t + \Delta t$), given they are free from recurrence at current time t, $(t, u]$. A patient i can be deemed to either have experienced the event if $\pi_i^*(u|t) < c$ or otherwise $\pi_i^*(u|t) > c$, with $c \in [0, 1]$, i.e., for a pair of patients $\{i, j\}$ who are randomly chosen where both their accrued biomarker readings up to time t are known, the AUC can be calculated for varying values of c, which is a measure of the discriminative ability of the joint model, this is given by,

$$\text{AUC}(u|t) = \Pr\left[\pi_i^*(u|t) < \pi_j^*(u|t) \mid \{T_i^* \in (t, u]\} \bigcap \{T_j^* > u\}\right]$$

i.e., it is expected that the joint model is to assign a higher probability of being event-free within the prediction window to the patient who has yet to experience the event.

Calibration is another vital component to assess model performance. It describes how well the model's predicted probabilities align to the observed data. It is therefore imperative to assess calibration to mitigate against poor predictions [91–93]. To assess calibration, Crowson and Austin and their respective colleagues propose a graphical calibration approach to constructing the curves and assessing calibration. The predicted probabilities at the horizon time of interest $\widehat{\pi_i^*}(u|t)$ can be extracted from the fitted joint model. This is typically presented by assessing the relationship between the log-hazard of recurrence and $\log\left(-\log\left(1 - \widehat{\pi_i^*}(u|t)\right)\right)$. From the fitted joint model, the estimated probability of recurrence just prior to time t can be estimated for each value of $\widehat{\pi_i^*}(u|t)$. The double-log (complementary log-log) transformation is considered for two reasons: 1) due to increasing the likelihood of a

linear relationship between the linear predictor and the probability of recurrence; and 2) to reduce the number of knots required for the restricted cubic splines. From these calibration curves, one can calculate the integrated calibration index (ICI) and other metrics. The ICI is defined as the mean absolute difference between the observed and predicted probabilities within the entire sample, $\text{ICI} = \frac{1}{N}\sum_{n=1}^{N}\left|\widehat{\pi_i^*}(u|t)^c{}_{t_0} - \widehat{\pi_i^*}(u|t)\right|$, where the latter term within the estimator is the predicted probability of recurrence immediately before time t_0 and the former term is the smoothed calibration curve predicted probability. Other metrics can be calculated such as the E50 which is the median absolute difference between the observed & predicted probabilities [91].

The prediction error (PE) focuses on assessing the calibration of the model, and it is defined as the expectation of the difference between the observed event status $N_i(u|t) = I(T_i^* > u|t)$ and the predicted event occurrence $\pi_i^*(u|t)$, at a specific time. A loss function can be incorporated within the expectation (e.g. the mean absolute prediction error (MAPE) or squared-loss functions). The latter is also known as the Brier score (BS), which is an overall measure of prognostic performance, defined as $\text{BS}(u|t) = \mathbb{E}\left[\left(N_i(u|t) - \pi_i^*(u|t)\right)^2 \middle| u > t\right]$. The BS can be shown to be decomposed as a combination of calibration and discrimination (with an additional uncertainty term) [94–96]. Under any loss function, as the difference between these two terms decreases and tends to zero, the closer the observed and predicted event align, resulting in better predictive performance of the model. In practice, one may want to consider predictions over a window of interest, rather than specified time points, by using weighted extensions of these estimators (e.g. a weighted average absolute prediction error (WAPE) or integrated BS [97,98]). For any of these predictive measures to be valid, the censoring distributions need to be corrected for, (e.g. using inverse probability weighted estimators [55,87,88,99]).

Alternative measures of accuracy can be utilised, such as the expected prognostic observed cross-entropy (EPOCE) [100]. The EPOCE quantifies the prognostic information from the joint model at the landmark time of interest. When estimated internally, leave-one-out cross-validation of the prognostic observed log-likelihood (CVPOL) is used to correct for over-optimism [101]. For external validation, no cross-validation is required. Proust-Lima et al. [66] argue the advantages of EPOCE over the previously stated measures, including no censoring

distribution nor a prediction window is assumed, direct comparison of two joint models can be made, and that it is more reasonable to evaluate directly on the likelihood density functions. Further formulation and discussion on this predictive accuracy metric can be found in [66,100].

2.7. Discussion

Joint models give more precise estimates of the possible treatment effects on the longitudinal biomarker (indirect effect) and time-to-event endpoint (direct effect), reduces bias in the treatment effect estimates, and enable investigation of the possible direct association of the biomarker trajectory upon the endpoint [33]. The joint modelling framework is known to be more efficient than standard Cox modelling (without time-dependent factors), as it reduces the bias in quantifying the treatment effect(s) in both submodels and reduces the corresponding standard error, compared to solely a Cox model.

Another benefit of joint models is how they can account for informative censoring. Patients could drop out from the study early due to worsening/increasing PSA and/or patient anxiety (or conversely, they were not concerned for their PSA, although unlikely they would drop out for these reasons). Informative censoring could also be that patients receive an intervention (e.g. PSA tends to grow exponentially before formal recurrence). Some clinicians may take the view that some sort of salvage therapy is required, so the patient recommences androgen deprivation therapy and as a result does not achieve formal biochemical failure and is censored at the date of therapy. Given the fully specified likelihood approach of the joint model mentioned in section **2.6**, these types of informative censoring bias are inherently dealt with. It utilises model-based estimators and adjusts for the censoring distribution which corrects for these censoring patterns [102]. In this thesis, I will assume noninformative right-censoring.

Chapter 3 – Literature Review

3.1 Publications related to this chapter

Joint models for dynamic prediction in localised prostate cancer: a literature review. **Harry Parr**, Emma Hall, Nuria Porta. *BMC Medical Research Methodology* volume 22, Article number: 245 (2022) https://doi.org/10.1186/s12874-022-01709-3

3.2 Introduction

In this chapter, I synthesise a review for published applications of joint models to localised prostate cancer over the last two decades. Given the rapid popularity and use of dynamic prediction models, this review serves as a reference to assess and reflect the applied and dynamic methods used in localised prostate cancer. The main review of the identified articles is given in the results section and summarised on **Table 3-1**.

Within each modelling framework, I present the approaches that these articles used for modelling of the time-to-event process(es), the functional form of PSA, estimation of dynamic predictions and model validation strategies. Finally, an appraisal and conclusion of these models are given.

3.3 Literature Search

The search strategy included linear combinations of, {"joint model*" OR "individual* prediction"} AND {"prostate cancer" OR "prostate-specific antigen" OR "PSA"} in the title or abstract, using Web of Science and PubMed databases up to and including June 2020. A flowchart depicting the study identification strategy is given in **Figure 3-1**. A total of 751 articles were identified from the initial search parameters, 703 and 48 articles came from Web of Science and PubMed respectively. Duplicated articles were removed, leaving 702 unique papers. Novel and seminal papers that involve the joint modelling of the longitudinal biomarker PSA and time-to-event of clinical recurrence in localised prostate cancer were selected by the author of this thesis, as the focus was to understand the PSA dynamics for this disease, which can be quite different from advanced prostate cancer. Further exclusions were made on inspecting the abstract, these included: advance/metastatic disease; different disease; no joint modelling undertaken, or alternative machine learning/artificial intelligence methods

used; simulated data used; predicting alternative endpoints such as time to diagnosis or death; no dynamic predictions derived; and whether focus was on methodology development.

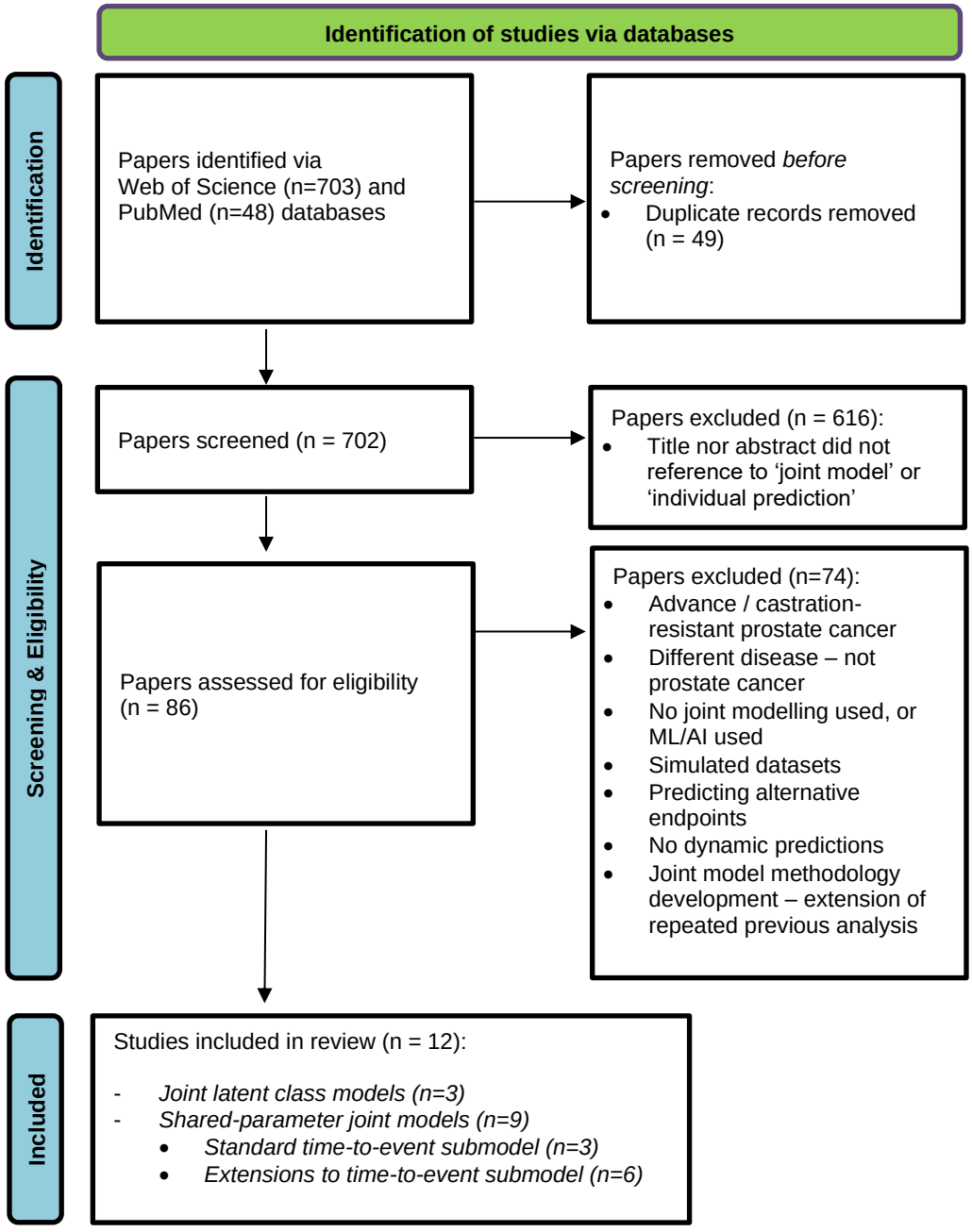


Figure 3-1 a flowchart for identifying studies of the literature review.

There were 12 identified relevant full-text papers that best illustrated the joint modelling framework and summarised its applications in localised prostate cancer; these were selected to be included within this chapter. **Table 3-1** summarises these 12 papers including details of the modelling framework used, sample sizes, parameterisations, the prediction windows of interest, whether validation was undertaken, and the code/software used.

In the following, I review and summarise these papers in detail around their model specification, estimation of dynamic predictions and model validations conducted. Where available, the corresponding software and code with packages can also be found [103,104]. Nine papers (75%) applied the SPJM framework, three of these presented the standard joint model for a time-to-failure endpoint in section **3.4**. Three papers (25%) described the JLCM approach, presented in section **3.5**. In section **3.6**, 6 of the 9 SPJM papers presented extensions to the time-to-event submodel incorporating cure, competing risks, and multi-state models for localised prostate cancer. I finish this chapter with a discussion (section **3.7**) and conclusion.

Table 3-1 a summary table of the articles where joint modelling was applied to localised prostate cancer to predict recurrence, in chronological order. Abbreviations include: JLCM = joint latent class model, SPJM = shared-parameter joint model, PSA= prostate-specific antigen, EBRT = external-beam radiotherapy, HT = hormone therapy, Gy = Gray, EPOCE = expected prognostic observed cross-entropy, CVPOL = cross-validated prognostic observed log-likelihood, IBS = integrated Brier score.

Paper [ref]	Modelling Framework	Sample sizes (N) & Events (E) for model development	Joint model parametrisation	Dynamic prediction landmark and prediction window	Validation undertaken	Code & software used
1) Pauler & Finkelstein, 2002 [105]	Bayesian change-point SPJM.	N=676, E=176	PSA data during the first two years was dropped from analysis due to rapid drops of PSA post-EBRT & HT. The random effects include the intercept and the slopes (before & after the change-point). The change-point is used to predict recurrence. Logged-PSA is modelled with covariates age, presenting PSA, T-stage, with a change-point indicator.	A change-point occurring within 10 years. Relapse landmark by four years with a prediction horizon of 10 years.	None performed.	C routine *dfpmin*, and S-PLUS *surv.fit* function.
2) Law et al., 2002 [106]	Frequentist cure SPJM.	N=458, E=92	Two models are fitted, joint-cure and logistic-Cox (no longitudinal PSA consideration). Nonlinear exponential- decay & growth modelled longitudinal logged-PSAs using presenting PSA, T-stage, and Gleason.	Not specified, estimated probabilities of recurrence are given for each patient at some time in the future.	Simulation study performed showing that joint-cure model has better sensitivity and discrimination compared to logistic-Cox model.	MATLAB

Paper [ref]	Modelling Framework	Sample sizes (N) & Events (E) for model development	Joint model parametrisation	Dynamic prediction landmark and prediction window	Validation undertaken	Code & software used
3) Yu et al., 2004 [73]	Cure SPJM (comparing Bayesian and Frequentist).	N=458, E=92	Modelled the current PSA value & slope trajectory. Random effects are modelled parametrically by exponential- decay & growth models adjusting for presenting PSA, T-stage, and Gleason.	Not specified, estimated probabilities of recurrence are given for each patient at some time in the future.	Not performed, comparisons are made between the two estimation methods and are shown to be similar to one another.	MATLAB & C++
4) Taylor et al. 2005 [107]	Bayesian cure SPJM.	N=934, E=140	PSA value & slope and a time-dependent hormone therapy commencement indicator is considered, adjusting for baseline covariates: presenting PSA, T-stage, Gleason, age, total dose (Gy), and treatment duration.	Landmarks from last contact, with a prediction window of four years.	Validation is performed on data of the same patients used for development, but with further follow-up. The model is shown to be well calibrated and accurately predict new PSA values and recurrence risk.	C++
5) Yu et al., 2008 [86]	Bayesian cure SPJM.	N=928, E=146	PSA value & slope and time-dependent hormone therapy commencement indicator is considered, adjusting for baseline covariates: presenting PSA, T-stage, Gleason, age, total dose (Gy), and treatment duration.	Landmarks from last contact, with a prediction window of four years.	Validation performed on data of the same patients used for development, but with further follow-up. The model is shown to be well calibrated and accurately predict new PSA values and recurrence risk. Kaplan-Meier plot shows the higher predicted risks go on to have more recurrences indicating its validity.	C++

Paper [ref]	Modelling Framework	Sample sizes (N) & Events (E) for model development	Joint model parametrisation	Dynamic prediction landmark and prediction window	Validation undertaken	Code & software used
6) Proust-Lima & Taylor, 2009 [84]	Frequentist JLCM.	*Model development and validation:* N=2,386, E=317	Baseline covariates included: presenting PSA, T-stage, and Gleason. The main and random effects are of the biphasic initial decline and long-term rise. Five latent classes were identified.	Landmarks taken at every six months from 1—3½ years, with a prediction window of three years.	External validation of prediction is performed on two external cohorts. A range of models are explored, the 5-JLCM shows consistently lower absolute and weighted prediction errors in both cohorts, using prediction windows of one and three years.	Not stated but presumably **R** using the *lcmm* package.
7) Jacqmin-Gadda et al., 2010 [77]	Frequentist JLCM.	N=459, E=74	Similar to [84] with biphasic longitudinal components for the logged-PSA, considering presenting PSA, T-stage, and Gleason. Four latent classes were identified to be best fitting where the proposed score test did not reject the null of conditional independence.	Only mean evolutions for each of the four classes are given with predicted recurrence-free survival. No windows are specified.	Simulation study performed to appraise score test, where the baseline hazard function was misspecified. This methodology was applied to a prostate cancer cohort.	Not stated but presumably **R** using the *lcmm* package.

Paper [ref]	Modelling Framework	Sample sizes (N) & Events (E) for model development	Joint model parametrisation	Dynamic prediction landmark and prediction window	Validation undertaken	Code & software used
8) Taylor et al. 2013 [108]	Bayesian SPJM.	*Model development and validation:* N=3,232, E=458	Covariates include presenting PSA, T-stage, and Gleason grade. Longitudinal parameterisation includes PSA value & slope, and time-dependent HT.	Landmarks are given from most recent PSA values with a prediction window of three years.	External validation is performed on a fourth dataset. Simpler visual approaches are undertaken, focusing on estimated risk of recurrence three years after treatment using a three-year prediction window. Patients are assigned to four risk groups, comparing the training and testing Kaplan-Meier plots. A sensitivity analysis is performed commencing hormone therapy as either censored and as an event. The model is deemed adequately calibrated with similar patterns being exhibited between training & testing datasets.	C

Paper [ref]	Modelling Framework	Sample sizes (N) & Events (E) for model development	Joint model parametrisation	Dynamic prediction landmark and prediction window	Validation undertaken	Code & software used
9) Proust-Lima et al., 2014 [66]	Frequentist JLCM & SPJM.	*Model development and validation:* N=1,178, E=200	Biphasic mixed-effect parameterisation of longitudinal logged-PSA. Baseline covariates: presenting PSA, T-stage and Gleason Four latent classes identified for the JLCM, SPJM included PSA value and slope association structure. All other components had the same model structure for direct comparison.	Landmarks taken at every six months from 1—3½ years, with a prediction window of three years.	Internal and external validation is performed. The EPOCE is estimated internally using CVPOL from 1 – 6 years after EBRT. The difference in EPOCE for 4-JLCM and SPJM shows the 4-JLCM to be a better prognostic model in the first four years. External EPOCEs and IBSs are shown over the follow-up period. The IBSs and EPOCEs show reduced errors for ≥ 3-JLCM and SPJM with little differences between the two approaches.	R: using the *lcmm* and *JM* packages – code is available on request from the authors.
10) Sène et al., 2016 [109]	Frequentist SPJM.	N=2,386, E=312	Similar to [84] with biphasic longitudinal components for the logged-PSA, considering presenting PSA, T-stage, Gleason, and corrected total EBRT dose (Gy). Several specifications of the time-dependent initiation of salvage HT, and the association structures of the longitudinal value and slope of PSA and random effects.	Landmarks from 1.2, 1.6, 2 and 2.6 years are given with a prediction window of recurrence within the next three years. The predicted recurrence probabilities are given under four scenarios of initiating salvage HT immediately, in 1 or 2 years, or not at all.	Internal validation is performed using cross validation for a prediction window of three years. The CVPOL, CV-BS, and CV-IBS are shown for the six model structures. A simpler random effects joint model is best and chosen for the absence of salvage HT; for immediate HT, the JM that separated the PSA trajectory before and after HT is deemed best.	R: *JM* package with modifications to source code.

Paper [ref]	Modelling Framework	Sample sizes (N) & Events (E) for model development	Joint model parametrisation	Dynamic prediction landmark and prediction window	Validation undertaken	Code & software used
11) Ferrer et al., 2016 [110]	Frequentist multi-state SPJM	N=1,474; E=941* * sum of all events	Similar to [84] with biphasic longitudinal components for the logged-PSA, considering presenting PSA, T-stage, and Gleason. The longitudinal PSA value and slope was modelled.	For the multi-state process, transition probabilities are given from each transition to any of the other four transitions from the end of treatment throughout follow-up.	A simulation study was undertaken to ensure the estimation process was correct. Diagnostic plots for the residuals and observed/predicted of the longitudinal model.	R: *nlme*, *survival*, *mstate* and *JM* packages, with adaptations. Code is readily available at the author's GitHub account.
12) Ferrer et al., 2018 [111]	Frequentist landmarking and cause-specific SPJM.	Not explicitly stated but assumed as in [110]. N=1,474; E=unknown.	Longitudinal logged-PSA modelled similarly as to [110]. Adjusting for: dataset cohort, age, T-stage, Gleason and presenting-PSA.	Predicted recurrence and competing risk of death probabilities for two patients at their landmarks of 1.3 to 2.5 years using a prediction window of 1½ and 3 years, comparing predictions from the SPJM and landmark modelling.	A simulation study is performed using the prostate patient cohorts to generate similar data. Evaluating robustness of JMs and landmark models, under different assumptions. JMs are generally more robust to deviations in assumptions than landmark models, other than a strong violation in the longitudinal PSA biomarker specification where the landmark model performs better.	R: *nlme*, *JM*, *survival, pseudo*, and *geepack* packages. Code is readily available at the author's GitHub account.

3.4 Shared-parameter joint models to predict recurrence in localised prostate cancer

In this section, I focus on three relevant papers that investigated PSA dynamics to predict recurrence in localised prostate cancer using the SPJM framework: Taylor et al. [108], Sène et al. [109], and Pauler & Finkelstein [105]. All three articles develop models in localised prostate cancer patients treated with EBRT in the absence of neoadjuvant hormone therapy. Taylor focused on developing a model and creating a clinical prediction tool online; Sène explored the effect on initiating salvage treatments at different time points and its effect on the predicted dynamic probabilities of recurrence. Pauler & Finkelstein used a change-point model to capture any jump in PSA.

3.4.1 Model specification

In Taylor et al. [108], the functional form over time of the longitudinal PSA mixed model assumes three phases: baseline/presenting PSA (B_0), and the short-term (decrease, B_1), and long-term (increase, B_2) evolutions of PSA, $Y_i(t) = \log[\text{PSA}_i(t) + 0.1] = B_0 + B_1 f_1 + B_2 f_2$, with $f_1 = \{(1 + \text{time})^{-\frac{3}{2}} - 1\}$ and $f_2 = \text{time}$. For each of the three phases, $B_{k=\{0,1,2\}}$, are matrices containing linear combinations of the fixed baseline covariates T-stage, Gleason grade and presenting pre-treatment PSA, along with subject-specific random effects parameters. A t-distribution with five degrees-of-freedom for the error term is assumed. Time to prostate cancer clinical recurrence is measured from the end of radiotherapy. In the survival submodel, the functional form $f(\boldsymbol{M}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha})$ is a linear combination of the value of PSA concentration and its slope at time t, $f(\boldsymbol{M}_i(t)\, \boldsymbol{b}_i, \boldsymbol{\alpha}) = \alpha_1 \text{PSA}(t) + \alpha_2 \frac{d\ \text{PSA}(t)}{dt}$. Additionally, the survival submodel included a time-dependent indicator variable for when salvage hormonal treatment (ST) is initiated to account for the subsequent drop in hazard of clinical failure. PSA values after ST were excluded due to the sudden decrease in PSA trajectory and did not feature in the mixed-effect model; however, clinical recurrences after ST were considered. A piecewise constant function is assumed for the baseline hazard.

Sène et al. [109] made similar modelling assumptions as Taylor et al. [108] for the functional forms in the mixed and survival submodels. The model adjusted for presenting PSA, Gleason score, T-stage, and total corrected dose of EBRT (in Gy), using the linear-quadratic model given in [112]. Initiation of ST was included as a time-dependent indicator variable to reflect

the potential decrease in risk of progression; five functional forms of ST were considered. Three different association structures of f were fitted: a linear combination of PSA value and gradient (with and without a logistic transformation for PSA), and the random effect structure, which evaluated the individual deviations from the overall population's PSA trajectories. A combination of those different parametrisations yielded 12 models with varied complexity.

Pauler & Finkelstein [105] proposed a change-point parameterisation in the longitudinal model for PSA, by incorporating an unknown change-point indicator variable for whether change in PSA has occurred. If a shift is indicated, a likely change-point time-range is estimated with a uniform distribution for PSA. A narrower posterior change-point range with larger differences in the slopes before and after the change-point indicate prostate cancer recurrence is likely (before the formal clinical failure endpoint). Trivariate normal and uniform priors are used for four random effects, which included: intercept; change-point time (uniform); the slope before and after the change-point. For the survival submodel, a piecewise exponential hazard function was used. Baseline covariates included age, presenting PSA, and disease stage. For the joint model, non-informative priors were chosen.

3.4.2 Estimation, prediction and validation

In Taylor et al [108], estimation was undertaken under a Bayesian framework using C software. The joint model was developed on three pooled cohorts (totalling $N = 2{,}386$ patients) and tested externally using a separate fourth dataset ($N = 846$ patients). Dynamic predictions for an individual patient's PSA trajectory and risk of recurrence for the next three years were shown: no formal validation measures were presented. The authors opted for simpler graphical inspections to study the model, owing to the complicated nature of the time-dependent ST events within the validation cohort. An online prognostic calculator was developed, enabling individual dynamic predictions of disease recurrence given PSA trajectories for future patients (http://psacalc.sph.umich.edu[1]).

In Sène et al [109], estimation was undertaken under a frequentist framework, and **R** software used for model development, again using the same three cohorts as in Taylor et al [108]. Internal approximated leave-one-out cross-validation was used to assess six of the 12 models'

[1] Last accessed in March 2023.

predictability, using BS and EPOCE accuracy measures [101]. The two best fitting models were the logistic-transformed PSA value and slope that separated the effect of PSA before and after ST, whilst the model with the random effect association structure performed best when assumed that the patient would not start ST within three years. Exemplar individualised dynamic predictions used a prediction window of three years on an intermediate risk patient. Different scenarios when ST would be initiated were used to illustrate the impact of delays in ST initiation on risk of recurrence. External validation was not performed.

One cannot make direct comparisons between the predictive performances of the two papers as they used different assessment methods (graphical approaches in Taylor, EPOCE & BS presented in Sène). In Sène et al., patients who did not receive hormone therapy nor ST within three years were mainly used in order to assess predictive performance. Sène noted that this may not be a representative situation for all patients, so they performed a sensitivity analysis using Taylor's approach to widen the sample on hormone therapy-free patients at the landmark prediction time only, then with subsequent ST initiation within the three-year prediction window, as either a recurrence event or dependent censoring. The relative predictive performance was largely unchanged in both papers under this approach and therefore can be considered robust.

In Pauler & Finkelstein [105], estimation was done in a Bayesian framework, using C and S-plus software. The joint model was developed on a cohort of $N = 676$ patients. As the majority of patients do not exhibit clinical failure, the slope after the change-point was non-significantly negative, indicating PSA trajectories generally remain constant over the follow-up period. The regression coefficients from the relative risk component are not straightforward to interpret due to the number of pairwise and three-way interactions. The authors noted that coefficients are in the expected directions. Sensitivity analysis was done on three differing definitions of recurrence based on PSA rises. They showed that regardless of rule followed, there was little difference to their optimal joint model. The AIC rose when considering only a relative risk model using indicator covariates for each rule, this provided justification on using the joint change-point model, as the longitudinal PSAs substantively improve the goodness-of-fit. The posterior distributions of four individual patient change-points were shown. For two patients who do relapse, sharp change-points are given between 2-4 years, who then go on to recur at

six and four years of follow-up. For stable PSA patients, the change-point is imprecise with very wide uniform posteriors. Individualised predictions are performed on two hypothetical patients showing each's posterior predictive distributions of time to relapse. Although discussed, the model was not validated.

3.5 Latent class joint models to predict recurrence in localised prostate cancer

In this section, I focus on relevant papers that investigated PSA dynamics using the JLCM framework. There are three papers of interest: by Proust-Lima & Taylor [84], Jacqmin-Gadda et al. [77] , and a third paper by Proust-Lima et al. [66], which is appraised separately in section **3.5.1,** as it compares the SPJM and JLCM.

Proust-Lima & Taylor [84] modelled the functional longitudinal PSA similarly to Taylor et al. [108] (described in section **3.4**). Baseline covariates T-stage, Gleason score, and pre-treatment PSA were included into both submodels. The survival submodel also includes an exogenous time-dependent indicator variable for initiation of ST, and a class-specific Weibull baseline hazard function.

Model development was performed on a single cohort of patients ($N = 1{,}268$), and external validation was performed on two additional smaller cohorts (with $N = 503$ and $N = 615$ patients respectively). Several JLCMs were fitted with ranging classes ($2—6$), with the five-class model (5-JLCM) producing the lowest Bayesian Information Criterion (BIC); the optimal model included estimation of 75 parameters. Predicted PSA evolutions and survival curves for each of the five classes illustrate how PSA trajectories with long-term rise of PSA correspond to greater risk of failure. Dynamic predictions were made within a prediction window of three years for two patients with contrasting baseline risk factors: a lower-risk patient who recurs and a higher-risk patient with no observed recurrence.

Within each external validation cohort, measures of predictive accuracy (weighted absolute prediction errors, WAEP) for the five-class JLCM were computed and compared to a relative risk model with baseline information only, and a two-stage landmark model. The JLCM was shown to be the best fitting at various landmark times, and accounting for the longitudinal biomarker reduced both the EP and WAEP, particularly at earlier landmarks.

For Jacqmin-Gadda et al. [77], the score test methodology (introduced in section **2.3.2)** is applied to develop a prognostic joint model for prostate cancer recurrence (with the same dataset used as in Taylor et al., [107]). They develop the JLCM similarly to Proust-Lima et al., [63,84]. They show that the more flexible 4-class JLCM did not reject conditional independence, whereas the less powerful alternative Wald test for dependence failed to reject the null for a 3-class JLCM.

3.5.1 Comparison between latent-class and shared-parameter joint models

A direct comparison is made between the two types of joint models applied to prostate cancer by Proust-Lima and colleagues [66]. Three prognostic baseline factors were adjusted for, logged initial-PSA, T-stage, and Gleason score using the same Michigan hospital cohort dataset. The three-component parameterisation of PSA in the mixed-effect model was done in the same manner to Proust-Lima & Taylor, and Taylor et al. [84,108] for both joint models for direct comparison. The developed 4-JLCM adjusting for PSA value and slope was chosen from information criteria and conditional independence being met. The BIC favoured the 4-JLCM compared to the shared-parameter JM.

For direct comparisons between the JLCM and SPJM, evaluation of dynamic predictions (for the entire follow-up) is made using the cross-validated EPOCE framework in the first six years. The 4-JLCM is superior to the SPJM in the first four years on internal validation, and slightly better in the first three years on external validation.

3.6 Extensions to the shared-parameter joint model

Here I present some further extensions to the joint model in the following subsections. In particular, I comment and review four papers with a cured fraction [73,86,106,107]; a competing risk joint model [111], where clinical recurrence is competing with a death unrelated to prostate cancer; and a multi-state joint model [110], whereby patients can go through a pathway of disease states throughout their follow-up.

3.6.1 Joint-Cure models

There can be a high proportion of patients who are recurrence-free after long follow-up, resulting in heavy censoring. This may compromise the predictions of a joint model given the

lack of events observed. It therefore may be appropriate to model these patients who appear to have prolonged event-free survival as 'cured', using a cure joint model. This is a natural extension to the SPJM is to incorporate a cure component to the time-to-event submodel, whereby patients are considered to be susceptible to experience the event under study (e.g. recurrence), or, on the contrary, to be cured after initial treatment, and thus never susceptible to recurrence. Allocation into the two groups is typically modelled using a logistic classifier submodel:

$$\Pr(D = 1|\, X_i) = \frac{\exp(\boldsymbol{\beta}^T X_i)}{1 + \exp(\boldsymbol{\beta}^T X_i)},$$

where $D = 1$ refers to the susceptible group (observed only when the event of interest occurs), X_i is the fixed baseline design matrix with their corresponding vector of coefficients, $\boldsymbol{\beta}$. Patients who have been allocated to the 'cured' group are coded $D = 0$.

There are four articles that consider a joint-cure model for the risk of clinical recurrence [73,86,106,107]. The four papers have a similar model specification: a nonlinear parametric exponential decay-growth (U-shaped) model is used to capture the log PSA trajectory $m_i(\mathrm{t}, \mathbf{r}_i) = r_{i1}\exp(-tr_{i2}) + r_{i3}\exp(tr_{i4})$, where $r_{i1,\ldots,4}$ are the random effects to be estimated. Those that have been allocated to the cure group (from the logistic incidence submodel) have $r_{i4}|(D = 0) = 0$, as this assumption reduces the PSA trajectory, $m_i(t, \boldsymbol{r}_i)$, to an exponential decay cure SPJM. The conditional failure time model is given by $h\big(t\big|D_i, X_i, \boldsymbol{r}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, g(m_i)\big) = h_0(t)\exp\big(\boldsymbol{\beta}^T X_i + \boldsymbol{\alpha} g(m_{i|D}, t)\big)$, where $g(m_i)$ can be given by including the PSA trajectory function and its slope given in [86,107].

Baseline covariates included pre-treatment PSA, T-stage, and Gleason score. Additionally Taylor et al. [107] considered PSA value & slope as time-dependent covariates, age, EBRT total delivered dose (in Gy) and treatment duration as baseline covariates. Yu and colleagues [86] included an exogenous time-dependent variable to indicate start of salvage hormone therapy, similarly to [108], and used a generalised Weibull model for the baseline hazard function. Both frequentist and Bayesian approaches are directly compared by Yu et al. [73].

In Law et al. [106], the joint-cure model is compared to the standard cure model without longitudinal time-dependent information, and to the shared parameter joint model without

the cure component. They showed better predictions and discrimination, together with reducing biases from informative censoring. Yu et al. & Taylor et al. [86,107] compared the predictions of the model with updated information on the same patients who were initially used to develop the model, that is whereby more longitudinal PSAs and events on the same patients are gathered.

The extended shared-parameter joint-cure model offers additional flexibility to model the inherent heterogeneity of patients who go on to have extended event-free survival. Yu et al. [86] directly compared joint models with and without a cure component. They showed a standard JM tends to overestimate the number of clinical events. They compared the two models using the conditional predictive ordinate and BIC, both favouring the additional cure submodel component, despite an extra 30 parameters needed to be estimated [86]. This however may over-parameterise the model without adequate event sizes [113]. Also, as the prostate cancer disease pathway is complicated, clinical input is recommended with regards to plausibility of the cure component and its definition.

3.6.2 Competing risks joint models

The event of interest may be precluded by the occurrence of a competing event, for instance, a non-cancer related death, before recurrence, being observed. It is well known that biases are elicited by censoring these competing event deaths [114,115]; joint models can be extended to consider the presence of a competing event.

Ferrer et al. [111] presented individual dynamic predictions and validated the robustness of the estimators in the presence of competing risk of death (from a non-related cancer cause), within a frequentist framework. A cause-specific proportional hazards submodel was proposed for each competing event, and thus the relationship of the longitudinal biomarker with each competing event can be assessed. Individual dynamic predictions were estimated and compared to landmarking estimators. Two simulation studies were performed using simulated data that was alike to the applied prostate cancer dataset. Each approach validated the estimators, then compared and assessed their robustness to misspecification of the joint model. Both the AUC and mean-squared prediction error were employed to characterise the predictive accuracy. An extension of the AUC was adapted to the competing risk setting, proposed by Blanche et al. [88]. It was shown that in almost all cases, the joint models were

superior to the landmark models. The landmark models were only superior to the joint models when the longitudinal biomarker was heavily misspecified. Ferrer's competing risk paper is the only study to present validation metrics, using simulated studies. Code is available at https://github.com/LoicFerrer/Individual-dynamic-predictions [2].

3.6.3 Multi-state joint models

The evolution of localised prostate cancer over time can be characterised by the occurrence of different events of interest, such as biochemical failure, local recurrence, distant recurrence, and death. One way to jointly model all these events is via multi-state models, in which the occurrence of the events of interest define the transition between different disease states [116]. As longitudinal processes such as PSA trajectories can have an impact on several of these event transitions, multi-state models can be generalised to the joint modelling framework.

Ferrer et al. [110] proposed modelling the longitudinal PSA process using a mixed-effect submodel, similarly to Proust-Lima and Taylor [84], Sène et al. [109], and Ferrer at al. [111]. They used a non-homogeneous Markov multi-state model for the intensity of the transitions between five states: 0) end of EBRT treatment, 1) local recurrence, 2) salvage hormone therapy, 3) distant recurrence, and 4) death (absorbing state). Intermediate states could be skipped (e.g. ending $EBRT_0$ → $death_4$), and backward transitions were not allowed. Two properties were considered: 1) the Markov property whereby the future process is only dependent on the present state and not the preceding transitions / states; 2) the non-homogeneous property ensures the time since entering the study influences the evolution of the process.

Each transition intensity was modelled assuming proportional hazards and incorporated the biomarker trajectory. For each transition from state i to j, only patients visiting the state i were included in the analysis. The baseline intensity function was modelled parametrically. The maximum likelihood framework was used to estimate the corresponding parameters.

The multi-state joint model was fitted with the same two study datasets as in Ferrer et al. [111]. Four covariates (presenting PSA, Gleason score, T-stage, and study cohort) were adjusted for in the models. A linear combination of PSA value & slope were used.

[2] Last accessed March 2023

Predictions were compared with the observed data. The observed values were averaged at each decile with corresponding predicted values computed, they show the observed values lay within the 95% CIs, with very similar predicted values. The predicted transition probabilities over time, in a given state to another other feasible state are presented, comparing similar parametric estimated probabilities to the observed. The only exception was between transitions 1→2 (from local recurrence to receiving hormone therapy) where the spike after EBRT was not adequately captured with the splines, it shows there is a very near-immediate initiation of hormone therapy after localised recurrence to control the disease. It is worth noting that PSA dynamics were only collected until the patient's first clinical event and thereafter were extrapolated according to their posterior trajectories.

Diagnostics of the joint multi-state model were evaluated visually. Residuals *vs* fitted values, observed and predicted PSA trajectories, and predicted *vs* non-parametric transition probabilities between states were presented. In general, they showed the model fits particularly well to the longitudinal, and multi-state submodels. The models themselves were not externally validated nor stated any predictive performance measures, only the estimation process via simulation studies. Although equations for obtaining individual dynamic predictions for patients were presented in the paper, these were not demonstrated with specific examples.

The code to apply these multistate models to a simulated dataset and adapt for use is freely available at https://github.com/LoicFerrer/JMstateModel[3] and could be used to derive patient predictions and be adapted for the reader's need.

3.7 Discussion

Over the last two decades there has been a plethora of research on PSA concentration and its association to recurrence, or prolonged event-free survival (effectively cure). In this chapter, twelve papers were reviewed and assessed, which report on joint models of longitudinal PSA trajectories and time-to-event endpoints that aim to describe how these trajectories impact and predict prostate cancer recurrence. I identified two broad frameworks (SPJMs & JLCMs) that were used in the identified papers. I reviewed and synthesised the methodologies of these two

[3] Last accessed in March 2023.

different approaches, applied to similar dataset cohorts of prostate cancer patients receiving EBRT without hormone therapy, which allow the methodology to be compared.

Due to the long-term nature of prostate cancer recurrence and progression, the datasets to develop the CDPJMs comprise patients treated in the 1980s. As long-term follow-up is necessary, the historical nature of the datasets is unavoidable but the impact of changes in clinical practice should be considered when utilising CDPJMs for contemporary patients.

There are limitations to this work, as this report was not initially intended to be a systematic review on all the available literature, but a synthesised summary of what I considered relevant articles of modelling both PSA longitudinally, and time-to-recurrence in localised prostate cancer, in preparation to apply these methods in my own application (**Chapters 4, 5, & 6**). For instance, I focused on specific key words within the title and abstract only, so there may have been missed reports if the use of these terms was not explicit in these fields. Further joint modelling papers not included here were due to, for instance, no dynamic predictions presented [117], a mix of non-radiotherapy treatments (e.g. radical prostatectomy); methodology development focused but repeated analysis referred to [100,118,119]; or exclusive use of simulated datasets [120]. It was noted that not all papers were expectedly populated by the search strategy [121]. In the localised prostate cancer setting, where PSA is used to monitor recurrence after radical treatment of disease, joint models have also been used in the context of prostate cancer screening [122–127] or advanced (metastatic) disease [61,128–132], however, as PSA dynamics differ greatly from localised disease, these scenarios were not considered. Joint models could also be extended to accommodate multiple longitudinal biomarkers, such as PSA and testosterone, or the sequential findings on MRIs, in a joint multivariate model [69,133,134].

Modern typical first-line treatment of localised prostate cancer includes hormone therapy before (neoadjuvant) and concurrently with EBRT [13,46], and PSA trajectories are known to be more homogeneous with combined treatment [108]. Furthermore, given recent advances in radiotherapy techniques and the use of moderate- and ultra-hypofractionation (fewer but larger doses of radiotherapy) [22,135], treatment exposures of radiotherapy are less than the average treatment durations presented in these papers. The tool in Taylor et al. [108], reviewed in section **3.4**, was developed in the absence of neoadjuvant hormone therapy, therefore

predictions from these models have limited applicability within current treatment pathways. Further development of these models for patients receiving hormone therapy are needed.

The papers reviewed provide a very good exposition and rationale to their model development and clinical usage. Regardless of the functional form used in the joint modelling framework, a fully parametric form was fitted for the mixed-effects model. There are possibly more appropriate and flexible forms that may exist, compared to the biphasic form for PSA trajectories they postulate throughout [63,66,77,84,108,109,111]. Many of the reviewed articles present an appraisal of their models, either by validation or contain a simulation study. External validation is seen as the gold standard, to ensure model suitability and generalisability in other patient populations and to assess overfitting [43]. However, when rigorous measures of predictive performance have not been reported in these papers, these would not be considered well validated by today's standards, for example (not) using the TRIPOD framework [40].

As with any specification of modelling, there are advantages and disadvantages to the joint modelling approach taken and several differences exist. For JLCMs, the maximum likelihood approach contains closed-form solutions and are computationally feasible to obtain. They are advantageous for the use of developing a predictive joint model for dynamic predictions, whilst not having to impose specific parametric assumptions for the biomarker's functional form (e.g. current value, slope, area), unlike SPJMs [66]. Robustness to deviations of the imposed functional form have been rigorously assessed in Ferrer at al. [111]. In their paper, they demonstrated that no method (joint modelling nor landmarking) was particularly robust to misspecification in the longitudinal biomarker. However, when there was heavy misspecification, landmarking methods outperformed joint modelling.

The SPJMs assume a homogenous population with a singular average PSA biomarker trajectory, whereas JLCMs account for further population heterogeneity through the latent classes. Both JLCMs and SPJMs account for the variability of the PSA biomarker through the random effects in the longitudinal submodel. The purpose of the random effects in the SPJM is two-fold, accounting for the correlation of the repeated measures in the mixed-effect model, and the association between the PSA biomarker and time-to-recurrence, whilst in the JLCM only the latent classes account for the association between the biomarker and event.

The disadvantages of the JLCMs approach include the possibility of having multiple local maxima for the maximum likelihood estimates, and several models are needed to be fitted in order to find the optimal number of latent classes (by comparing multiple information criteria) [67]. Some of these issues can be circumnavigated via parallelisation of the computation for more optimal resourcing, for example making use of parallel computing by using search grid methods for JLCMs as computations are independent (see the mpjlcmm function from **R** package *lcmm*); or implementing multiple MCMC chains performed in parallel using Bayesian SPJMs (jm function from **R** package *JMbayes2*) [103,136].

Both Frequentist and Bayesian paradigms were used for the SPJMs, whereas for the JLCMs only frequentist methods were reviewed. In their direct comparison of JLCMs and SPJMs [66], the authors showed that the JLCMs had less assumptions and performed better. However when adjusting for the same patient cohort dataset, baseline covariates, prediction times, and biphasic components for the longitudinal PSA component, the prognostic accuracy measures for EPOCE in Sène et al. [109] using SPJMs appear superior than those obtained with the JLCM in Proust-Lima et al. [66].

All models reviewed in this paper can produce dynamic predictions for prostate cancer prognosis. The JLCMs do not assume a specific association structure nor quantify those associations (like the SPJMs do); they describe the trajectories in a heterogeneous population. If the main goal is to quantify the associations assuming a homogenous population, then SPJMs are recommended. There is not one overarching or standout framework to always use by default. The choice of model may be primarily driven by the research question and personal choice. If the purpose is solely for prediction, then combining several frameworks for dynamic predictions using some weighted model averaging methodology could be applied [137]. Indeed, one type of framework may outperform another at certain time intervals and then vice-versa at different time windows. Each model has its own advantages, depending on the end goal of the reader. It is hard to compare each model's framework with another in terms of superior predictive performance as not all these papers present these metrics.

This review focused on radiotherapy, however there are other treatments for prostate cancer including (neoadjuvant) hormone therapy, prostatectomy and combinations therein, though optimal timing of these combinational therapies appears unclear [138,139]; Sène addressed

optimal initiation of ST [109]. There have been recent advances in using sophisticated machine learning/artificial intelligence (ML/AI) techniques on imaging data to predict whether patients require biopsies, or to predict clinical failure or death under these alternative treatment pathways. Some recent articles include development of artificial neural networks, support vector machines, and random forests for predicting diagnoses [140,141], optimal timing of biopsies [142] , and clinical failure [143] or death [144]. However, it is not apparent that the longitudinal nature of time-varying markers like PSA have been considered, nor produce dynamic predictions. A review of these AI and ML methods is given in Tătaru [145]. Some authors refer to joint modelling itself is an AI approach [146]. Other studies have suggested combining the boosting approaches of machine learning to joint models, to create a unified framework using mechanistic data-driven approaches [147]. ML/AI techniques are not a panacea and need to be correctly developed and incorporate all available information, be rigorously validated, and to have clinical utility [148–151]. Reporting guidance, based on TRIPOD & PROBAST statements, have been developed for AI & ML (TRIPOD-AI/ML & PROBAST-AI/ML) [152–155]. A further discussion on the future of AI & ML techniques and examples can be found in **Chapter 7**.

The aim of this literature review chapter and publication [41] was to inform the modelling procedure in the subsequent results chapters using data from the CHHiP trial. The reviewed articles are informative of the joint modelling framework, i.e., a frequentist or Bayesian approach (which was compared at length by Yu et al. [73]); the parameterisation of each of the submodels for the joint model, using either a shared-parameter or latent class joint model. The Bayesian shared-parameter joint model will be used throughout this thesis due to a relatively homogenous patient population and inclusion criteria of CHHiP. Additionally, this framework allows one to calculate the strength of association of the structure via the $\boldsymbol{\alpha}$ log-hazard ratios.

Pertaining to the Bayesian / frequentist approaches, there was little difference in the parameter estimates between the expectation–maximisation likelihood and MCMC approaches for the joint-cure model in [73]. This was confirmed in the data in a preliminary sensitivity analysis, where I fitted a joint model with both approaches and no large differences in the estimated parameters were found [156].

In a Bayesian framework, instead of estimating an unknown *fixed* parameter (as is in a frequentist approach) the parameters are treated as *random variables* where its distribution is explored (using MCMC), and probabilistic statements about these unknown parameters can be made on its posterior distribution, i.e., 95% credible intervals is interpreted as a 95% probability of including the true parameter. Confidence intervals do not have this interpretation, that is, over many runs 95% of the computed confidence intervals will contain the true parameter (i.e., in this framework the true parameter either lies within the interval or not, as probabilities cannot be assigned directly to the parameters themselves). Posterior distributions can be directly used to calculate the probability of differing hypotheses, say an increase of survival of x-months/years, or superiority of a treatment over another. When performing inference, this interpretation is possibly more intuitive to one's own probabilistic rationale, rather than under null hypothesis significance testing, where the probability is obtaining the data is as least as extreme as the one observed by chance (p-values) under the null (assuming equal treatments) [157].

Having the posterior distributions and the corresponding random effects available is advantageous as draws from these existing distributions can be sampled from, to elicit dynamic predictions for say a new patient whose random effects are unknown. This process of updating predictions given new data underpins the basis of Bayesian statistics and follows more naturally under this framework. With the utility of Bayesian MCMC approaches, it is also not necessary to use asymptotic approximations, are typically computationally not so burdensome and can overcome convergence, unlike in a maximum likelihood framework.

In my preliminary analyses, I used non-informative priors on all parameters, but there is flexibility to incorporate informative priors if so required (e.g. clinical expertise or by previous experience or study results). Given that the preliminary results were largely similar to the frequentists approach with non-informative priors, I did not feel the need to pursue prior elicitation and, for the advantages outlined above, in the remainder of this thesis I will opt to use a Bayesian approach with non-informative priors.

For the shared-parameter joint model, it is vital to apply the correct, or appropriate, structure of the functional form relating the longitudinal trajectory's impact on the risk of recurrence. The parameterisation of the mixed-effects submodel was done in differing ways, (e.g. the

parametric form of capturing the longitudinal PSA using an exponential decay-growth model, biphasic, or using cubic splines). The latter is advantageous due to capturing the observed patient-specific nonlinear PSA trajectories, which can vary between-patients, without having to impose specific parametric forms over time; these are proposed in [158]. Its flexibility can be controlled by increasing the number of internal knots hence capture further nonlinear evolutions of the biomarker. The optimal number of knots can generally be attained by information criterion [159]. Using natural cubic splines to model the PSA trajectory is the approach that will be taken for the remainder of this thesis, because of its flexibility and expected differing PSA trajectory due to neoadjuvant and concurrent hormone therapy, as well as radiotherapy.

Chapter 4 – Development of a Personalised Clinical Dynamic Predictive Joint Model to Characterise Prognosis for Patients with Localised Prostate Cancer Patients: Analysis of the CHHiP Phase III Trial

4.1 Publications and presentations related to this chapter

A Personalised Clinical Dynamic Prediction Model to Characterise Prognosis for Patients with Localised Prostate Cancer: analysis of the CHHiP Phase III Trial.

Harry Parr, Nuria Porta, Alison C Tree, David Dearnaley, Emma Hall, *International Journal of Radiation Oncology - Biology - Physics* (2023), https://doi.org/10.1016/j.ijrobp.2023.02.022

4.2 Introduction

This chapter will focus on the development of a shared-parameter clinical dynamic prediction joint model using data from CHHiP (**C**onventional or **H**ypofractionated **H**igh dose **i**ntensity modulated radiotherapy in localised **P**rostate cancer), a phase III randomised control trial.

Radical treatment with neoadjuvant hormone therapy and IMRT is less invasive and generally better tolerated in terms of long-term quality of life than radical prostatectomy [160–162]. Following the publication of three randomised controlled trials, including CHHiP, showed that moderately hypofractionated radiotherapy was non-inferior to conventional (2 Gray/fraction) radiotherapy, hypofractionation is now used as a standard of care in Europe and North America [12–14,163,164].

Investigation of pre-treatment prognostic factors of the risk of clinically confirmed biochemical failure have been previously explored in CHHiP [13]. In this work, in addition to (baseline) pre-treatment information, I propose the use of joint modelling methodology [41,55,84,102], as explored in **Chapter 2**. I incorporate PSA values collected over time (the longitudinal process) to obtain updated predictions of the risk of clinically confirmed

biochemical or clinical failure (the time-to-event process), as new information becomes available. This could lead to a more personalised approach to follow-up care and management. For instance, if a patient remains recurrence-free for a prolonged period, and the PSA trajectories would classify the patient as having a low recurrence risk, then a possible recommendation could be to reduce the patient's follow-up schedule, resulting in less burden for both patients and clinics. Conversely, if the patient's risk increases, it may enable the clinician to initiate more intensive follow-up, scans, and/or direct alternative therapies as appropriate [32].

The objective of this chapter is to develop a clinical dynamic predictive joint model utilising longitudinally collected PSAs to then predict the risk of future recurrence in the CHHiP trial. I present dynamic predictions of the developed model on prognosis for two patients contrasting in their baseline prognostic factors, then evaluate the predictions and performance with internal validation by bootstrapping to correct for biases within apparent validation. I then propose PSA thresholds that are indicative of good prognosis, or minimal < 5% risk of recurrence.

4.3 Methods & Materials

4.3.1 Study design & procedure

CHHiP is an international, multicentre, randomised, phase III, non-inferiority trial. Men with localised prostate cancer (T1b-T3aN0M0) and risk of seminal vesical involvement ≤30% were randomised (1:1:1) to receive conventional radiotherapy 74Gy in 37 fractions (f) over 7.4 weeks, or one of two hypofractionated radiotherapy schedules: 60Gy/20f in four weeks or 57Gy/19f over 3.8 weeks. The protocol mandated hormone therapy in men with NCCN intermediate and high-risk disease, for at least 3 months (maximum 6 months) before start of radiotherapy and continued until the end of radiotherapy; this was optional for low-risk patients. Bicalutamide monotherapy or LHRHa plus possible short-term anti-androgen were permitted according to patient and physician's choice. PSA values were recorded pre-hormone therapy and pre-radiotherapy; 12 weeks after initiating hormone therapy; then at weeks 10, 18, and 26 after start of radiation therapy; and then at intervals of 6 months after end of radiotherapy for 5 years; then annually thereafter. The trial was registered

(ISRCTN97182923), approved by the London Multicentre Research Ethics Committee (04/MRE02/10) and by the institutional research board of each participating international site. This study was conducted in accordance with principles of good clinical practice; full details of the trial design have been described previously [13].

4.3.2 Outcomes

Prostate cancer recurrence was defined as the composite of biochemical or clinical failure, or death due to prostate cancer. Biochemical failure was defined using the Phoenix definition of a PSA value $>$ the nadir $+$ 2ng/mL and confirmed by the local investigator [13,18]. Clinical failure included: recommencement of hormone therapy, local recurrence, lymph node or pelvic recurrence, and distant metastases. Time-to-recurrence was calculated as the time between the patient's closest pre-treatment PSA before hormone therapy (time origin $t = 0$), and the first primary endpoint event. The median time between the closest pre-treatment PSA and randomisation was 15 weeks. Patients who were alive and recurrence-free or died due to causes unrelated to prostate cancer were censored at their last known follow-up date, with administrative censoring of longitudinal follow-up taking place at 10 years after time origin.

For the present study, only patients who received hormone therapy and had complete baseline prognostic information, including age; hormone therapy type received; Gleason score (GS); and T-stage; as well as at least one post-treatment PSA longitudinal value, were included in the model. Complete-case analysis was undertaken for the baseline prognostic factors. I based analyses on a data snapshot taken on October 9, 2019.

4.3.3 Specification of the joint model

A Bayesian shared-parameter joint modelling framework was used to develop the CDPJM. I specify a mixed-effects submodel to model PSA trajectories over time, and a Cox hazard submodel to model the time-to-recurrence endpoint. The shared parameters link the two models together, allowing us to quantify how a specific PSA trajectory is associated with risk of prostate cancer recurrence. Model development has been conducted in line with TRIPOD guidance [40]; the TRIPOD checklist can be found in ***Appendix A, Supplementary Table A1***.

4.3.3.1 Mixed-effects submodel

Let $y_{ij} = \mathrm{Y_i(t_{ij})}; \mathrm{i} = 1, \ldots, \mathrm{N};\ \mathrm{j} = 1, \ldots, \mathrm{n}_i\}$ be the i^{th} patient with the j^{th} longitudinal log-transformed PSA measurement at time $\mathrm{t_{ij}}$, where N are the number of patients and n_i are the total number of longitudinal measurements per patient respectively. This process is assumed to follow a mixed-effects submodel, which captures the correlation structure of the repeated measurements, the biological variability, and the unbalanced PSA panel data between patients. It is typically defined by a linear predictor with a mix of time-dependent and independent main and random effects. The design matrix for the main effects is X_i with corresponding parameter estimates $\boldsymbol{\beta}$. The design matrix for the random effects is Z_i, with a corresponding vector of parameter estimate covariates $\boldsymbol{b}_i$ which capture the random intercept and slopes ($\boldsymbol{b}_{ik}\boldsymbol{z}_{ik}^T(t)$).

$$
\begin{aligned}
Y_i(t) = {} & \log(\mathrm{PSA}_i(\mathrm{t}) + 0.1\ \mathrm{ng/mL}) \\
= {} & m_i(t) + \boldsymbol{\epsilon}_i(t) \\
= {} & \boldsymbol{\beta} X_i(t) + \boldsymbol{b}_i Z_i(t) + \boldsymbol{\epsilon}_i(t) \\
= {} & \beta_0 + b_{0i} + \sum_{k=1}^{K} \left(\boldsymbol{\beta}_k \boldsymbol{B}_k(t, K) + \boldsymbol{b}_{ik} \boldsymbol{z}_{ik}^T(t) \right) + \sum_{k=K+1}^{K+9} \boldsymbol{\beta}_k \boldsymbol{x}_i^T + \boldsymbol{\epsilon}_i(t)\,.
\end{aligned}
$$

PSA is log-transformed to conform to the distributional assumptions, with a small offset term added (to avoid the logging of any zeros in the data). The trajectory function is denoted $m_i(t)$, i.e., the true but unobserved PSA value for the i^{th} patient at time t. Values are contaminated with some measurement error, which are time-dependent and are assumed to follow $\boldsymbol{\epsilon}_i(t) \sim N(0, \sigma^2)$.

The nonlinear PSA trajectory over time is captured using natural (restricted) cubic splines $\boldsymbol{B}_k(t_{\mathrm{PSA}}, K)$ with $K - 1$ number of internal knots. Implementing these splines is advantageous as they allow nonlinear PSAs to be flexibly modelled, without imposing specific parametric assumptions, such as the exponential-decay-growth, or biphasic parameterisations [63,107]. Natural cubic splines are local polynomials, which split the follow-up period into a finite number of intervals, and these knots are placed at various follow-up times, with corresponding K $\boldsymbol{\beta}_{k\in\{1,\ldots,K\}}$ coefficients. The number of knots was selected via information

criteria, and the mean squared error (MSE) estimator $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{\text{actual}_i} - \hat{Y}_{\text{predicted}_i}\right)^2$ using K-fold cross validation, over a range of internal knots to assess the bias–variance trade-off, whilst adjusting for baseline covariates. These metrics were compared to select the optimum number of internal knots, subject to the upper boundary at 10 years.

The remaining nine coefficients ($\beta_{K+1}, \dots, \beta_{K+9}$) correspond to the baseline prognostic factors and covariates (like treatment), included in the model as fixed effects: fractionation arm (reference: 74Gy/37f; 57Gy/19f, 60Gy/20f); Gleason score (reference: ≤ 6; 3 + 4, 4 + 3, ≥ 8); T-stage (reference: T1; T2, T3); hormone therapy received (reference: LHRHa; or bicalutamide), and continuous mean-centred age.

The random effects, $\boldsymbol{b}_i$ are included to capture the individual variability of each patient's presenting PSA (random intercept) and deviation over time (random slopes). Deviations of PSA from the predicted trajectory are captured through these random effects. The random effects are assumed to have a multivariate normal distribution with zero mean vector and $q = K + 1$, $q \times q$ variance-covariance matrix D (i.e., $\boldsymbol{b}_i \sim \text{MVN}(\mathbf{0}, D)$). A diagonal covariance structure is opted for in D to aid computation for higher dimensional spline structure in the mixed-effect submodel, $\text{diag}(D) = \{b_0, \dots, b_{K+1}\}$.

4.3.3.2 Cox submodel

In the Cox hazard submodel, it is assumed that the risk of recurrence depends on the trajectory of the longitudinal PSA biomarker. Thus, the trajectory parameterised via the mixed-effect submodel, considering the entire longitudinal history up to a time point t for each patient, is imputed into the Cox parameterisation as a linear predictor, with a specified association structure where the features of the longitudinal PSA biomarker outcome are included:

$$\begin{aligned} h_i(t \mid M_i(t),\, \boldsymbol{w}_i) &= \lim_{\Delta t \to 0} \frac{\Pr\{t \le T_i^* < t + \Delta t \mid T_i^* \ge t, M_i(t), \boldsymbol{w}_i\}}{\Delta t} \\ &= h_0(t)\exp\{\boldsymbol{\gamma}^T \boldsymbol{w}_i + f(M_i(t), \boldsymbol{\alpha}, \boldsymbol{b}_i)\}, \quad t > 0. \end{aligned}$$

The time-to-event process thus depends on $M_i(t) = \{m_i(s),\ 0 \le s \le t\}$, the *true but unobserved* longitudinal trajectory up to time point t. The functional form $f(M_i(t), \boldsymbol{\alpha}, \boldsymbol{b}_i)$ can be parameterised by a range of parametric structures, to assess and quantify the association between the longitudinal biomarker and the risk of recurrence. Typical association structures

feature the PSA value or the PSA rate-of-change/gradient, the accumulated area under the PSA biomarker trajectory, or a linear combination therein - see section **3.4.1** for description of these. For example, the underlying value and rate-of-change association structure of the PSA trajectory (at time t) could be associated with the hazard of recurrence at that same point in time. I considered these various association structures and chose the final structure using deviance information criterion and computational feasibility, whilst also considering intuitive interpretation. I consider a simultaneous linear combination of the value, and gradient of PSA, $f(M_i(t),\, \boldsymbol{\alpha}, \boldsymbol{b}_i) = \alpha_1 m_i(t) + \alpha_2 \frac{dm_i(t)}{dt}$. The inclusion of instantaneous gradient/rate-of-change, as well as PSA concentration, allows comparisons of PSA trajectories between like-for-like patients with similar PSAs at time t, but with differing velocities of PSA. The quantification of the association between the two outcomes is given by $\boldsymbol{\alpha}$, a log-hazard ratio where a unit increase in the log PSA and its gradient corresponds to an increased risk of recurrence by Hazard Ratio $= \exp \boldsymbol{\alpha}$.

A vector of baseline risk prognostic factors is $\boldsymbol{w}_i$ with a vector of corresponding log-hazard ratios $\boldsymbol{\gamma}$. The same baseline covariates as in the mixed-effect model were considered. The baseline hazard is defined to have a fully specified joint distribution. A function for $h_0(t)$ is typically defined parametrically using flexible penalised basis-splines $h_0(t) = \exp\{\psi_{h_0,0} + \sum_{q=1}^{Q} \psi_{h_0,q} B_q(t, \boldsymbol{v})\}$, where $Q = 10$, i.e., a linear combination of B-splines $B_q(t,\boldsymbol{v})$ with q^{th} basis function and a vector of spline coefficients ψ.

4.3.4 Estimation

The CDPJM was developed in **R** software (v4.1.0) using the *JMbayes2* package (v0.1-64–0.2-3) [136]. The individual submodels were fitted using maximum likelihood estimation with the *survival* (v3.2-11) and *nlme* (v3.1-152) **R** packages [165,166]. The fully specified joint model is then estimated in a Bayesian paradigm implementing MCMC algorithm sampling. The final selected association structure parameterisation was obtained by running 27,500 MCMC iterations with a burn-in of 2,500, using default Bayesian priors and control arguments. To ensure convergence, four simultaneous iterative independent chains were used, and the potential scale reduction factor ($\hat{R}$ statistic) was calculated [74]. This statistic evaluates the ratio of the average variance of samples of each chain, to the pooled variance samples; when the

chains have converged to a common distribution, $\hat{R}$ is at unity. It is recommended that all covariates have $\hat{R} < 1.1$ after adequate samples taken, with differing initial values of the chains. Computation was performed on a high-performance computer running CentOS 8 Linux distribution and Windows 10 Intel Core i9-8950HK CPU.

4.3.5 Dynamic predictions

For each individual patient, their longitudinal PSA biomarker values up to the landmark time point t are considered, when it is assumed that they are recurrence-free. The goal is to make predictions about their prognosis, within some clinically relevant prediction window in the future $[t, u]$; $u > t$, say two-, five-, or ten years from present landmark time t. The joint model estimates the probability of recurrence within time u, given the information available up to time t.

The probabilistic statement of their conditional cumulative probabilities of recurrence is $\pi_l(u \mid t)$ for up to the horizon time u is defined by,

$$\pi_l(u \mid t) = \Pr(T_l^* \leq u \mid T_l^* > t;\ y_l(t_{lj}), \mathrm{D}_n, \boldsymbol{\theta})$$

where $\mathrm{D}_n = \{T_i, \delta_i, \boldsymbol{y}_i, \boldsymbol{w}_i; i = 1, \ldots, n\}$, i.e., the sample of data from where the model was estimated, and $\boldsymbol{\theta}$ denoting the fitted joint model parameters**.** As more values of PSA are acquired, the landmark time t_j becomes greater, i.e., $t_1 < t_2 < \cdots < t_{j-1} < t_j$. The posterior probability of recurrence can be *dynamically* updated given new landmarks, or values of t, and the corresponding PSA value at that more recent time point t' are used to obtain the updated conditional posterior probabilities $\pi_l(u|t')$ [55].

4.3.6 Assessing predictive performance and risk thresholds

Predictive performance of the joint model was evaluated at varying landmark times, by assessing its discrimination via time-dependent AUC, and its calibration via the ICI metrics [85,91]. The ICI is the absolute mean difference between the predicted and observed event probabilities. Overall prognostic performance was measured by estimating prediction error (PE) with the Brier score, which is the expectation of the squared difference between the predicted and observed event probabilities, i.e., an overall measure of prognostic performance comprised of both calibration and discrimination [84,94]. Higher AUC metrics indicate

superior discrimination; as the ICI and Brier are both loss functions, smaller measures indicate closer predicted and observed agreement and better model calibration. Internal validation of the proposed CDPJM was pursued by internal bootstrapping (50 repetitions) to account for any over-optimism and misspecification, and to correct biases accordingly. I then compared the CDPJM predictions at future landmark times, with the predictions obtained when no longitudinal PSA biomarker information is available (i.e., at $t = 0$), to assess the improvement that longitudinal PSAs make.

As well as personalised predictions, it is often useful for clinicians to have threshold values of PSA which give acceptable risk profiles following radiotherapy and short-course hormone therapy. Linear regression is used to quantify the association of PSA values from zero (baseline) to five years to correlate to the predicted risk of recurrence by eight years.

4.4 Results

4.4.1 Dataset for model building

The CHHiP trial randomised 3216 participants, of which data from 3071 (95%) individuals were used to develop the statistical CDPJM. There were 104 participants who were excluded who did not receive hormone therapy ($n = 90$) or had missing hormone therapy allocation ($n = 14$); an additional 5 were removed who received maximal androgen blockade; 3 with at least one baseline prognostic factor missing; 9 with no baseline pre-treatment PSA available, and 24 with missing PSA values beyond baseline over time (non-mutually exclusive). **Table 4-1** presents the baseline characteristics of the included patients. The median follow-up of this subset was 8.6 years (IQR=6.3–10.1), with median 9.4 years (IQR=8.3–10.4) for censored patients and 5.5 years (IQR=3.9–7.4) for patients with a recurrence event. The median number of PSA values per patient was 16 (IQR=13—18); censored: 17 (IQR=16–18), recurrence: 14 years (IQR=11–16).

Of the 3071 patients, 607 (20%) had recurrence, a composite endpoint of biochemical ($n = 541$, 18%), clinical failure ($n = 65$, 2%), or prostate cancer death ($n = 1$). There were an additional 148 patients who exhibited PSA values that met the biochemical failure threshold but were not confirmed by a clinician nor subsequent PSA observation. A further 355 (12%) patients died due to causes unrelated to prostate cancer. These patients were censored at the time of

last follow-up for the primary analysis. This outcome was not considered a competing risk as the estimated cumulative incidence function accounting for competing risks and without (1 minus Kaplan-Meier estimate) yielded almost overlapping curves (e.g. the maximum difference was found at 10 years, between 0.216 and 0.23, respectively) [167]. Further consideration of these competing events is focused on in **Chapter 6**.

Table 4-1 baseline characteristics of CHHiP patients (N=3071) considered in model development, stratified by outcome. LHRHa – Luteinizing-Hormone-Releasing-Hormone analogue + possible anti-androgen; [1] n (%); Median (IQR).

Baseline Factors	Overall, N = 3071	Censored, N=2464	Recurrence, N=607
Allocated fractionation			
74Gy/37f	1017 (33%)	816 (33%)	201 (33%)
57Gy/19f	1025 (33%)	794 (32%)	231 (38%)
60Gy/20f	1029 (34%)	854 (35%)	175 (29%)
Gleason score			
≤ 6	1022 (33%)	882 (36%)	140 (23%)
$3+4$	1354 (44%)	1098 (45%)	256 (42%)
$4+3$	598 (19%)	421 (17%)	177 (29%)
≥ 8	97 (3%)	63 (3%)	34 (6%)
Clinical T-stage			
T1	1088 (35%)	931 (38%)	157 (26%)
T2	1713 (56%)	1344 (55%)	369 (61%)
T3	270 (9%)	189 (8%)	81 (13%)
Hormone Therapy			
LHRHa	2668 (87%)	2141 (87%)	527 (87%)
150mg bicalutamide	403 (13%)	323 (13%)	80 (13%)
Age (years)	69.1 (64.5–73.2)	69.3 (64.6–73.3)	68.7 (64.2–73.0)
Baseline/presenting PSA (ng/mL)	10.3 (7.3–14.6)	10.0 (7.1–14.0)	11.6 (8.5–16.0)

4.4.2 Modelling of PSA trajectories

In **Figure 4-1(a)**, PSA levels and boxplot distributions are presented, aggregated by years since starting treatment and outcome, with patients still at risk in the table below. There is much more variability and an increase in PSA values for those patients who recur at any time, compared to those that are alive and free from recurrence at their last follow-up. Presenting PSA values ($t = 0$) are higher for patients who recur. Apparent separation between the distributions of PSA is evident from year three and onwards. **Figure 4-1(b)** shows the smoothed reverse-year PSA trajectories (i.e., the PSA course in the years before a recurrence

or end of follow-up) of those that are recurrence-free and those patients who develop recurrence, stratified by fractionation arm, in the preceding two years before their last PSA for each outcome. Patients who develop a recurrence have higher presenting PSA levels, and do not achieve the same PSA reduction after treatment as patients who do not recur. In the final two-years before patients who developed recurrence, PSA increases at an exponential rate, compared with recurrence-free / censored patients whose PSA remains at a very low plateau. Those patients that receive the conventional fractionated dose appear to have slightly lower PSA between two-to-one years before recurrence, compared to the hypofractionated arms.

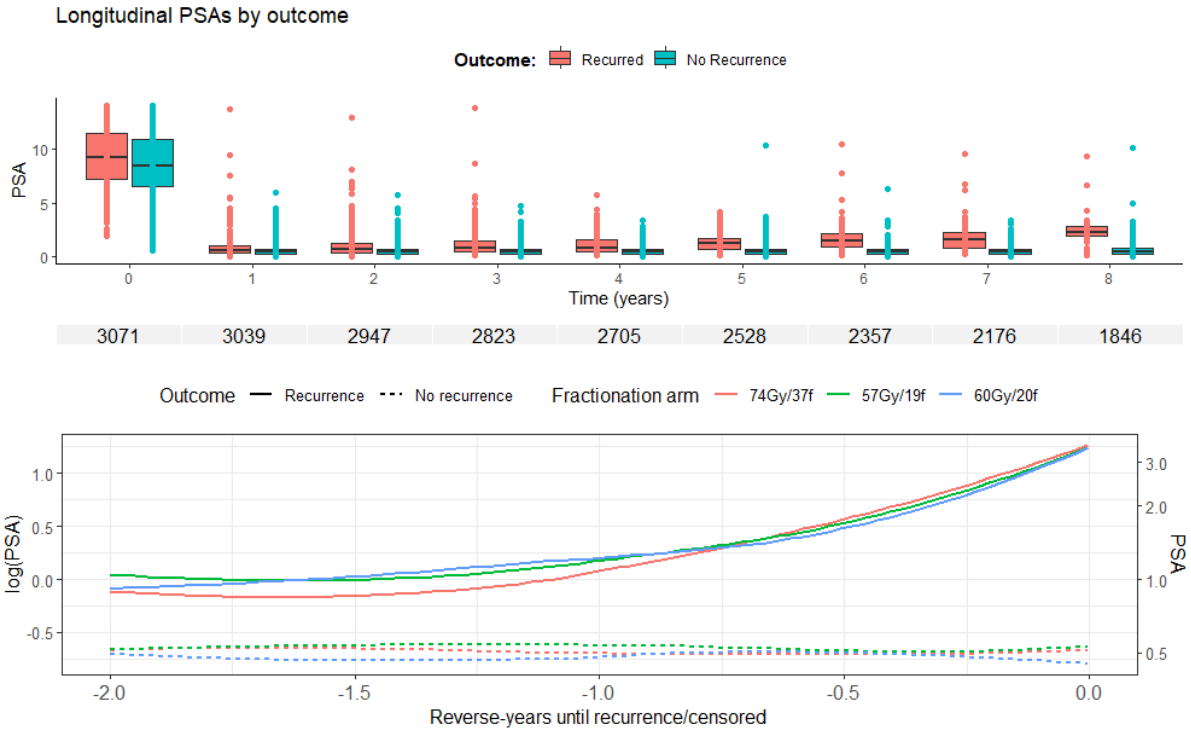


Figure 4-1 top (a): aggregated PSAs and boxplots by year and outcome since starting treatment, some non-recurrent PSA flares are observed. Patient numbers still at risk are presented below the plot. Bottom (b): smoothed reverse-year PSA trajectory plot, stratified by fractionation arm, lowess smoothers shown with the solid lines indicating recurrence and dashed lines indicating no recurrence. In the nonrecurring patients, a few PSAs >5 ng/mL are recorded; these PSAs were considered bounces/flares and therefore did not achieve the protocol's definition of clinically confirmed biochemical failure.

Knot selection for the mixed-effect submodel was performed to find the optimal number of knots that balanced between fitting the nonlinearity of PSA flexibly, whilst assessing the variance-bias trade-off. A range of internal knots were assessed, from one to four ($K = \{2, 3, 4, 5\}$). Likelihood-ratio testing, along with the AIC & BIC information criteria were obtained, and 5-fold cross validation for each number of internal knots was performed to

check whether there was a reduction in the MSE and that the model was not overfitting with an increasing number of knots in the testing data [159]. These results are shown in **Table 4-2**. The model with 4 internal knots were chosen; more knots were not considered due to computation demand and previous work has shown more knots is rarely required in practice [168]. Breakpoints were placed at fixed vigintiles, 20th% 0.75yrs; 40th% 1.9yrs; 60th% 3.51yrs; 80th% 5.36yrs.

Table 4-2 knot selection procedure fitted with maximum likelihood mixed-effect models for each internal knot for the natural cubic splines in the fixed and random effects, whilst adjusting for baseline covariates (treatment received, T-stage, Gleason, age). The LRT & p-value is comparing to the row above it. df=degrees of freedom, AIC=Akaike's information criterion, BIC=Bayesian information criterion, MSE=mean squared error, CV=cross validation, LRT=likelihood-ratio test.

Internal knot	df	AIC	BIC	MSE (5-fold CV)	log-likelihood	LRT	p-value
0 (linear)	14	125411	125533	1.04	-62691	*NA*	*NA*
1	16	120901	121041	0.97	-60435	4513	<0.0001
2	18	109429	109587	0.84	-54697	11476	<0.0001
3	20	94256	94430	0.70	-47108	15178	<0.0001
4	22	82796	82988	0.63	-41376	11464	<0.0001

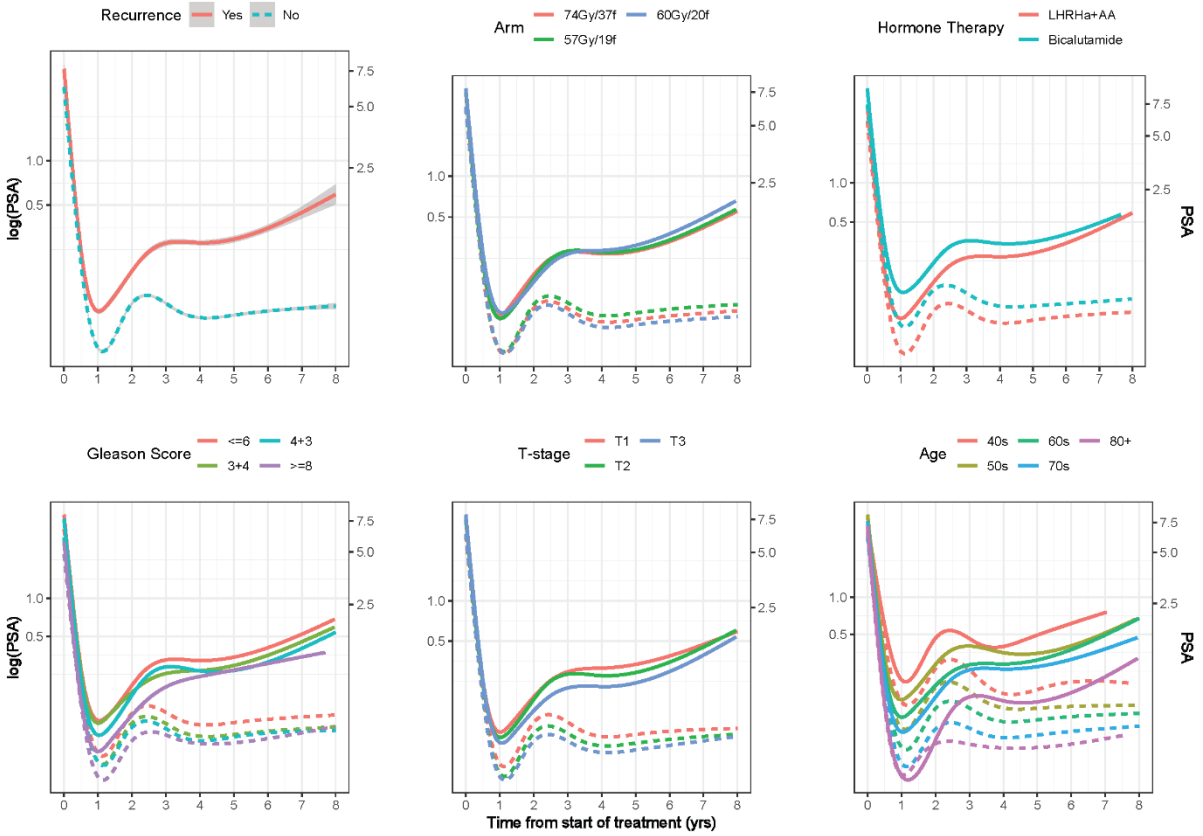


Figure 4-2 the predicted effect plots of PSA, stratified by outcome (solid – recurrence, dashed – censored) and each baseline subgroup over time. The top-left panel depicts the overall average PSA trajectories for each outcome. The natural cubic spline smoother is depicted.

Figure 4-2 shows the mixed-effect joint model predictions and how each baseline factor impacts on the PSA trajectory, by outcome. Initial high levels of PSA at diagnosis, which drop for both groups during treatment, are seen. When treatment stops, PSA recovery/bounce is seen at 1-2 years after start of treatment, the slight bump around 2 years is likely due to the effects of testosterone recovery. For those that go on to remain event-free, a slight decrease of PSA is seen and then a stable plateau.

For fractionation schedule, there is generally little difference between the PSA trajectories for each schedule in the first year, and then PSA slightly deviates post-two years with systematically lower predicted PSA values in the 60Gy/20f arm for those with no recurrence, but highest predicted PSA values for those that have cancer recurrence. Visually there does not appear to be much predicted difference in the GS and T-stage for the lower risk factors, but GS≥8 and T3 subgroups appear to exhibit lower PSA trajectories. Patients who received

LHRHa appear to have lower predicted PSA values than those receiving bicalutamide; noting that allocation to hormone therapy was not randomised, with most patients (87%) receiving LHRHa. The biggest effect on PSA trajectories is age at diagnosis, with younger patients (ages 40-49, $n = 6$) exhibiting higher post-treatment PSAs; for those who do not relapse, a stable PSA after 4 years is seen across all age groups. The fixed and random parameter estimates for the final mixed-effects joint submodel are presented in **Tables 4-3** & **4-4**, respectively.

Table 4-3 fixed effect model parameters from the joint mixed-effect submodel. HT = hormone therapy.

$\boldsymbol{\beta_j}$	Covariate	Mean	SD	2.5%	97.5%	p-value	$\widehat{R}$
0	Intercept	1.93	0.05	1.83	2.03	<0.001	1.000
1	Spline: [0, 0.75] yrs	-1.92	0.05	-2.02	-1.81	<0.001	1.000
2	Spline: [0.75, 1.90] yrs	-2.64	0.05	-2.74	-2.54	<0.001	1.000
3	Spline: [1.90, 3.51] yrs	-0.79	0.05	-0.88	-0.69	<0.001	1.001
4	Spline: [3.51, 5.36] yrs	-5.45	0.11	-5.67	-5.24	<0.001	1.000
5	Spline: [5.36, 10] yrs	-0.44	0.05	-0.54	-0.34	<0.001	1.000
6	Arm: 57Gy/19f	0.01	0.03	-0.04	0.06	0.690	1.000
7	Arm: 60Gy/20f	-0.02	0.03	-0.07	0.03	0.450	1.001
8	Gleason: 3+4	-0.10	0.02	-0.15	-0.05	<0.001	1.001
9	Gleason: 4+3	-0.08	0.03	-0.14	-0.02	0.007	1.000
10	Gleason: ≥ 8	-0.30	0.06	-0.43	-0.18	<0.001	1.000
11	T-stage: T2	0.04	0.02	-0.01	0.08	0.103	1.001
12	T-stage: T3	0.02	0.04	-0.06	0.10	0.674	1.000
13	HT: 150mg bicalutamide	0.20	0.03	0.14	0.27	<0.001	1.000
14	(Age-69) yrs	-0.01	0.00	-0.01	0.00	<0.001	1.000
	$\boldsymbol{\sigma}$	0.46	0.00	0.45	0.46		1.004

Table 4-4 the estimated symmetric variance-covariance matrix, D, of the random effect from the fitted joint model. The diagonal elements in bold indicate the standard deviations of the random effects.

Random Effects	$\boldsymbol{b_0}$	$\boldsymbol{b_1}$	$\boldsymbol{b_2}$	$\boldsymbol{b_3}$	$\boldsymbol{b_4}$	$\boldsymbol{b_5}$
$\boldsymbol{b_0}$	**0.46**					
$\boldsymbol{b_1}$	-0.52	**0.91**				
$\boldsymbol{b_2}$	-0.31	0.76	**0.74**			
$\boldsymbol{b_3}$	-0.22	0.46	0.68	**0.94**		
$\boldsymbol{b_4}$	-0.25	0.52	0.77	0.53	**1.37**	
$\boldsymbol{b_5}$	-0.13	0.23	0.51	0.90	0.47	**0.98**

4.4.3 Joint modelling time-to-recurrence

The best-fitting association structure $f(M_i(t), \boldsymbol{\alpha}, \boldsymbol{b}_i)$ using the above longitudinal parameterisation (with four internal knots) was investigated. The PSA value and then a linear combination of PSA value its gradient was fitted to assess the goodness-of-fit. Incorporating the slope of PSA improved the information criteria (DIC: Value = 82893 *vs* Value + Slope = 82877; wAIC: Value=83610 *vs* Value + Slope = 83367; LPML: Value = -41835 *vs* Value + Slope = -41698) despite the additional complexity of incorporating the derivative of PSA in fitting the model; the association structure $f(M_i(t),\ \boldsymbol{\alpha}, \boldsymbol{b}_i) = \alpha_1 m_i(t) + \ \alpha_2 \frac{dm_i(t)}{dt}$ is used hereafter.

*Table 4-5 comparing CHHiP hazard ratios from the standard Cox submodel and joint model. Age was median-centred (minusing 69 from all ages), * indicates 95% Bayesian credible intervals and $\hat{R}$ from the joint model.*

Covariate	N	HR_{Cox} [95% CIs]	HR_{JM} [95% CIs*]	$\widehat{R}_{JM}$
Fractionation arm 74Gy/37f	1017	Reference level		
57Gy/19f	1025	1.14 [0.95,1.38]	0.99 [0.65, 1.50]	1.001
60Gy/20f	1029	0.87 [0.71, 1.07]	1.01 [0.71, 1.44]	1.006
Gleason score ≤ 6	1022	Reference level		
$3+4$	1354	1.36 [1.11, 1.68]	1.81 [1.33, 2.49]	1.001
$4+3$	598	2.26 [1.81, 2.83]	2.76 [1.95, 3.95]	1.003
≥ 8	97	2.62 [1.80, 3.82]	2.49 [1.27, 4.96]	1.002
Tumour Stage T1	1088	Reference level		
T2	1713	1.54 [1.27, 1.86]	1.47 [1.10, 1.97]	1.002
T3	270	2.15 [1.64, 2.81]	2.41 [1.52, 3.76]	1.004
Hormone therapy LHRHa	2688	Reference level		
150mg bicalutamide	403	0.97 [0.76, 1.23]	0.70 [0.48, 1.00]	1.003
Age *(median-centred)*	3071	0.99 [0.98, 1.01]	1.05 [1.03, 1.08]	1.001

Table 4-5 presents the hazard ratios of the baseline covariates and compares the HRs with the standard baseline Cox model to the joint model. Conditioning on the PSA trajectory, fractionation schedule did not show a statistically significant difference in the treatment effect from the conventional arm (all p-vals>0.95), in line with the primary analysis [13]. Worsening of GS & T-stage, and older age at diagnosis, were associated with increased risk of recurrence

(all p-vals<0.01). Patients who received bicalutamide appeared to have lower risk of recurrence, although this was not statistically significant (p-val=0.053), in line with previous results [17]. All chains aligned in agreement from $\hat{R}$ being very close to unity.

Table 4-6 measuring the strength of association parameters between the two outcomes, log-hazard ratio ($\boldsymbol{\alpha}$) per unit increase in log(PSA) and its slope, with 95% credible intervals (CIs).

Association structure	log-HR	95% CIs	$\hat{R}_{JM}$
$\boldsymbol{\alpha_1 \log \text{PSA}(t)}$	$\alpha_1 = 4.52$	[4.07, 4.99]	1.007
$\boldsymbol{\alpha_2 \frac{d \log \text{PSA}(t)}{dt}}$	$\alpha_2 = 2.08$	[1.74, 2.43]	1.008

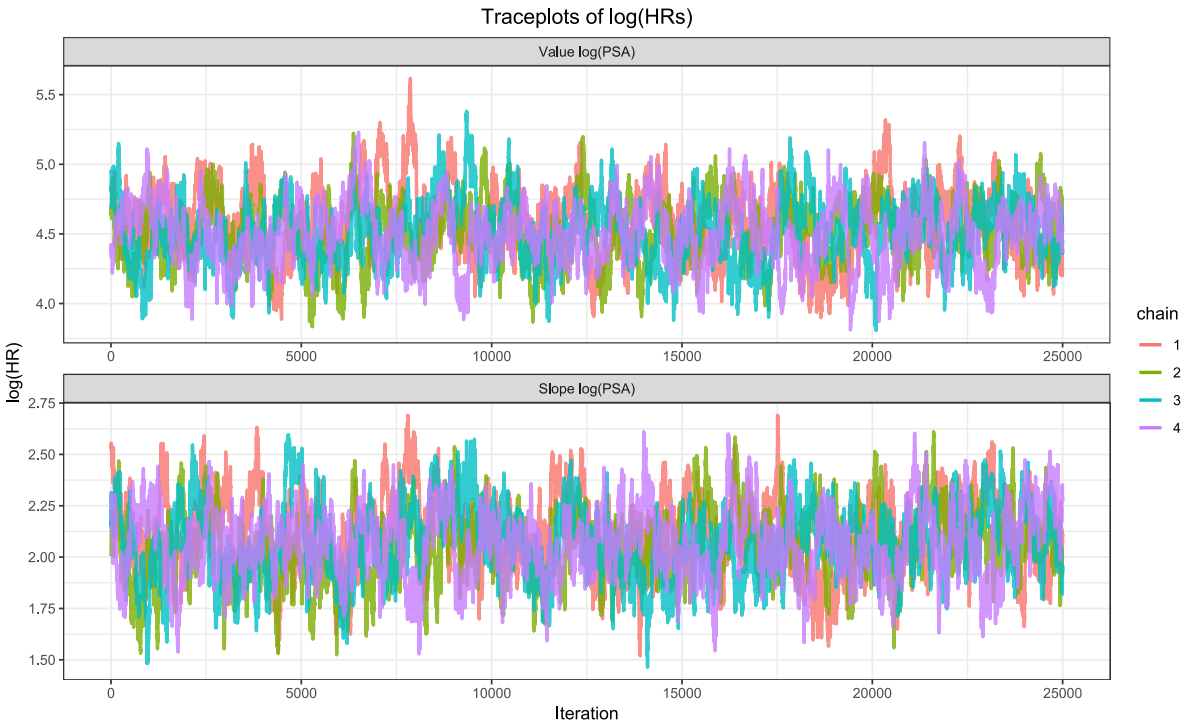


Figure 4-3 traceplots of the four $\boldsymbol{\alpha}$ log-hazard ratio parameter chains, for the value and slope association structure for PSA.

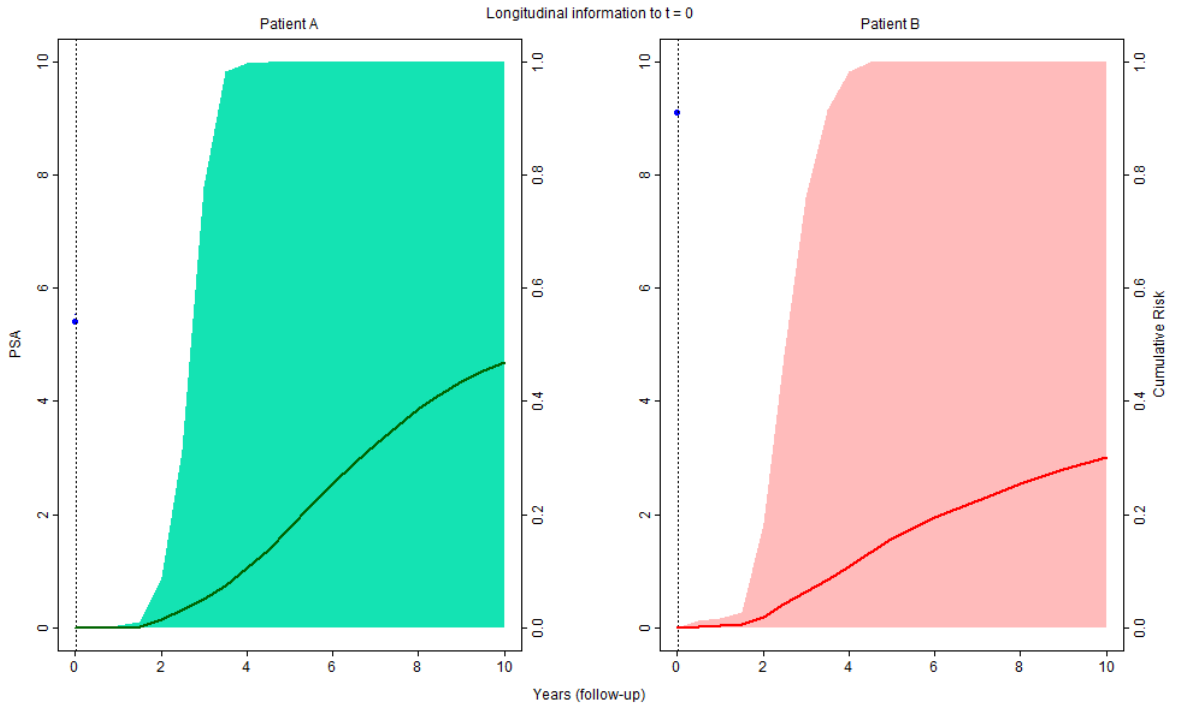
The log-hazard ratios $\boldsymbol{\alpha}$ and $\hat{R}$ diagnostics with their corresponding MCMC traceplots of convergence are presented in **Table 4-6** & **Figure 4-3**, respectively. For the association with the mixed-effect model, the log-hazard ratio parameter estimates for both the PSA value and PSA gradient are 4.52 (95% CI=4.07—4.99), and 2.08 (95% CI=1.74—2.43), respectively, indicating that both absolute PSA value and its gradient at a given time as parameterised in the mixed-effects model are highly predictive of recurrence. The traceplots are akin to a 'hairy caterpillar' indicating alignment and no apparent divergence.

4.4.4 Dynamic predictions

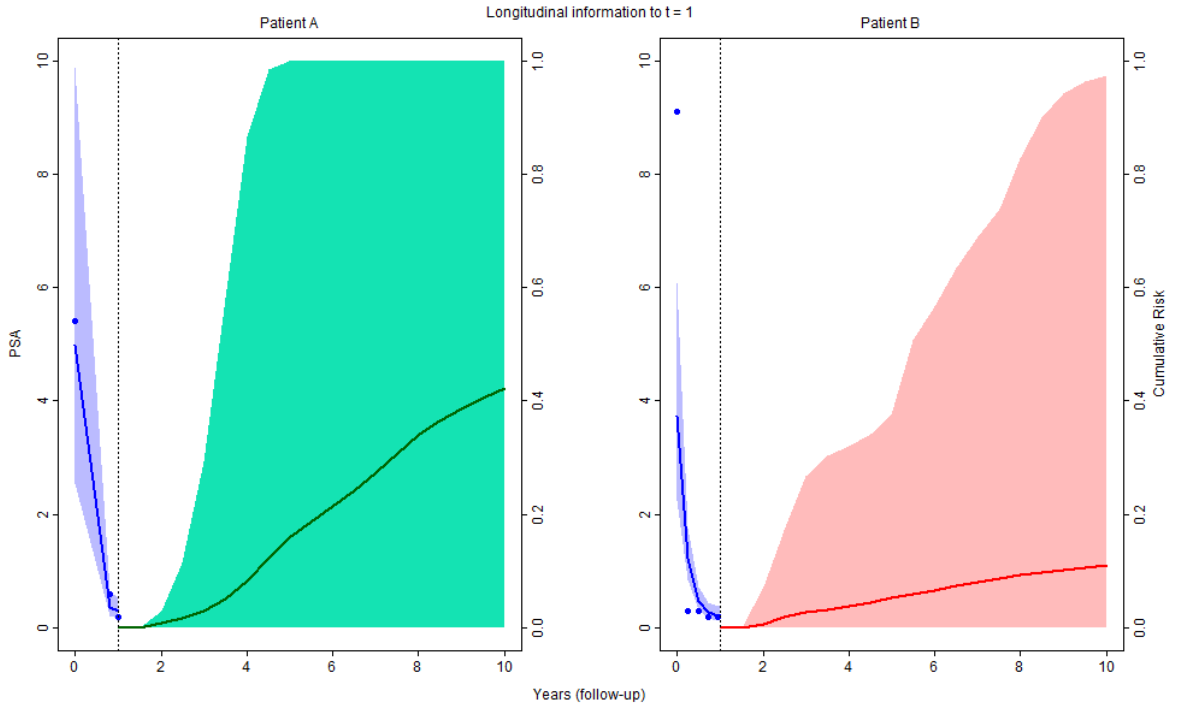
I demonstrate how the model updates prognosis over time on two selected patients who received the same treatment (57Gy/19f radiotherapy schedule, LHRHa hormone therapy) and PSA follow-up schedule, similar age at diagnosis and contrasting NCCN risk groups at presentation (patient *A*: GS = 8, T3, presenting PSA = 5.3ng/mL, *vs* patient *B*: GS = 6, T1, presenting PSA = 9ng/mL), and outcome. Dynamic predictions for these two patients are presented in **Figure 4-4** over five panels (**V—Z**) for different prediction landmark times ($t = 0, 1, 3.5, 4.5, 5$ years), to predict risk ten years after initiating treatment. On each panel, the left side of each figure depicts PSA (in blue, observed PSA values in dots, whilst the line depicts estimated predicted PSA) and the right side shows the point estimate of the cumulative risk of recurrence up to a ten-year horizon from the landmark time (in green the curve for *A* who does not experience recurrence, the red for *B* who does). The shaded areas show the 95% credible intervals of the estimated predictions for each outcome.

At baseline (at $t = 0$ years, **Figure 4-4 V**), patient *A* has poorer baseline prognostic factors and worse ten-year prognosis (~45% recurrence risk) than patient *B* (~30% recurrence risk), despite having a lower presenting PSA. For both patients, using only presenting PSA gives very wide credible risk intervals for the predictions beyond two years. A year since starting treatment (**Figure 4-4 W**), both patients exhibit a similar drop in PSA with patient prognosis slightly improving for *B*. In **Figure 4-4 X** (landmark $t = 3.5$), *A*'s PSA remains low whilst *B*'s PSA level starts to increase beyond the plateau. In **Figure 4-4 Y** (landmark $t = 4.5$), A's PSA continues to remain low and stable, with their risk substantially dropping, whilst B's PSA continues to increase thereby further increasing his risk of recurrence. In **Figure 4-4 Z** after 5 years follow-up, *A*'s PSA continues to be very stable around 0.1 ng/mL, thus his updated prognosis is very good, with reduced credible intervals for his predictions, compared to *B*'s, whose post-treatment PSA presents more variability and increases over time. The risk of recurring by 10 years for *A* is very small (~5%) compared to *B*'s risk of recurrence (>60%), with jumps in estimated risk at each previous landmark after a year. This is driven by the accrued PSA levels before 5 years, approaching biochemical failure. Patient *A* was recurrence-free by 9 years follow-up, whilst *B* had a recurrence by 5½ years.

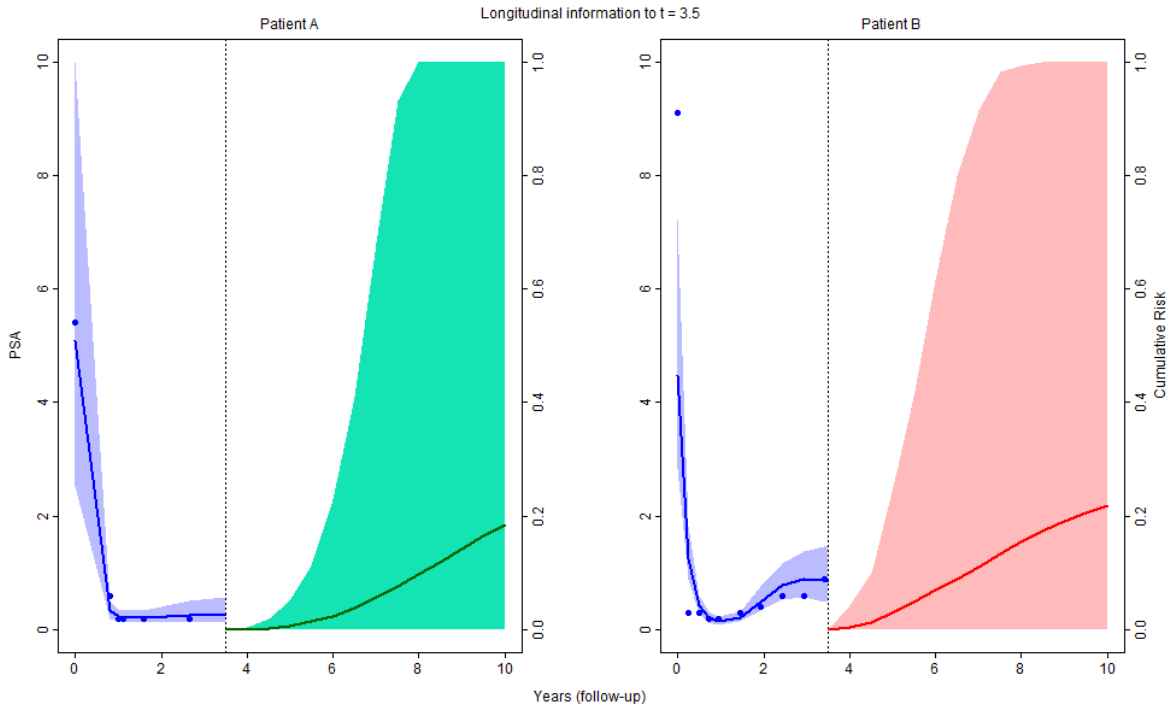
4V


Longitudinal information to t = 0
Patient A
Patient B
PSA
Cumulative Risk
Years (follow-up)

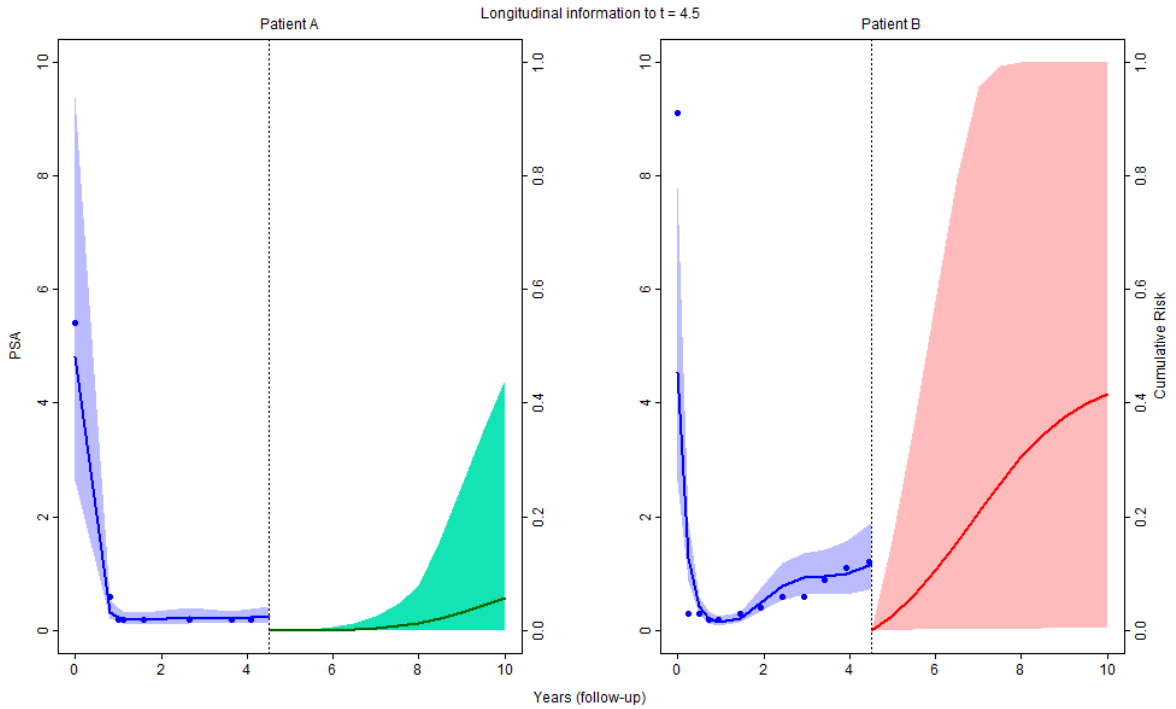
4W


Longitudinal information to t = 1
Patient A
Patient B
PSA
Cumulative Risk
Years (follow-up)

4X


Longitudinal information to t = 3.5
Patient A
Patient B
PSA
Cumulative Risk
Years (follow-up)

4Y


Longitudinal information to t = 4.5
Patient A
Patient B
PSA
Cumulative Risk
Years (follow-up)

4Z

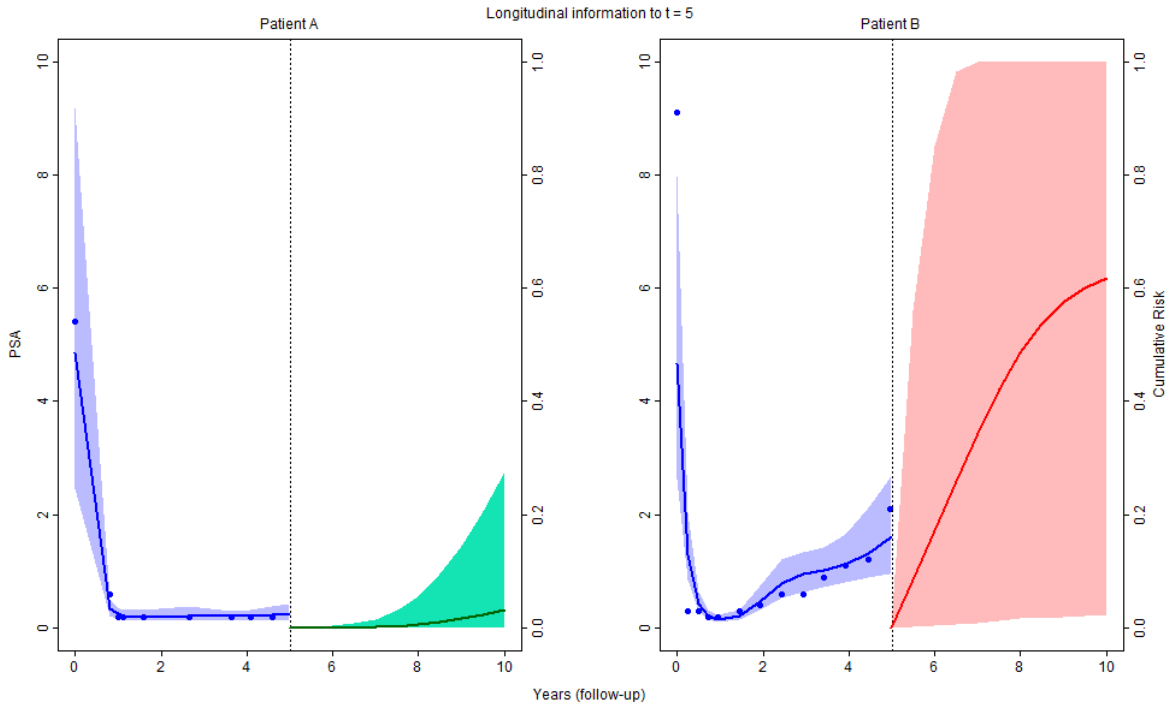


Figure 4-4 dynamic predictions of two patients: A & B, over five panels (V—Z). Patients A & B are ages 63 and 64 respectively and both received the same treatments, with contrasting prognostic factors. The left-hand side of each plot shows their modelled PSA values over time and the right-hand side shows their cumulative risk of recurrence at particular landmarks by ten years after initiating treatment. The 95% credible intervals are shown (shaded).

4.4.5 Assessing predictive performance

The CDPJM's calibration and discrimination for predictions of risk to recurrence by 8 years is assessed. Presented in **Table 4-7**, the 50-times repeated bootstrapped optimism-corrected metrics (mean for each time point) for the CDPJM to predict recurrence at landmark times from years zero ($t = 0$, baseline) to seven ($t = 7$), with a fixed horizon prediction of eight years. Discrimination improves as more longitudinal PSA information becomes available after three years' worth and AUC was maximised after five years of follow-up (AUC=0.84, 95% bootstrapped confidence interval (bCI) = 0.81—0.87). Similarly, calibration and the Brier score improves considerably after four years. The overall corrected AUC is 0.70, (95% bCI = 0.51—0.86); ICI=0.05, (95% bCI = 0.014—0.089); Brier=0.10, (95% bCI = 0.025—0.164). The apparent (non-resampled) validation metrics where $\pi_i(u = 8|t)$ (a fixed horizon of eight years), and fixed prediction windows (varying horizons) of $\pi_i(t + 2|t)$ & $\pi_i(t + 5|t)$ can be found in ***Appendix A*** (chapter 4)***, Supplementary Table A2***.

Table 4-7 optimism-corrected model metrics from landmark times t=0—7 predicting at a horizon time of 8 years. Discrimination – AUC (area under the curve); calibration – ICI (integrated calibration index); overall predictive performance – (Brier score). Mean, [95% bootstrapped CIs] refers to the bootstrapped replications of the posterior means. The ICI & Brier are loss functions (where lower is better), with higher AUC measures indicating better discrimination. Ns are patients remaining at risk at the development landmark.

		Optimism-corrected metrics		
Landmark t_{years} for prediction interval [t, 8]	**N still at risk**	**AUC**	**ICI**	**Brier**
$t = 0$ ***(baseline)***	3071	0.525 [0.500—0.553]	0.056 [0.043—0.068]	0.16 [0.154—0.166]
$t = 1$	3039	0.58 [0.556—0.6]	0.06 [0.045—0.072]	0.153 [0.147—0.16]
$t = 2$	2947	0.612 [0.583—0.644]	0.083 [0.069—0.098]	0.153 [0.145—0.16]
$t = 3$	2823	0.651 [0.632—0.677]	0.061 [0.049—0.069]	0.123 [0.113—0.132]
$t = 4$	2705	0.748 [0.728—0.767]	0.045 [0.036—0.052]	0.097 [0.089—0.106]
$t = 5$	2528	0.797 [0.767—0.821]	0.038 [0.031—0.048]	0.068 [0.062—0.075]
$t = 6$	2357	0.838 [0.807—0.868]	0.024 [0.019—0.029]	0.047 [0.039—0.054]
$t = 7$	2176	0.806 [0.756—0.873]	0.016 [0.013—0.019]	0.027 [0.022—0.033]

4.4.6 PSA risk thresholds

In **Figure 4-5**, I perform linear regression analysis between the predicted risk of recurrence by eight years from the joint model, given the accrued longitudinal biomarker information up to landmark time t ($t = 0, 1, \ldots, 5$ years), and the latest PSA value available prior to the landmark time. At all landmarks there is a strong positive correlation between latest PSA value and predicted risk of recurrence by eight years. As the latest PSA value nearest to landmark time increases, predicted prognosis worsens with an increased recurrence probability. The recurrence risk threshold is minimised at the origin for each landmark time t. This gives an approximate level of an 'acceptable' average PSA threshold on a continuous scale.

For instance, at the start of treatment (landmark $t = 0$ years) in **Figure 4-5** (top-left), for a minimal PSA, the lowest predicted risk is 13% (y-intercept) and a relatively small R^2 value as there is some heterogeneity here at baseline time origin with wider 95% prediction interval (PI) bands (8-year recurrence risk PI 0% to 29% at the intercept). As follow-up continues ($1 \leq$

$t \leq 2$ years), PSA drops to the nadir (the lowest recorded PSA) which is near-zero. The intercept implies a minimal PSA predicts a recurrence risk of 11% and 7% at landmark time one and two years respectively. At landmark times 3, 4, and 5 years, the regression intercepts are negative (a nil PSA implying an infeasible negative risk), though their magnitudes are very small; PSA levels less than 0.23, 0.34, and 0.41ng/mL respectively are indicative of having a small (< 5%) risk of recurrence by 8 years.

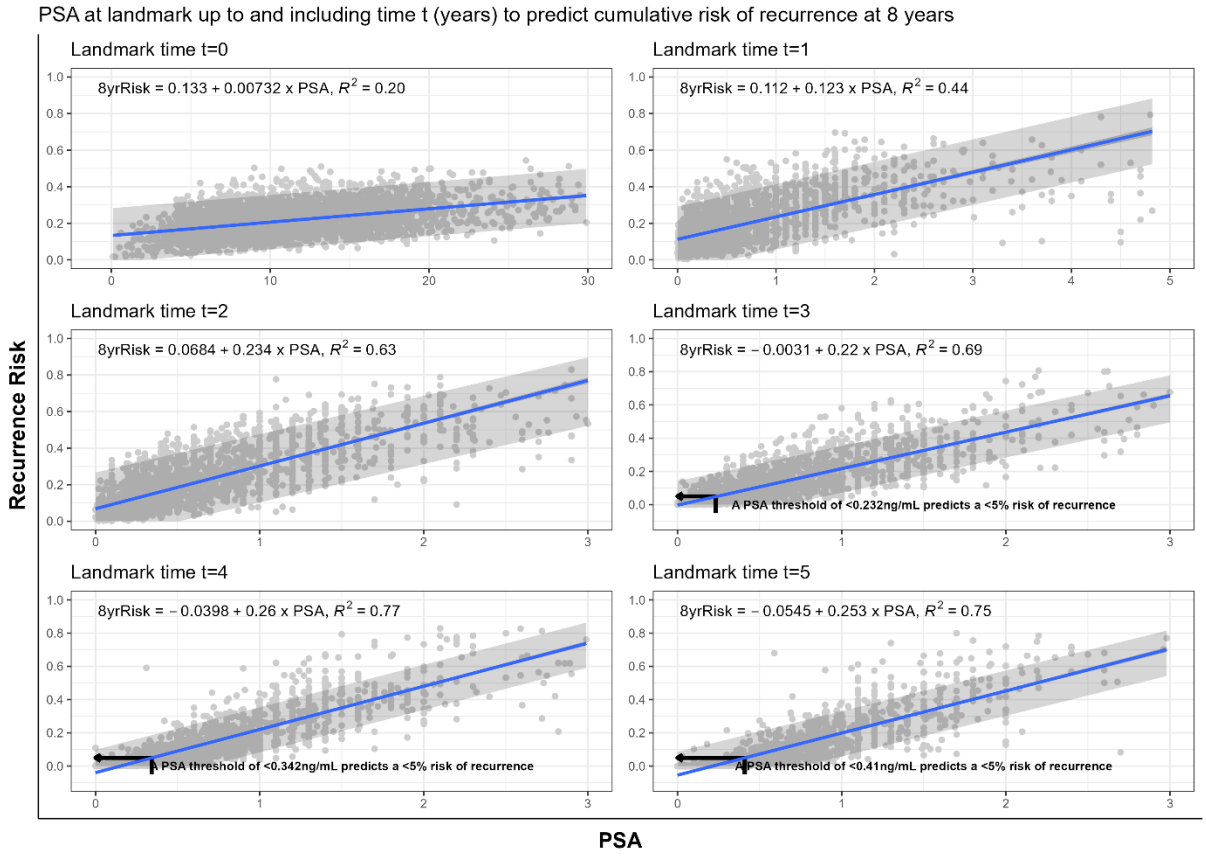


Figure 4-5 scatter plots of PSA predicting prognosis/recurrence risk at 8 years (horizon), each panel represents landmarks 0 – 5 years. Each grey dot indicates a patient's PSA (nearest to that landmark time) and risk at each landmark time. PSAs ≤ 3ng/mL are considered after t=1. The blue line indicates regression fit with the corresponding equation and R^2 labelled in each panel, with 95% confidence intervals. The wider grey bands indicate 95% prediction intervals. At the intercept (or less) indicates the predicted recurrence risk for a nil PSA; for the regression lines at t=3,4,5, each PSA threshold is labelled that predicts a <5% risk.

4.5 Discussion & conclusions

In this chapter, I have developed a dynamically updated clinical prediction joint model for the risk of prostate cancer recurrence in patients treated with both hormone therapy and IMRT in the CHHiP trial. It was shown that incorporating longitudinal PSA values collected over time into the model, in addition to baseline prognostic factors and treatment schedules, aids and improves prediction of individual patient prognosis. I explored and quantified the effect of hypofractionation (3Gy/f) compared to conventional fractions (2Gy/f) on patients' longitudinal PSA trajectories and on recurrence. There was no statistical evidence of a difference between either of the hypofractionation schedules, compared to the conventional fractionation arm, in terms of the PSA trajectories or reducing recurrence risk as expected due to the non-inferiority hypothesis of CHHiP's study design.

PSA levels typically started to rise exponentially approximately 1½-2 years before formal confirmed biochemical failure. The association of PSA values and its rate-of-change were both highly significant and predictive of recurrence. The rationale to include PSA gradient is that there may be non-recurring patients who have a higher post-radiotherapy PSA value but continues to be stable (non-increasing over time), compared with a patient who may have a lower PSA post-radiotherapy that continues to increase post-treatment. With the entire PSA trajectory captured and supplied to the CDPJM, PSA to nadir is directly modelled and has previously been shown to be an important predictor of event-free survival [169]. Similarly, inference in changes of the minimum (nadir) PSA between patients can be made. The nadir often occurred by two years from the commencement of treatment, with PSA value and gradient at nadir both being close to zero (e.g. take two similar patients where their only clinical difference is a PSA nadir of nil and 0.1ng/mL after a year of starting treatment). The predicted recurrence risk by eight years of the patient with a higher nadir is 4.85%, over doubled from the risk of the patient with a lower nadir, 2.06%. However, in absolute values, this is still a small increase, and the patient with a higher nadir would still be considered to have a good prognosis. A nadir of 0.2ng/mL at a year after commencing treatment would increase their eight-year recurrence risk to 14%.

I also attempted to quantify the relationship between PSA values at particular landmarks from starting treatment and subsequent recurrence. This is not straightforward as it is difficult to define precise or best cut-offs for PSA, which need clinical (and patient) value judgements. For example, some implausible predicted risk values from the regression parameter estimates (the intercept) between landmark times 3—5 years after treatment was seen. One way to remedy this is to force the intercept to be zero, alternatively more careful consideration and sophisticated methods could be applied here (e.g. beta-regression or zero-inflated models).

However, linear regression was used to give an indication of upper PSA bounds to predict a recurrence risk of $< 5\%$ by 8 years at various landmarks in a simple 'rule-of-thumb' without overcomplicating the interpretation. It is worth noting that the extracted risk predictions are obtained from the joint model. It is not only the PSA value by the landmark time, which is considered, it is also its rate-of-change and its history modelled by the mixed-effects submodel. Using only the most recent PSA value predictor at the landmark time is a simplified approach, as the raw concentration is a proxy to each patient's PSA trajectory by the landmark time point. Additionally for personalised predictions it is difficult to give one-size-fits-all cut-offs, and a balance must be made to the weighting and importance of false-positives and false-negative predictions. The data suggest that PSA levels ≤0.23, 0.34, and 0.41ng/mL at 3, 4, and 5 years respectively give a reasonable indication of having a <5% risk of recurrence by eight years. For those same landmark times and risk threshold, 27%, 40%, and 51% proportion of patients have this PSA threshold (or less).

Similarly, the logging of the PSAs was performed to help the model conform to distributional assumptions. Depending on the PSA assay used, there may have been some censoring in the PSA response where PSAs were read as zero, but not truly zero and in fact below the lower limit of quantification. In this thesis, a simple mitigation for these zeros was to artificially add a small offset term (0.1ng/mL) to all PSAs so that the logarithm could be calculated straightforwardly. Other possible methods could have been used such as only offsetting the 33 zeros in the data, using the reciprocal of the maximum PSA value, or more sophisticated methods such as imputing those censored values from the truncated part of the appropriate lognormal distribution [170].

It is encouraging that these thresholds are consistent with previous findings in mono-EBRT [171,172]. Yock et al similarly state that a 5-year PSA ≤0.5ng/mL has very good prognosis (97% progression-free rates by 8 years) [172]. These studies have some differences with this work, they are over 20 years old, lower radiation doses were delivered, with no hormone therapy, and they used PSA categories at a fixed time at 5 years using a simplified Kaplan-Meier landmarking approach [171,172]. This may explain the slightly lower threshold that was found; the continuous method for ascertaining these thresholds is more flexible, without arbitrarily categorised PSAs.

Conversely a PSA of 1ng/mL at 5 years gives a predicted risk of recurrence of 20% (95% PI = 6%–33%). The reason that prediction intervals were reported, rather than the smaller confidence interval, is to have a prediction range for new patients entering this treatment pathway. There is reasonable heterogeneity in **Figure 4-5** scatter plots at the earlier landmark times (indicated by the lower R^2 values), with most patients within the 95% PI bands. It is worth noting the individualised predictions directly from the joint model will give bespoke credible intervals. This chapter supports the importance of presenting, nadir, and post-treatment recovery levels of PSA. Findings from this CDPJM suggest that patients with a PSA $\lessapprox$ 0.23ng/mL and stable (low or nil gradient) PSA by five years are very unlikely to exhibit any future recurrence / clinical failure, by $t = 8$ years. This is consistent with findings in the context of prognosis after brachytherapy [173].

I chose to validate up to a fixed horizon time of eight years, given that the median time at the date of data snapshot (Oct 2019) was 8.6 years, despite being able to extend this as seen in the dynamic predictions. A fixed horizon approach was chosen to exemplify how predictions improve as more PSA information is collected. The model also allows predictions at fixed prediction windows, such as two and five years from fixed landmark times (e.g. given a recurrence-free patient's data up to three years, what is the predicted risk of recurrence in the next two or five years). These apparent (non-resampled) metrics are presented in ***Appendix A*** (chapter 4) ***Supplementary Table A2***. There is relatively little difference in the validation between the two methods (apparent *vs* bootstrapped-corrected) AUC and Brier scores, with

some notable differences in the earlier landmarks. There are slightly bigger differences in the first two years for the ICI metrics.

Diagnostics of the joint model were performed (***Appendix A*** (chapter 4), ***Supplementary Figure A1***). The longitudinal component conforms to the assumptions, though there were some departures observed in the tails of the quantile-quantile plot, suggesting t-distributed residuals could be appropriate. The random effects themselves conform to normally distributed residuals. The Cox submodel proportional hazards assumption for baseline covariates were checked and found not to be violated; a joint model including time-varying $\alpha_1(t)$ showed some departures of proportional hazards for the time-varying PSA process, which become reasonably constant after the nadir. Previous work has shown the joint model is highly robust to departures in the proportional hazards assumption [111].

A limitation of the study is the inherent association of the longitudinal process and the outcome, as biochemical recurrence, confirmed by a clinician, was included in the definition of the outcome. This is because the primary endpoint in CHHiP captured failure-free survival, which is time free of any event that would trigger further treatment for the patient (or prostate cancer death). For this reason, for patients whose biochemical failure triggered treatment, it was not always possible to confirm clinical radiological progression.

Furthermore, I acknowledge the relative complexity of the joint model, namely in the mixed-effect submodel. The parameterisation of the longitudinal mixed-effects model is complicated with four internal cubic spline knots over time to capture the exhibited nonlinear PSA, with 15 main effect parameters needing to be estimated, with a total of 71 fixed parameter estimates. The complexity of the model is seemingly warranted, however. Further investigation of pairwise interactions of the baseline variables with time was performed (not presented). Most notably the differences between age and hormone therapy received groups were seen, however with an additional 45 parameters to be estimated, and as most curves were reasonably parallel, this was considered adding unnecessary complexity and did not improve the goodness-of-fit.

Differing parametrisations could have been considered, such as a parametric form however the PSA trajectories exhibited here are vastly different from those undergoing radiotherapy monotherapy-only treatments in **Chapter 3** – Literature Review that would not be translatable such as the biphasic parameterisation. Another parametrisation could have been the change-point model, where change-point knots could have been constructed at locations of interest, e.g. at the end of treatment phase, nadir, post-nadir recovery. In the context of CHHiP these treatment phases are known, however they may not be known (or missing) in other studies (RT01) or observational studies.

In addition, I performed a post-hoc calculation of the minimum sample size required for the time-to-event outcome as indicated by Riley & colleagues, to ensure there is an adequate sample size for the development of the prediction model [174,175]. I calculated the minimum sample size based on the Cox submodel only. Given the required adjusted Cox-Snell R^2, number of parameters, rate (number of events per person-year of follow-up), horizon time point of 8 years, and the mean follow-up time, it was found to be well within the allotted number of patients and events available (required: N=2828 & events=559 *vs* available: N=3071 & events=607) (full details and calculations can be found in ***Appendix A*** (chapter 4)). The Cox submodel (used in the joint model) used itself in calculating the $R^2_{CS_{\text{apparent}}}$ and resulting $R^2_{CS_{\text{adjusted}}}$. Future work could quantify the improved efficiency of joint modelling, given it uses both the time-to-event & longitudinal process [33], compared to baseline-only sample size calculations of the Cox proportional hazards model.

When conditioning on baseline covariates and PSA trajectory, it appeared that receiving bicalutamide magnified a reduced risk of recurrence, compared to LHRHa (not randomised) when using the regular Cox model. However these patients were younger and selected to have bicalutamide to reduce impact of side effects, as well as having a lower proportion of positive core biopsies and hence better prognosis [17]. Conversely, PSAs for these patients remained higher than for LHRHa patients (**Figure 4-2**). Surprisingly, it was seen in patients with worse prognostic factors (Gleason ≥8 & T3) had the lowest PSA trajectories, however there were relatively few patients in these subgroups ($n = 97$ and $n = 270$ respectively).

I assumed the PSA value & gradient association structure, linking PSA trajectories and its impact on the time-to-event process, but further association parameterisations could be considered, such as incorporating the cumulative area under the PSA biomarker over time, random-effects structure, and time-varying $\boldsymbol{\alpha}$ extensions [176]; further combinations and interactions between prognostic factors could additionally be considered therein. Clearly this can lead to questioning other association structures. Two approaches can be considered to overcome this, including penalising the parameter estimates using Bayesian shrinkage priors to automate variable selection [177]. This adds a further degree of complexity and I chose an intuitive approach, similar to previous studies [107,108,111]. Alternatively, the use of joint latent class models could be explored: this other joint modelling framework does not rely on explicit assumptions on the association structure linking the two submodels [41,84].

Dynamic prediction models that incorporate longitudinal PSA levels to predict risk of recurrence in prostate cancer have been previously explored as discussed in **Chapter 3**. A full review of these relevant studies can be found by this author with similar dynamic prediction windows and expected predictions to this chapter [41]. However as stated, these previous studies use PSA dynamics after standard EBRT has ceased and without neoadjuvant or concurrent hormone therapy; therefore, in their setting the PSA dynamics are different, with lower PSA values at $t = 0$, slower PSA decrease to nadir, and elongation of its trajectory, compared to the PSA dynamics observed from pre-hormone treatment PSA values captured here. As recruitment of these previous studies recruited patients from the 1980s, there have been significant advances in treatment, with 5- and 10-year survival rates doubling (in the UK) since then [178]. Additionally, the majority of CHHiP patients received neoadjuvant and concurrent hormone therapy, with hypofractionated radiotherapy regime, therefore this model and analysis is applicable to the current global standard of care.

I compared the prognostic performance of the CDPJM to other published articles. Arguably most similar to this work is Taylor et al who propose a joint model using real-time evaluation of predicting recurrence of prostate cancer [108]. The longitudinal PSA biomarker was modelled using a biphasic exponentially decreasing-increasing parametric function. Some of their parameter estimates were remarkably similar to this work, namely in the log-hazard

ratios to the PSA level, T-stage and Gleason. Their prediction time focuses on a window of no more than three years ahead, whereas I presented a fixed horizon prediction time of eight years (see ***Appendix*** A (chapter 4)***A***, ***Supplementary Table A2***). Their exhibited PSA trajectory is an elongated tick shape, typical of radiotherapy-only treatment. Direct model comparison cannot be made due to their differing validation appraisal methods, and lack of androgen deprivation therapy. There is likely not much difference in predictability between mono-radiotherapy and dual-therapy at earlier landmarks. However, the nadir may occur later for monotherapy patients, which could slightly decrease the predictability compared to dual-therapy at the nadir. In Proust-Lima et al [66], integrated Brier scores are used that are not directly comparable to my measures, in Proust-Lima et al [84] prediction errors are comparable to the ICIs presented in this chapter.

As follow-up continues, and there are an ageing population of patients in CHHiP, deaths from causes unrelated to prostate cancer may represent a competing risk for the outcome of interest but were found not to be a huge issue in this snapshot, therefore these deaths were treated as censored in this model. Extensions exist for joint models accounting for competing risks, but extracting dynamic predictions and assessing their predictive performance is not trivial [88,179]. This extension is explored in **Chapter 6**. Other extensions to this model could include an additional multivariate longitudinal process (e.g. with both PSA and testosterone), which is known to be prognostic in later disease stages [180], or novel biomarkers of early detection of recurrence, such as circulating-tumour DNA fraction [181]; or additional histopathological prognostic factors, such as Ki67 [182]. All these, however, were not routinely collected in this trial.

In the dynamic prediction example, it was seen that patient *B* in **Figure 4-4** had poor prognosis evident from their increasing PSA from 3½ years, despite their relatively good baseline prognostic factors. Although having worse baseline prognostic factors, PSA trajectory indicated patient *A's* prognosis was good, and continued to be so after 4 years of follow-up; the model could be further extended to recommend and reduce follow-up frequency and burden. For instance, amongst the patients who were recurrence-free and alive at 5 years, the median time to failure for patients who recurred after 5 years was 7 years. In this cohort, the

median predicted cumulative incidence of recurrence is 4% by year 6, 12% by year 7, 20% by year 8, 27% by year 9 and 34% by year 10. Amongst patients who do fail by 7 years $[t = 5, u = 7]$ $(n = 136)$, their median cumulative risk of failure is 30%; compared to a median of 2% risk of failure for equivalent patients who are censored by 7 years $(n = 423)$. This demonstrates the predictive difference of the two outcomes and the two-year lead-time capabilities of the model, suggested by the reverse-time plot (**Figure 4-1**(b)).

Development of a clinical calculator would allow the clinician to visualise each patient's personalised risk of recurrence over time; if the risk surpasses an unacceptable threshold, further investigation could be considered, and personalised follow-up schedules could be designed [32,134]. To achieve this, in the subsequent **chapter (5)** the CDPJM undergoes robust external validation so its clinical utility and generalisability can be assessed in differing patient populations, where alternative treatment modalities and similar PSA dynamics are expected. For instance, I suggest for future work to explore the applicability of the proposed CDPJM with stereotactic radiotherapy or using longer hormone therapy schedules [46,135]. It may be that, with differing treatments and disease stage, alternative model development and/or recalibration of the baseline hazard is required. Additional work might also include decision analysis to quantify net benefit at various thresholds, versus a 'do-all-or-nothing' approach for every patient [183].

To conclude, I quantified the impact of an increase in PSA value and rate-of-change on prostate cancer recurrence, adjusting for baseline prognostic factors and treatments in the CHHiP trial. The model will be applicable to future patients who undergo neoadjuvant and concurrent short-course hormone therapy with either conventional or hypofractionated IMRT. As expected, PSA trajectory is indicative of predicting recurrence, as previous studies have shown. I also assessed the performance of the prediction model, which showed good calibration and discrimination, optimal after 4-5 years' worth of accrued longitudinal PSA biomarker information to predict recurrence by 8 years. I demonstrated the practical aspect of these models in performing dynamic predictions from the relevant patient population, which can help to guide patient care and allocate limited resource more effectively. Additionally, I proposed clinical thresholds at various landmarks, with simple continuous calculations to

determine alternative PSA thresholds given the recurrence risk clinicians (and their patients) might be willing to accept, which is easily applicable in clinical practice.

Chapter 5 – External Validation of the Clinical Dynamic Predictive Joint Model for Recurrence in Localised and Locally Advanced Prostate Cancer

5.1 Introduction

The aim of this chapter is to externally validate the CDPJM developed with CHHiP data in **Chapter 4** to appraise the model and its suitability applied in other healthcare settings using data from two randomised clinical trials (RCT): RT01 [47] and RADAR [46].

In this chapter, I wish to assess the joint model's generalisability in other 'unseen' populations, as assessing model performance on the development sample (i.e., internal validation) often shows over-optimistic results as the data is used twice, to develop and to validate, regardless of when corrections like cross-validation or bootstrapping have been applied. Moreover, there is not usually enough heterogeneity to infer how the model will perform in the wider population. Crucially, the process of external validation can provide valuable insights into the model's strengths and weaknesses; for instance, to ascertain if the model performs well in certain subgroups, or after an allotted duration of follow-up; or conversely, to ascertain the biases elicited in the modelling process, or by differing study protocols. If issues are identified, then steps can be taken to improve the performance of the CDPJM and ensure it is robust and reliable.

Broadly, the RADAR & RT01 trials have similar inclusion criteria and treatment modality to CHHiP. Patients with either localised or advanced localised prostate cancer were given neoadjuvant and concurrent androgen suppression, together with radical radiotherapy. RT01 was CHHiP's 'predecessor' trial, comparing dose-escalated conformal radiotherapy to the contemporary standard-of-care control group; this experimental arm then became the control arm in CHHiP. RADAR is a 2x2 randomised factorial design trial, comparing the efficacy of neoadjuvant androgen suppression therapy duration (6 *vs* 18-months) with the possible

addition of zoledronic acid; all patients were treated with conventional fractionated radiotherapy.

I hypothesised that, given RT01's similar inclusion criteria to CHHiP, external validation measures of predictive performance should perform reasonably comparably to CHHiP. However, RT01 is a higher risk population and will provide a more robust validation of the model. RADAR included patients with locally advanced disease, having worse prognosis, potential for miscalibration was anticipated in the baseline hazard (effectively the intercept for the survival submodel), with the CHHiP CDPJM potentially underpredicting the observed risk of recurrence in the RADAR cohort.

5.2 Methods & Materials

5.2.1 External cohorts

External validation was conducted under TRIPOD guidance [40]. Anonymised data from two external RCTs, RT01 (N=834) and RADAR (N=1,051), was obtained, with data-sharing agreements put into place with each trials' research groups to enable collaborative research. Both trials were registered (ISRCTN47772397, ISRCTN90298520), approved by the relevant Research Ethics Committees (MREC/97/2/16, 03/06/11/3.02) and by the institutional research board of each participating international site. These studies were conducted in accordance with principles of good clinical practice; full details of the trials' designs have been described previously [46,47]. **Figure 5-1** compares each of the trials' treatment and follow-up schema.

RT01 (managed by the Medical Research Clinical Trials Unit, London, UK) was a phase III, open-label, international RCT. It recruited 843 men between January 1998 – December 2001 with T1b–T3a, N0, M0 prostate cancer, and presenting PSA below 50 ng/mL. Patients were randomly assigned 1:1 to either standard dose (64Gy/32f), or dose-escalated (74Gy/37f) conformal radiotherapy to assess superiority of the latter. All patients received neoadjuvant (for 3–6 months before the start of radiotherapy) and concurrent androgen deprivation therapy. The experimental dose-escalated fractionation arm was the control arm in CHHiP.

RADAR (Randomised Androgen Deprivation and Radiotherapy, TROG 03.04, managed by the Trans-Tasman Radiation Oncology Group, Newcastle, Australia) was a phase III, open-

label, 2x2 factorial design trial that between Oct 2003 and Aug 2007 recruited 1,071 localised and locally advanced prostate cancer patients T2a—T4, N0, M0. It compared two interventions: 1) 6 months *vs* 18 months of androgen deprivation therapy given with radiotherapy, and 2) given with or without zoledronic acid. As the addition of zoledronic acid was not beneficial (shown in the primary analysis via no interaction effects between androgen suppression and zoledronic), treatment groups were collapsed to focus on comparisons between the androgen suppression durations. Patients were given dose-escalated conformal EBRT (not randomised) of either 66Gy/33f, 70Gy/35f, or 74Gy/37f; or 46Gy/23f combined with a high-dose-rate brachytherapy boost over 19.5Gy/3f.

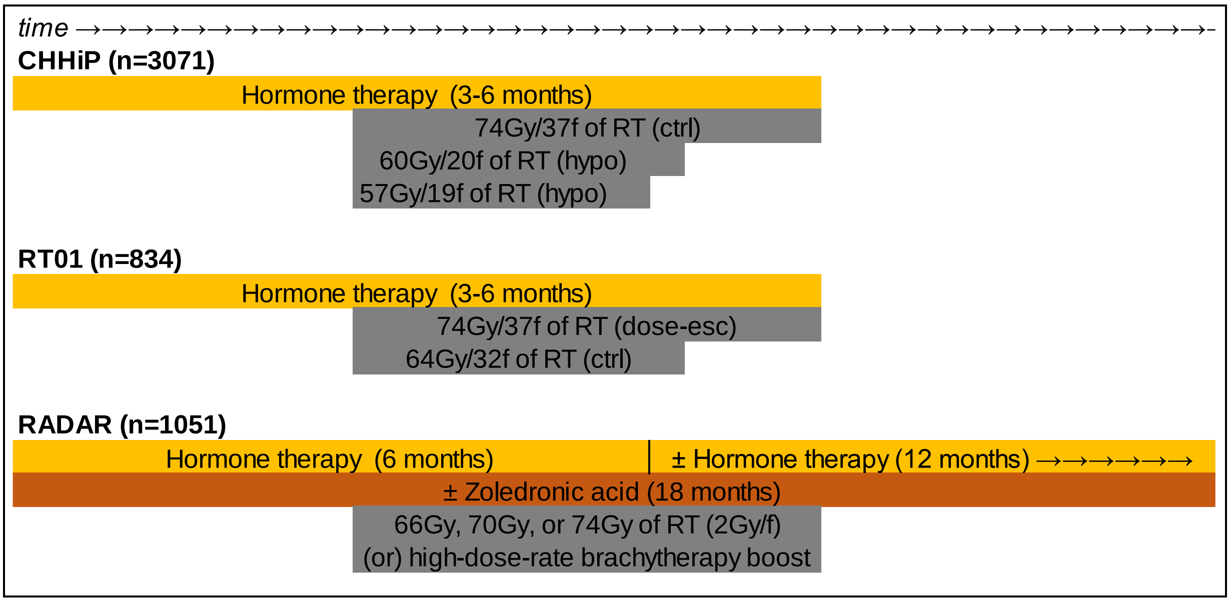


Figure 5-1 comparison of the trial treatments and follow-up schemas for CHHiP, RT01 & RADAR.

5.2.2 Outcomes

In RADAR, the primary endpoint was PSA progression (changed from prostate cancer-specific mortality after a protocol amendment in 2011). Biochemical failure was defined using the Phoenix definition of a PSA value > the nadir + 2ng/mL (Roach et al., 2006).

In RT01, the coprimary endpoints were biochemical progression-free survival (bPFS), local progression, and overall survival [47,184]. Biochemical failure was defined as an increase in PSA concentration to greater than the nadir by at least 50% and greater than 2ng/mL 6 months or more after the start of radiation therapy.

For the external validation of the CHHiP model, I have followed the same event definition, and defined the primary event as the composite of biochemical or clinical failure or death due to prostate cancer. For each trial, biochemical failure was defined using its trial-specific definition. Clinical failure included: recommencement of hormone therapy, local recurrence, lymph node or pelvic recurrence, and distant metastases. Time-to-recurrence was calculated as the time between the patient's latest pre-treatment PSA (presenting value, pre-hormone therapy, time origin $t = 0$), and the first primary event. Patients who were recurrence-free at their last follow-up visit were censored; this includes deaths unrelated to prostate cancer.

5.2.3 Statistical analysis

Predictions are obtained from the external cohorts using the CDPJM developed for CHHiP in **Chapter 4,** but reparametrised to exclude treatment (fractionation and hormonal therapy received); all other covariates remained as included previously. This approach was taken for the generalisability to future patients when treated with varying hormone therapy and radiotherapy schedules, so the model represents the average treatment effect. This reparameterisation made little difference to the model information criteria. Some bookkeeping is required on the external cohorts, to ensure the covariables of the model are in the correct format, analogous to CHHiP, to extract the dynamic predictions $\pi_i(u|t)$. For RADAR, Gleason grade 5 (Gleason score 9 or 10) was merged into Gleason score of ≥ 8, as the development of the original CHHiP CDPJM had only a maximum Gleason score of 8. In RT01, only total Gleason score was recorded, a score of 7 included either 3+4 or 4+3, therefore allocation to one or the other was imputed (see section **5.2.4**).

To evaluate predictive performance in the external cohorts, different prediction intervals are considered, with varying landmark times t from zero to seven years, to predict the probability of recurrence by eight years, $\pi_i(8|t)$. Additionally, I have also computed predictions using a fixed prediction window of two, $\pi_i(t+2|t)$, and five years, $\pi_i(t+5|t)$, ahead of the landmark time, and varying landmark times t up to eight and five years, respectively. Performance metrics were compared to the bias-corrected internal validation metrics proposed in **Chapter 4.4.5**. Graphical calibration plots are used to evaluate the calibration-in-the-large, using ten-quantile error bar groups. Possible miscalibration is assessed and resolved by recalibrating the

baseline hazard using the predictions from the separately fitted cohort-specific CDPJMs of RADAR & RT01. The ICI [91] is calculated to quantify the improvements in those recalibrations. All analyses were done using **R** (v4.1.0) with the *JMbayes2* package (v0.2-3—0.3-0) [136,185]. Similarly to **Chapter 4**, the TRIPOD checklist [40] can be found in ***Appendix B, Supplementary Table B1***.

5.2.4 Multiple imputation for missing Gleason levels

The Gleason scoring system, developed in the 1960s, is an ordinal scale ranging from 2 to 10, with higher scores indicating more aggressive and with worse prognosis tumours. The original Gleason grading system used a two-tier system, where the primary and secondary patterns of cancer growth were assigned a grade from 1 to 5, and the Gleason score (GS) was calculated by summing the two grades. The primary grade corresponds to the dominant pattern of the pathology of the tumour (>50% of the total specimen). There have been several changes to this scoring system over the years, with the latest modification confirmed in 2014 by the International Society of Urological Pathology to use a Gleason Grade Grouping (ISUP GGG) [186]. GS have been re-allocated into prognostic grading groups, scores of <6 are not recognised as individual Gleason scores of 1 & 2 no longer exist [187]. A GS = 6 corresponds to a GGG of 1; GS=3+4 → GGG=2; GS=4+3 → GGG=3; GS=8 → GGG=4; GS = 9 or 10 → GGG=5 [188,189].

As RT01 is an older protocol, designed in the 1990s, the Gleason pathology was recorded differently to today's standards. Moreover, only the sum score was recorded, but not its individual components, i.e., its primary and secondary grade. Specifically, the issue lies in a score of 7, either comprising of 3+4 or 4+3 that cannot be distinguished. It is known that a Gleason score 3+4 (GGG=2) has better prognosis than a Gleason score 4+3 (GGG=3), demonstrating histological heterogeneity for a GS of 7. One way to circumnavigate this issue and to maximise the efficiency of the external validations is to impute these two grades for a GS=7 in RT01. Multiple imputation by chained equations (MICE) are proposed [190].

The Gleason grade distribution from CHHiP is used to impute GGG to those patients with GS = 7 in RT01 to distinguish between 3+4 and 4+3. For simplicity, the imputation is conditional on GS=7 (i.e., only uses CHHiP patients with GS=7 and known GGG) and uses logistic

regression imputation for GGG=2 or GGG=3 using 50 sets of imputations. The imputation was based on the known T-stage, age, presenting pre-treatment PSA, follow-up time, and whether the endpoint was experienced. After the multiple imputations were performed, clinical expertise was sought to ensure that reasonable proportions of Gleason scores of 3+4 and 4+3 were imputed, as seen in current clinical practice for prognosis of these population. Without concerns from my clinical supervisors, these imputations were used in this analysis. A further sensitivity analysis is done by constraining all patients who had a GS=7 to be assigned all to 3+4, or all to 4+3, assuming a minimal change in predictive performance due to the predictive power of the joint model primarily being driven by the accrued longitudinal PSAs.

5.3 Results

5.3.1 Baseline characteristics comparison

The combined sample size for the external validation cohorts was N=1,885. For the original 1,071 recruited RADAR patients, 20 were excluded as 17 did not receive any radiotherapy, and 3 did not received it as per protocol (received either 50Gy or 76Gy). For RT01, 9 patients were removed: 5 had missing T-stage and Gleason score, and 4 had no PSA recordings available. Patients with a total GS of 7 and missing gradings were imputed as described previously in section **5.2.4**. Baseline characteristics per study are presented in **Table 5-1**, and compared to CHHiP. Both external validation studies exhibit worse prognostic characteristics, particularly in RADAR, that has no T-stage 1 patients, and more patients with Gleason score ≥ 8.

*Table 5-1 baseline characteristics of the development (CHHiP) and external validation cohorts (RADAR & RT01) total N=4956; * indicates imputation via MICE. MICE = multiple imputation by chained equations.*

Covariate	CHHiP, N = 3,071[1]	RADAR, N = 1,051[1]	RT01, N = 834[1]
Age	69 (64, 73)	69 (63, 73)	67 (63, 71)
Baseline PSA	10 (7, 15)	14 (9, 25)	13 (8, 20)
T-stage			
T1	1,088 (35%)	0 (0%)	206 (25%)
T2	1,713 (56%)	669 (64%)	479 (57%)
T3†	270 (9%)	382 (36%)	149 (18%)
Gleason Score			
≤ 6	1,022 (33%)	99 (9%)	537 (64%)
$3+4$	1,354 (44%)	342 (33%)	103 (12%)*
$4+3$	598 (19%)	247 (24%)	88 (11%)*
≥ 8	97 (3%)	363 (34%)	106 (13%)

[1] Median (IQR); n (%); * MICE imputation; [†] *RADAR combine T3 & T4*

5.3.2 Follow-up characteristics comparison

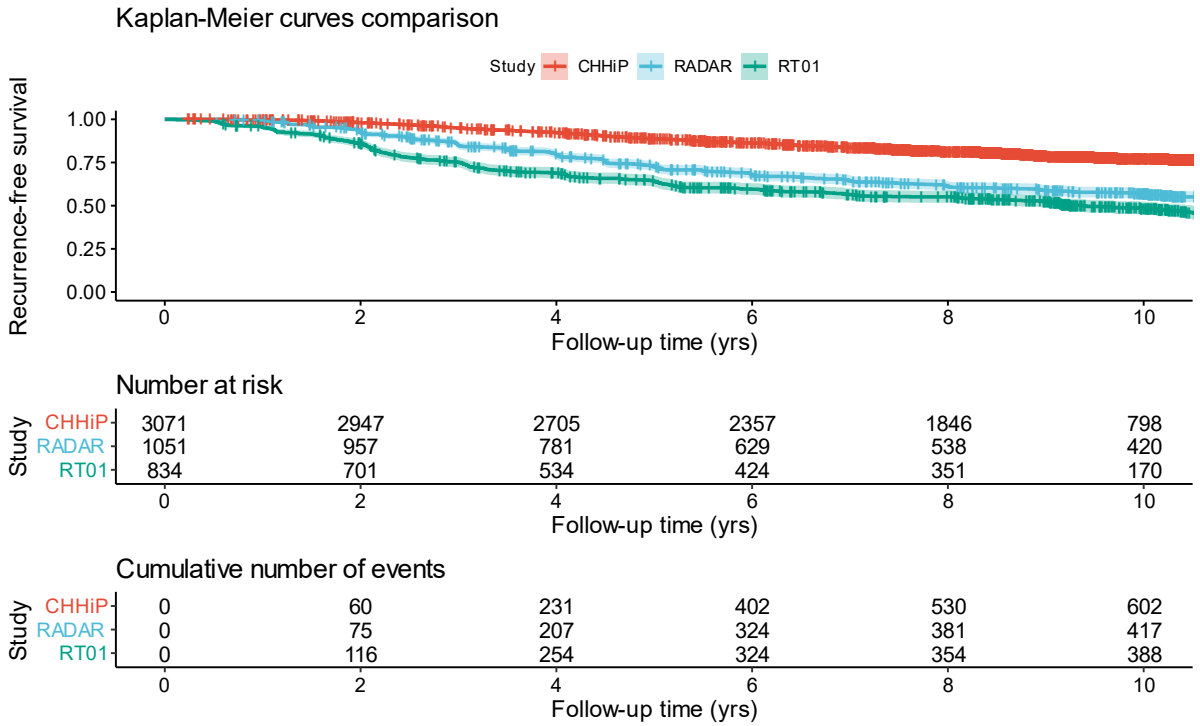


Figure 5-2 Kaplan-Meier curves comparing recurrence free survival in CHHiP (development cohorts), with RADAR & RT01 (validation cohorts).

The Kaplan-Meier estimates of recurrence-free survival for each trial are presented in **Figure 5-2**. Median follow-up time for RADAR is 8.2 years (IQR=4.0–11.0) and 6.1 years (IQR=2.6–9.5) for RT01. CHHiP has better prognosis compared to RADAR and RT01 cohorts. For RADAR there were a total of n=443 (42%) primary endpoint events, while RT01 had n=419 (50%) events. See ***Appendix B, Supplementary Figure B1***, which shows the recurrence-free survival for each RADAR hormone duration of 6 and 18-months.

There was a total of 26,697 PSA recordings for external validation (RADAR: 18,708; RT01: 7,989), with a median of 18 readings per patient (IQR=12–22). The average longitudinal PSAs trajectories for the three trials are depicted in **Figure 5-3**, stratified by outcome. In general, the external cohorts exhibit similar dynamics, with a steep drop in response to treatment, to PSA stabilisation and plateau after 3—4 years in those patients who do not recur. For recurrence (bottom panel), PSA gradually increases after the initial treatment decrease seen in the first 1—2 years. PSAs in RT01 are as expected similar to CHHiP; the lower PSA trajectory seen in RADAR are a consequence of the longer hormonal therapy schedules used in the factorial design of the trial. CHHiP & RT01 hormone schedule was 3—6 months, RADAR mandated at least 6 months.

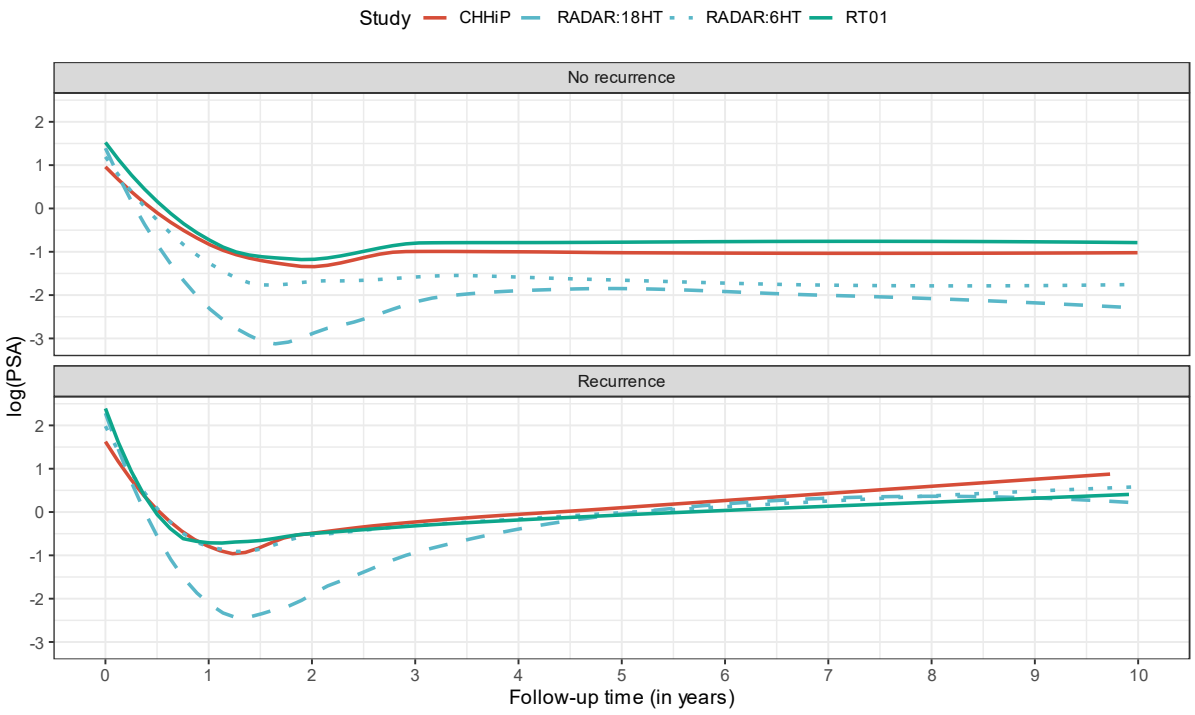


Figure 5-3 averaged logged-PSA trajectories for the three studies with RADAR stratified by hormone duration: (6 or 18 months), over follow-up, separated by outcome, top – no recurrence, bottom – recurrence. Lowess smoothers are depicted.

5.3.3 Predictive performance comparison

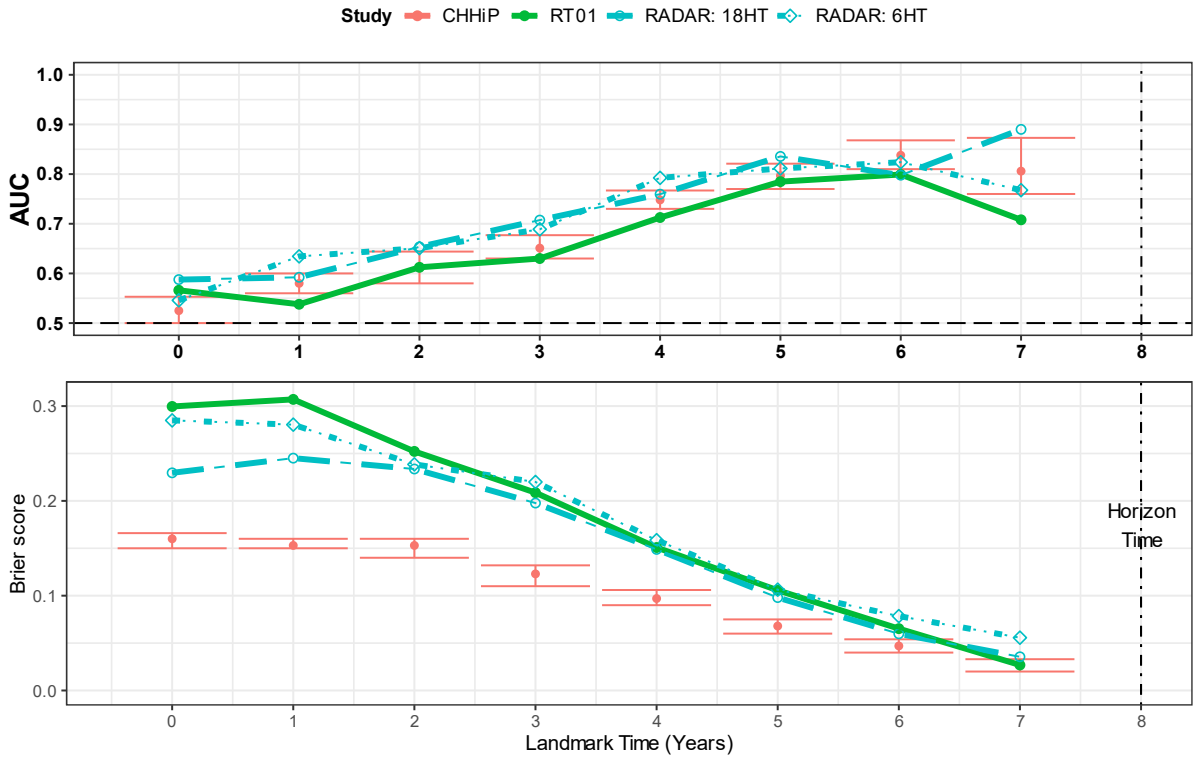


Figure 5-4 comparing the internal (CHHiP – red, 50 bootstrapped samples) and external validation cohort performance metrics (RADAR – blue stratified by hormone treatment duration: 6 & 18 months; RT01 – green). The AUC (top panel) assess discrimination, i.e., to distinguish between patients who do and do not have recurrence of their cancer, based on their accrued PSA. A higher AUC values indicate improved discrimination. The bottom panel assesses overall prognostic performance through the Brier score loss function (lower values are better). These are based on patient follow-up landmarks from zero to seven years, to predict recurrence by a horizon time of eight years, $\pi_i(8|t = 0, \ldots, 7)$. AUC = area under the receiver operating characteristic curve.

Figure 5-4 presents the AUC and Brier score metrics for the prediction of recurrence by 8 years, based on patient follow-up information from 0 to 7 years, $\pi_i(8|t = 0, \ldots, 7)$. It compares the 50x bootstrapped bias-corrected internal validation metrics of **Chapter 4.4.5** model development (presented as error bars) of the full JM to be used in clinic, and the two external cohorts RADAR (split between randomised hormone therapy duration: 6 and 18 months), and RT01. For AUC (top panel), one can see the external cohorts are largely comparable with the CHHiP biased-corrected internal validation. The model applied to RADAR's hormone durations both show similar performance (to one another) and has marginally better-than-expected discriminatory performance compared to the bias-corrected internal validation results for the first six years of PSA accrued landmark times. RADAR 6- and 18-month hormone therapy schedule gives an AUC of 0.77 and 0.89 respectively at a landmark of 7 years of accrued PSA

follow-up, when predicting recurrence by 8 years. Overall discrimination for RT01 patients is comparable to the bias-corrected metrics over the entire follow-up period with some over-optimism in CHHiP, with a drop in AUC at the seven-year landmark time.

For the overall prognostic performance measured via the Brier score prediction error loss function (**Figure 5-4**, bottom panel), in the earlier landmark years there is over-optimism of the bootstrapped bias-corrected samples compared to the external cohorts. For the external cohorts, RADAR has a marginally lower Brier score, with the 18-month schedule being consistently lowest and nearest to CHHiP, compared to 6-month hormone therapy and the overall RADAR cohort (averaged between the two hormone treatment durations), and RT01 in the first two years; these scores are then remarkably similar from landmark years three and onwards and align more closely to bias-corrected CHHiP at the latter landmark times (six years onwards). In general, as more longitudinal PSA information is accrued, the overall AUC and Brier scores are improved, indicating an improvement of predictive performance of the model in these cohorts, aligning more so to CHHiP. After seven years of accrued prognostic information, RT01 & RADAR 18-month hormone therapy has a minimal prediction error comparable to the bias-corrected internal validation.

To evaluate the predictive performance of RADAR's patients who are most similar to CHHiP, I compared the subgroup of T-stage 2 & Gleason score of 7 between RADAR's 6-month hormone therapy duration and this same subgroup for CHHiP for $\pi_i(8|t = 0, \ldots, 7)$. This is depicted in **Figure 5-5**. By and large, the predictive performance of RADAR was similar to the apparent performance in this subgroup of CHHiP patients, with some expected over-optimism in CHHiP, i.e., RADAR's predictive ability was slightly less. This subset of RADAR patients did outperform CHHiP in AUC for landmark years 1 and 2, (RADAR AUC = 0.64 *vs* CHHiP AUC = 0.61 at both landmarks). AUC was maximised for this RADAR subgroup at 5 years follow-up (AUC = 0.80) and CHHiP at 6 years (AUC = 0.89). Both the Brier and ICI scores (non-recalibrated) follow a similar pattern with the predictive error loss function metrics being larger for RADAR for both metrics for all landmark times. There is a moderate prediction error for both metrics up to landmark of 3 years. After 3 years of follow-up, these loss metrics start

to notably decrease, implying improved prognostic performance and calibration, with RADAR subgroup aligning more closely to CHHiP at the latter timepoints.

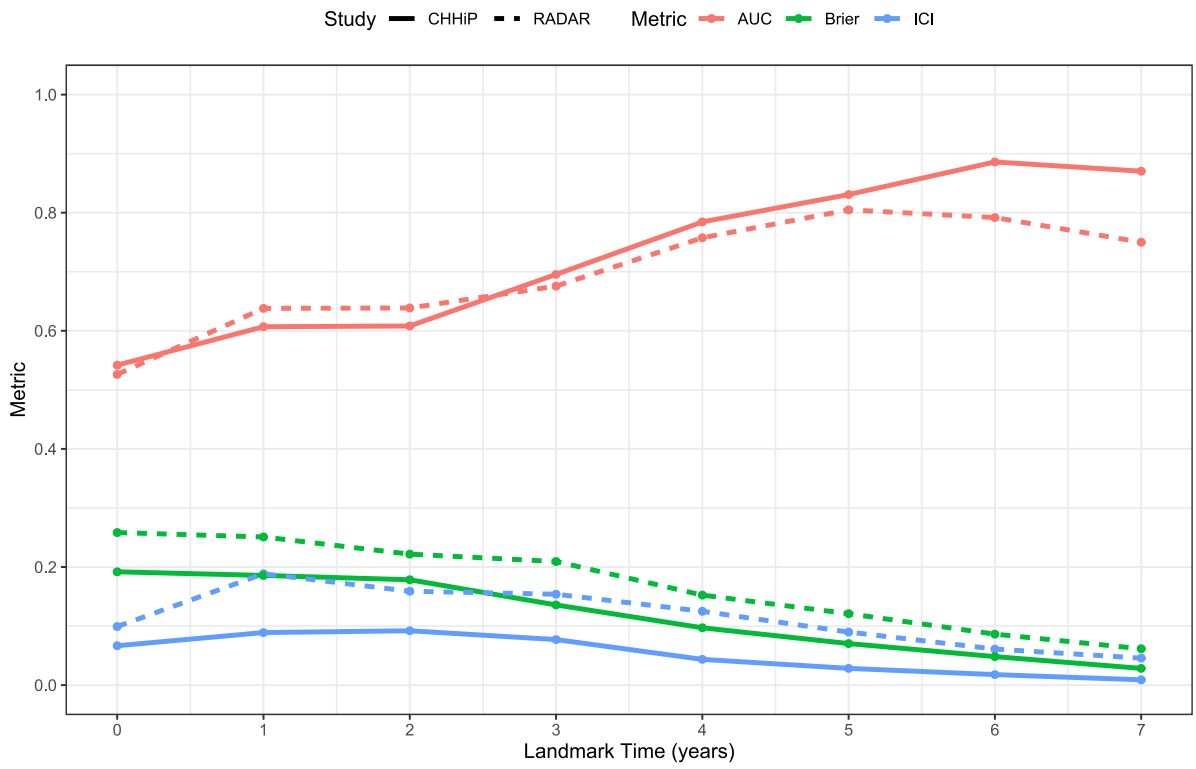


Figure 5-5 comparing the predictive performance for $(u = 8|t = 0, \ldots, 7)$ between RADAR (6-month hormonal therapy schedule) and CHHiP, for the subgroup of T-stage=2 and Gleason score = 7 patients in both cohorts. AUC= area under the receiver operating characteristic curve, ICI=integrated calibration index.

There was some miscalibration in the predicted *vs* observed probabilities for the external cohorts, which was expected due to the slightly differing patient population, and varying treatment modalities. To resolve this, the joint model's baseline hazard function is recalibrated for each study, using the baseline hazard of the joint model when fitted to each of the external datasets. The original ICI (before recalibration) is compared with the ICI of the recalibrated predictions and the percentage difference is calculated. The resulting improvement is shown in ***Appendix B, Supplementary Table B2***.

The recalibrated ICIs are visually depicted in **Figure 5-6** for each external cohort RADAR & RT01 and prediction interval procedure. There is generally higher ICI (less well calibrated) in the RADAR cohort. The lowest ICIs (better calibration) are generally lowest for $\pi_i(t + 2|t)$, i.e., only predicting 2 years ahead of the current landmark time. There is some fluctuation in

the first few landmark years, which then start to decrease from 3 years onwards. For RT01, calibration at $\pi_i(t+2|t)$ and $\pi_i(8|t)$ are remarkably similar at landmarks 4—6 years. Similarly for RADAR, all three prediction interval procedures are similar from landmark 4 years and onwards. For the prediction windows $\pi_i(t+5|t)$ and $\pi_i(8|t)$, the recalibrated ICIs are closely aligned in each cohort with more concordance between the two cohorts at the latter landmarking times for all prediction window procedures. For improved calibrations, this suggests accrued PSA data for at least 4 years is recommended.

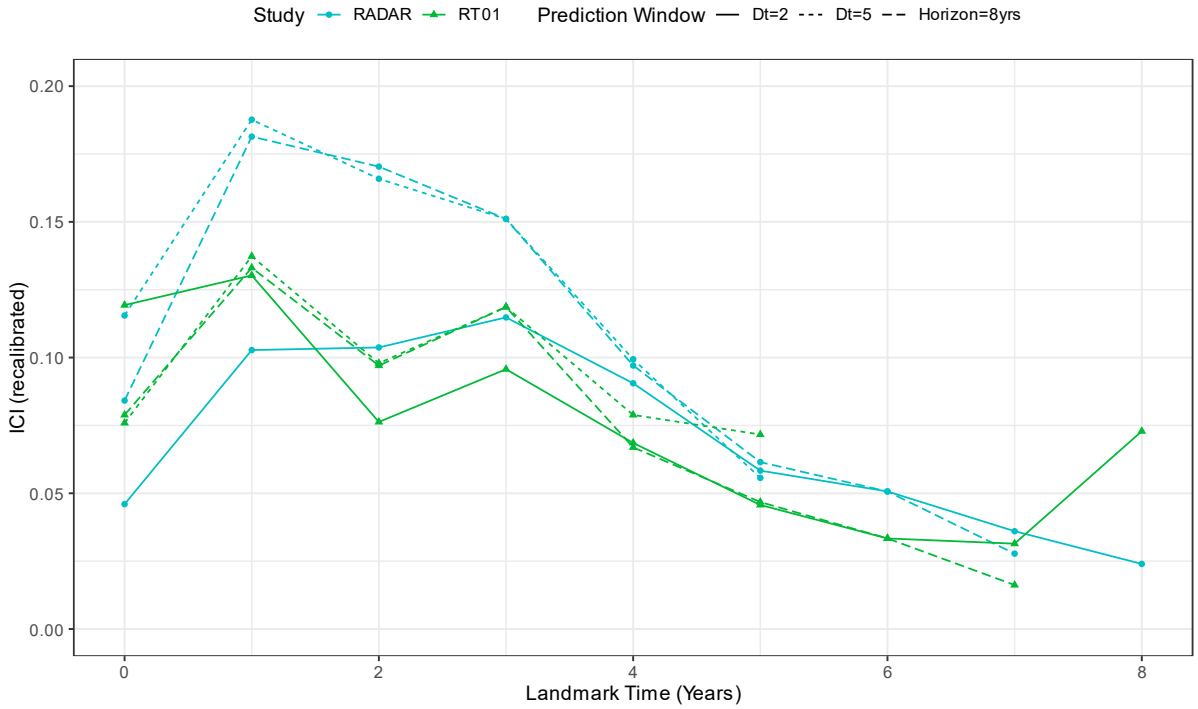


Figure 5-6 comparing prediction windows of the recalibrated index of the two external cohorts: RADAR & RT01.

To assess calibration-in-the-large, graphical calibration plots are presented (**Figure 5-7** & **Figure 5-8**) comparing the predicted and observed probabilities of recurrence $\pi_i(u|t)$, in RADAR and RT01, at various landmarks ($t = 0, \dots, 7$) and horizon times ($u = 8, u = t+2, u = t+5$). These figures display predictions before and after recalibration of the baseline hazard for each trial, to visually assess the improvement of the recalibrations.

Graphical calibration plots to predict $\pi_i(8|t)$ for RADAR (**Figure 5-7**) show systematic underprediction in the earlier landmark t times, and gradually re-align, more closely to what is observed, particularly from landmarks five years and onwards. The proportion of patients

with higher observed/predicted probabilities of recurrence are closely aligned. When recalibrating the baseline hazard (2nd and 4th rows), probabilities are better calibrated at the smaller predicted probabilities. Similar results are observed for RT01 (**Figure 5-8**).

Calibration plots (seen in ***Appendix B***) for $\pi_i(t+2|t)$ in RADAR (see ***Supplementary Figure B2***: original calibration & ***Supplementary Figure B3***: recalibration) show that there are considerable underpredictions in the first few years that tend to stabilise after 5 years of accrued PSA. Recalibration tends to improve, seen visually mostly at the latter landmark times of 5 years onwards and tends to resolve predicted recurrence probabilities. For RT01 $\pi_i(t+2|t)$ (in ***Supplementary Figure B4***: original calibration & ***Supplementary Figure B5***: recalibration) similarly there are underpredictions, less so compared to RADAR, in the earlier landmark times that again resolve after 4 years and more so when recalibration is applied.

Calibration plots for $\pi_i(t+5|t)$ in RADAR (in ***Supplementary Figure B6***: original calibration & ***Supplementary Figure B7***: recalibration), show again that there are some underpredictions in the first 4 landmark years, which appear to resolve at 5 years of accrued longitudinal PSA data. Recalibration of the model does not completely improve calibration-in-the-large, though. For RT01 $\pi_i(t+5|t)$ (in ***Supplementary Figure B8***: original calibration & ***Supplementary Figure B9***: recalibration) there are some considerable miscalibration underpredictions for all landmark times, however recalibration appears to better resolve the latter landmark times of 4 & 5 years, although this is not perfect: by and large the 95% confidence intervals of the error bars lay within the 45-degree line.

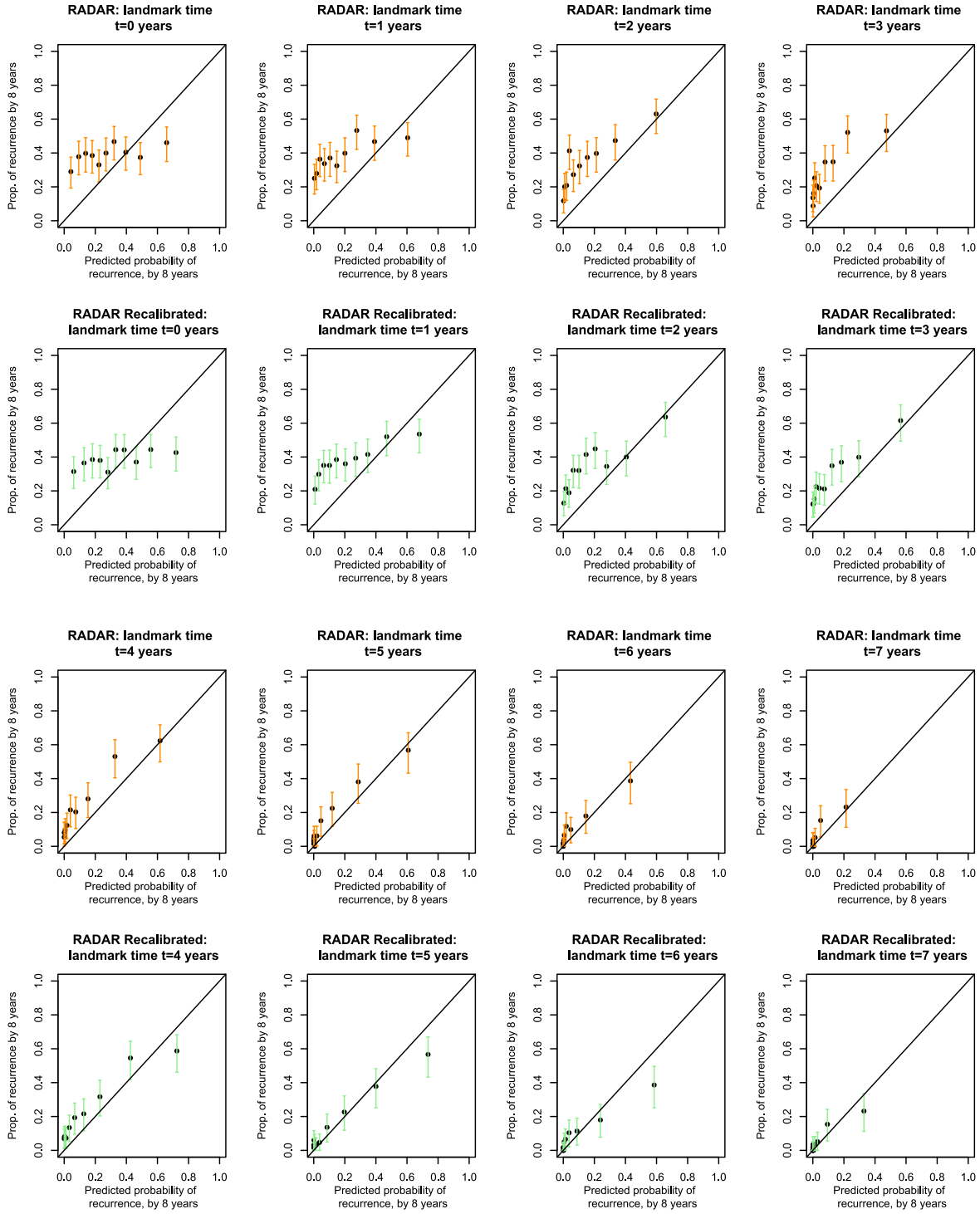


Figure 5-7 visually assessing calibration-in-the-large via graphical calibration plots of the RADAR cohort, before (1st & 3rd rows, orange) and after recalibration (2nd & 4th rows, green), for a fixed horizon of eight years at landmark zero to three years (top panel) and four to seven years (bottom panel).

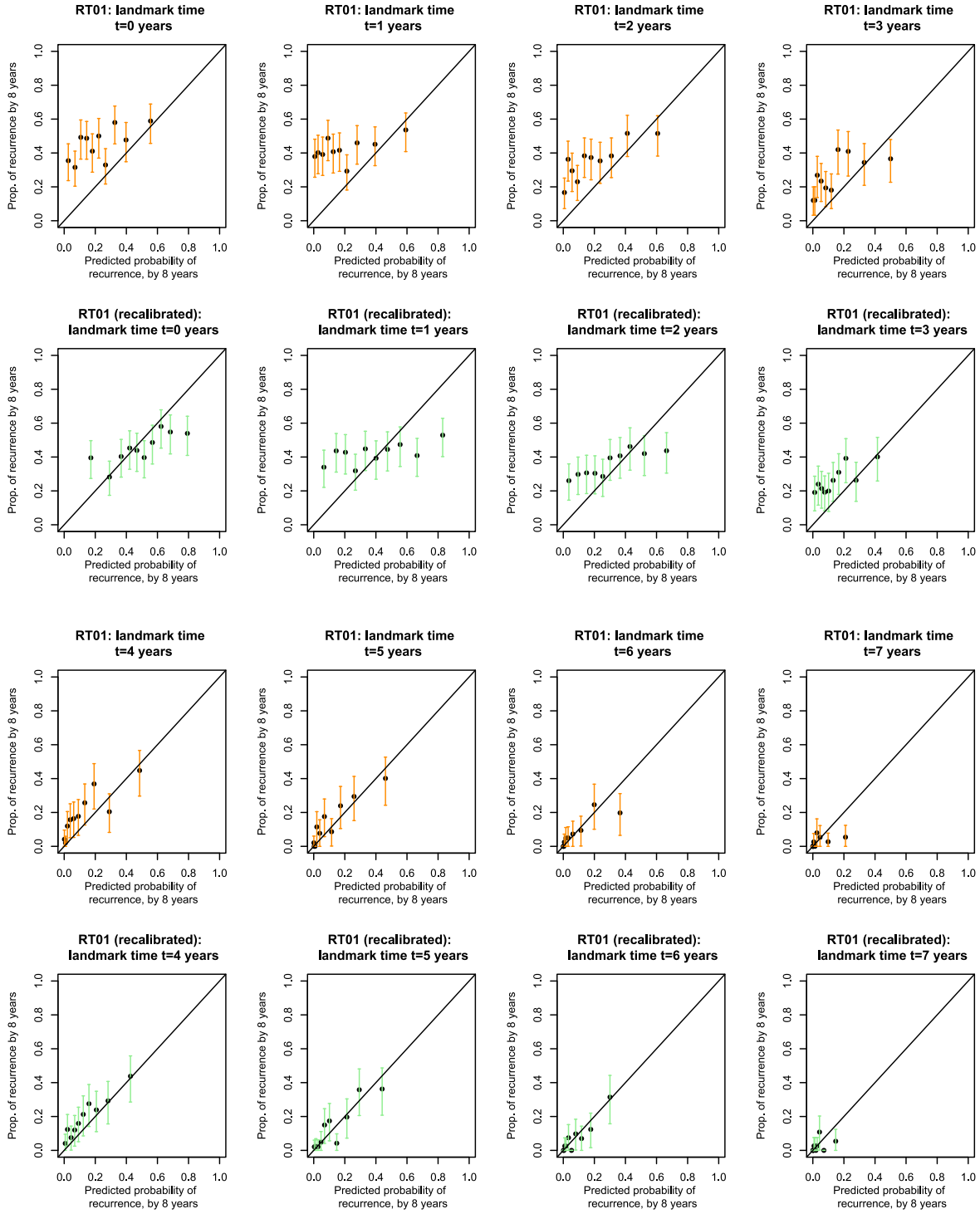


Figure 5-8 visually assessing calibration-in-the-large via graphical calibration plots of the RT01 cohort, before (1st & 3rd rows, orange) and after recalibration (2nd & 4th rows, green), for a fixed horizon of eight years at landmark zero to three years (top panel) and four to seven years (bottom panel).

5.3.4 Gleason imputation sensitivity analysis for RT01

I assess here different approaches to imputing Gleason grade 3+4 or 4+3 to patients with reported Gleason score of 7 in RT01, using MICE. I compared the predicted outputs of those imputations to the predicted outcomes when they are assigned all to 3+4 or 4+3. The hypothesis here is that these assignments for a Gleason of 7 should not make a huge overall difference to the predictive performance of this cohort, as the predictions are mainly driven by the accrued longitudinal PSA biomarker. **Figure 5-9** depicts the external predictive performance of the metrics of AUC, Brier, and the recalibrated ICI, comparing the MICE imputations to forced 3+4 or 4+3 imputations, using a fixed prediction horizon of 8 years, from landmarks 0 to 7 years. Visually there is virtually no difference to any of the predictive metrics regardless of the parameterisation of imputation undertaken, as the lines for each metric and imputation method are almost superimposed onto one another.

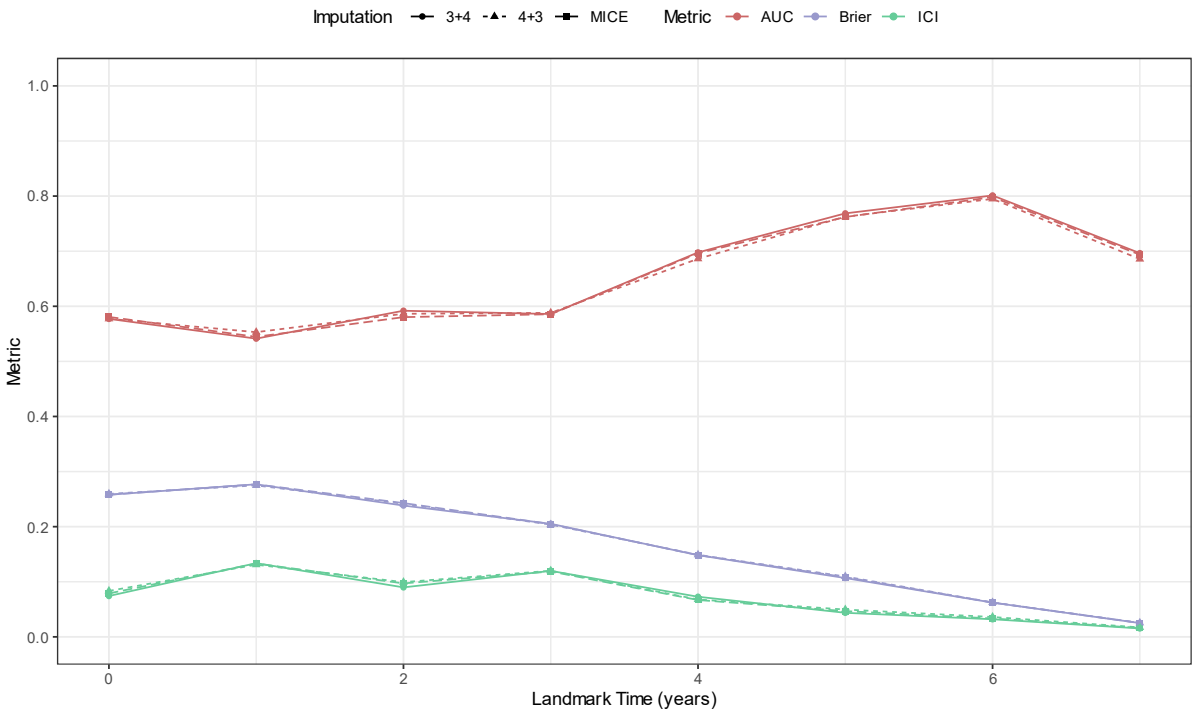


Figure 5-9 predictions assessed using AUC, Brier, and recalibrated ICI, of RT01 Gleason scoring imputed levels via MICE, comparing to Gleason scores of 3+4=7 or 4+3=7. AUC = area under the receiver operating characteristic curve; ICI = integrated calibration index; MICE = multiple imputation by chained equations.

5.4 Discussion

In this chapter, I have performed a rigorous external validation of the CDPJM using unseen data of almost 1,900 patients from the two RCT cohorts. In these cohorts, their baseline covariates and accrued longitudinal PSA values were accounted for and used to extract predictions from the CDPJM. I assessed the CDPJM's predictive performance in patients with alternative treatment pathways and appraised its generalisability in differing populations with localised (RT01), or locally advanced (RADAR) prostate cancer. This is important because it allows us to assess how well the model generalises to new data & patients and can provide an estimate of its performance on real-world applications when using PSA to inform prostate cancer prognosis post-treatment.

To assess the model's predictive performance in these external cohorts, I evaluated discrimination and calibration abilities via the AUC, Brier score, ICI, and graphical inspection. These external metrics were compared to the internal validation metrics shown in **Chapter 4,** which appear to give similar and reasonable discrimination, calibration, and overall prognostic performance; with RADAR out-performing the other trials on the former. This may be due to the more advanced nature of the disease, so that it is more straightforward to distinguish those pairs of patients who are predicted to relapse, or do not. For the Brier score, there were bigger discrepancies between the internal bias-corrected and external cohort metrics, certainly in the earlier landmark times, although these differences appeared to reduce over follow-up. As the Brier score can be decomposed of both discrimination and calibration components [94], and given that the AUC metrics were similar to the internal validity, this indicates there was some miscalibration.

Calibration was widely assessed at various prediction times, using a fixed prediction horizon of eight years, and a fixed prediction window of two and five years. That is, "if I have yet to display recurrence of prostate cancer after t years, what is my prognosis in the subsequent two or five years?". Recalibrating the baseline hazard for each external cohort largely reduced miscalibration. Calibration-in-the-large improved from landmark four years and onwards of accrued PSA data, regardless of the prediction combination considered. For RT01 there may be more miscalibration in the latter landmark times (e.g. $\pi_i(8+2|8)$) as there were fewer patients at risk ($t = 8, n = 351$) with even fewer events thereafter. Moreover, PSA collection

post-five years was sparse (last PSA time median of 4.5 years, (IQR=2.1—8.1), adherence to PSA testing was 90% at 5 years, 76% at 10 years [47] and these missing values beyond this may not have been completely random, or at least systematically missing, and may go some way to explaining the drop in AUC at the 7-year landmark. RADAR appeared to be more robust to these departures given its longer follow-up, and over double the number of PSA observations, despite having only a slightly larger sample size (20% increase). This goes to show the importance of maintaining rigorous follow-up and data collection, which provides a wealth of information many years after initial treatment.

Another pertinent component of external validation is that it allows utility of the model to be assessed in other geographical locations than the one where the model was trained, as it is the case for RADAR (Australia). This is important because models can sometimes be affected by differences in the characteristics of the data from different geographical regions, such as variations in the population, healthcare systems, and environmental factors. This may go some way to explaining the larger calibration disparities of RADAR compared to RT01 and CHHiP, which were largely based in the UK, as well as RADAR being a higher risk patient population.

In RT01, the definition of biochemical failure was different to today's Phoenix definition. There were more biochemical failure events using their threshold of 2ng/mL (assuming a rise from the nadir concentration by 50% or more), which is lower than the nadir + 2 ng/mL Phoenix definition threshold. A brief sensitivity analysis performed using the Phoenix definition resulted in fewer events, from 365 originally reported [47] to 293. As these lower values were not centrally reviewed (nor with a confirmatory PSA), it may have included some bounces, so there may have been even fewer true PSA failures. This could be a reason why my model underpredicted probability of recurrence in the RT01 calibration plots, due to the 2ng/mL definition 'inflated' number of events in this trial; furthermore, achieving a failure of 2ng/mL was quicker, compared to the training CHHiP dataset Phoenix threshold, and therefore could elicit some lead-time bias. Previous work has shown similar sensitivity and specificity for each of these biochemical definitions to predict subsequent clinical failure, therefore this validation should be generalisable to slight variations of the biochemical failure definition [191]. Given the external validation was in line with the bias-corrected measures, the predictions of biochemical failure are generalisable, and given the similar PSA trajectories between CHHiP

and RT01, the PSA thresholds given in **Chapter 4.4.6** look to be applicable. As RADAR is a more high-risk patient population, these thresholds may not hold. However, the lower PSA trajectories exhibited, due to the longer hormone schedules, that it may be the case that lower PSA thresholds are indicative of prolonged event-free survival, for example a PSA $\leq$ 0.17 ng/mL for the 6-month and a PSA $\leq$ 0.14 ng/mL for the 18-month hormone schedule, after PSA recovery seen from 3 years and onwards (**Figure 5-3**).

Another important consideration when one is developing a clinical dynamic predictive tool is that there is an adequate sample size and number of events, as large sample sizes are needed to provide more reliable and accurate results. When there are low sample sizes, and consequently events, the results of the study may not be representative of the overall population, and findings may not be generalisable. There has been recent research on calculating the minimum sample size needed for external validation of a clinical prediction model, proposed by Riley and colleagues [192]. Though their study is not quite applicable to the dynamic nature of this thesis (as they address models with baseline predictors only), they do address predictions at multiple time points. I replicated their methods in a simulation study. Assuming a target standard error of ~0.1 for the calibration slope, this corresponds to a minimum required sample size of 20,000 (see ***Appendix B***). This is of course beyond the external cohort sizes considered here. The high sample size required can be attributed to the high censoring rate observed in the development cohort, with an 83% censoring rate by 8 years of follow-up. In the external cohorts the censoring rates are lower compared to CHHiP; the cohorts are roughly ⅔^{rds} of the development sample size and have a combined total of 43% more events. As these external cohorts are finite in size, it may be possible in future to have incorporated many more patients following CHHiP's current moderately hypofractionated treatment regime from observational routinely collected clinic data / electronic health records, to further expand the sample size as required from Riley and colleagues' minimum external sample size method. Though with most data sharing arrangements, permissions to obtain these pose challenges and are not always readily available. It is worth noting that this was conducted using the linear predictor distribution from the survival submodel only; it is known that joint modelling is more efficient and therefore has higher power and thus yields smaller samples sizes [33,193]. The degree of efficiency has not been assessed for sample size

calculations of external validation cohorts, as far as this author is aware. This could be an area of future work to ascertain whether joint modelling procedures do considerably reduce the minimum required sample size for external validation of a predictive model and quantify by how much.

Another limitation of this external validation work was the lack of hypofractionated treatment regimens used as external cohort studies. Other large trials that could be incorporated to assess the CDPJM in external hypofractionated cohorts would be the PROFIT and RTOG-0415 RCTs [12,14]. However, in these trials, patients were not treated with hormone therapy, unlike in RADAR & RT01.

Recalibration was done by fitting cohort-specific joint models, as was done by Tomer and colleagues [194]. This is different from other methods, like the one proposed by Crowson that uses the linear predictor offset method, which is more straightforward to do in the standard baseline survival-only modelling framework. The cohort-specific fitting method is necessary to fit a parametric baseline hazard, estimated using penalised B-splines, for both cohorts and extract each of the baseline hazards to impute into the reduced CDPJM.

The most computational aspect is modelling the trajectory function of the longitudinal PSAs, via the mixed-effects model. To capture the nonlinear nature of PSA, natural cubic splines are used: the same parameterisation was used to develop the cohort-specific joint models, as in **Chapter 4**, i.e., to use four internal knots. This may have been too many and perhaps gives rise to overfitting in the smaller datasets, particularly in RT01 where there were only around 8k PSA observations, compared to CHHiP's 46k. In order to accommodate fewer recordings, the reduced model could have been simplified to incorporate a reduction in the number of knots, or a fully specified parametric form used, such as a biphasic- or exponential-decay-growth model; this could allow for further degrees of freedom available to estimate the baseline hazard B-spline estimation via the penalty matrix and/or number of knots. However, I would suggest that this does not overly impact the parametric estimation of the underlying baseline hazard, which is the only component extracted.

In RT01, only the total Gleason score was collected, rather than its individual scores, therefore there was a requirement to impute patients with a known Gleason score of 7 to either 3+4 or

4+3. For this, I used the MICE approach, and compare its predictive performance in section **5.3.4** to simpler approaches. Regardless of imputation method used (MICE, or all to 3+4 then 4+3), either approach made almost no difference to the predictive performance of the model, albeit this was only examined under a fixed prediction horizon of eight years. This lack of difference in the predictive performance would be expected to carry through to alternative prediction windows. This was as expected given that the predictions are, by and large, driven by the acquired longitudinal PSAs. It is also encouraging that the sensitivity analysis showed little difference in these metrics, demonstrating that if there are some differences in the collection of baseline prognostic factors, which will inevitably change over time, these can be accommodated for.

Another pertinent caveat is that the Gleason pathology scoring system itself has also changed considerably over the last two decades. In RT01, Gleason scores that were allocated ≤ 6 then are likely now to be classed ≥ 7, as shown in a recent pathological review of CHHiP, where the proportion of Gleason scores of 6 or below have substantially reduced. Further work to incorporate these changes to RT01 would require central pathological review to re-score Gleason to current standards. Regardless, it has been demonstrated that the model can be adapted, with the amendment of the assigned categorical levels to flexibly deal with those temporal changes to the Gleason classification system, with no detrimental effect on predictive performance. The predictions for individuals may change themselves, but the ranking of the linear predictor does not change much and therefore very little differences are seen in the appraisal of the validation when these differing parameterisations of the Gleason are considered.

To conclude, in this chapter I successfully externally validated the CDPJM to predict recurrence of prostate cancer. To my knowledge, this is the first rigorous external validation of a CDPJM for the prognosis of localised prostate cancer treated under both short- and long-course hormonal therapy and radical radiotherapy. Predictions and calibrations are optimal after at least 4 years of accrued longitudinal data follow-up. Recalibration is advised for external cohorts that have worse prognosis and/or no hypofractionation received; recalibration is not expected to be required for patients undergoing current standard-of-care hypofractionation. If recalibration is necessary for new patients, then given their type and dose

of radiotherapy, hormone duration, and baseline risk factors, the cohort they are most similar to can be used for their underlying baseline hazard. At the time of development, only the CHHiP dataset was available. Now that all the cohorts are available, they could be utilised in further model development. Cohort-specific CDPJMs were required to be developed to estimate their baseline hazard for recalibration; one could argue to create an ensemble of predictive models with (Bayesian) model averaging, and / or to create a CDPJM developed using all the cohorts, with some patients held back for external validation [45].

Chapter 6 – Competing Risks Joint Models

6.1 Introduction

In previous chapters, I developed the CDPJM and validated its dynamic predictions for recurrence of disease in localised prostate cancer. I considered recurrence to be a composite event of biochemical failure, clinical recurrence or death due to prostate cancer. Deaths due to other causes in the absence of recurrence were considered as non-informative censored observations. It was noted in **Chapter 4** – Development of a Personalised Clinical Dynamic Predictive Joint Model to Characterise Prognosis for Patients with Localised Prostate Cancer Patients: Analysis of the CHHiP Phase III Trialthat, given there was little difference in the cumulative incidence estimators obtained with Kaplan-Meier ($1 - KM$) or competing risks methodology, given this and relatively few of these competing events, it was reasonable to treat those deaths as censored. Assuming non-informative censoring, however, may be unrealistic: indeed, when censoring at a given time, it is assumed that the event of interest will occur at some point in the future. But if a patient dies due to non-disease causes, their risk of recurrence is zero; it is erroneous to treat them as censored as it becomes a missing data issue where there is some (incorrect) contribution to the risk of recurrence.

In general, in the localised prostate cancer setting, it is unlikely that the first disease-related event observed would be a death, as recurrence will normally be characterised first by biochemical failure or progression to advanced disease. However, as this is an ageing population, and patients are followed for a long time, comorbidities occur, and patients die due to other causes. In this setting there is one event, non-cancer related death, which is a competing event for recurrence (**Figure 6-1**), because its occurrence may preclude the observation of the event of interest. Competing risks methodology would account for events that may hinder the observation of interest and provide unbiased estimates of the cumulative probability of recurrence over time [167,195,196].

Joint models have also been extended to the competing risk setting [69,88,111,179]. In this chapter, I explore the impact of using a competing risk joint model in the dynamic predictions for risk of recurrence, comparing these with the predictions obtained by the standard CDPJM developed in **Chapter 4** – Development of a Personalised Clinical Dynamic Predictive Joint Model to Characterise Prognosis for Patients with Localised Prostate Cancer Patients: Analysis

of the CHHiP Phase III Trial. For both frameworks, I compare the parameter estimates of the models, the predictions for individual patients and the goodness-of-fit for each model. I then extend the appraisal metrics presented in **Chapter 2** to evaluate predictive performance assuming a competing risks joint model.

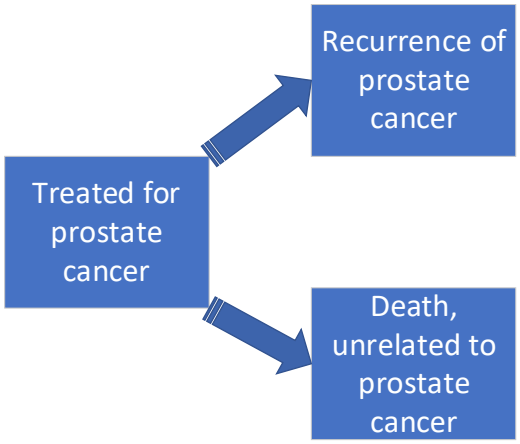


Figure 6-1 a graphical representation of a competing risk model with two causes (K=2): recurrence of prostate cancer (k=1), or death unrelated to prostate cancer (k=2).

6.2 Methodology

The interest is in the joint distribution of the time-to-failure T^* and the cause of failure δ (also denoted type of event). This distribution can be characterised by the cause-specific hazard functions for each type of event, representing the hazard of failing from a given cause in the presence of the competing event:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T^* < t + \Delta t, \delta = k \,|\, T^* \geq t)}{\Delta t}$$

The cumulative incidence function $\mathrm{CIF}_k(t)$ is the probability of failing from cause k by time t, and can be inferred from the cause-specific hazards,

$$\mathrm{CIF}_k(t) = \Pr(0 < T^* \leq t, \delta = k) = \int_0^t h_k(s) \cdot S(s)\, ds.$$

where $S(t) = P(T^* \geq t) = \exp\left(-\sum_{k=1}^{2} \int_0^t h_k(s) ds\right)$ is the survival function of T^* (which depends on both cause-specific hazards).

The observed data in this setting is, for each patient, $T_i = \min(T_i^*, C_i)$ (where, as before C_i indicates the censoring time) and the event indicator taking values $\delta_i \in \{0,1,2\}$ with 0

corresponding to those who are censored, 1 to the primary event of interest recurrence, and 2 to the competing event non-cancer death. The observed data permits estimation of these functions for each cause, recurrence from biochemical/clinical failure, and unrelated death [167,196].

6.2.1 Competing risks joint model

The longitudinal model is as defined in **Chapter 4** – Development of a Personalised Clinical Dynamic Predictive Joint Model to Characterise Prognosis for Patients with Localised Prostate Cancer Patients: Analysis of the CHHiP Phase III Trial with no changes in its parameterisation. For a competing risk submodel, the cause-specific hazard function of interest (recurrence, $k = 1$) and also for the competing event (unrelated death, $k = 2$) is defined,

$$h_{ik}(t|\boldsymbol{M}_i(t), \boldsymbol{w}_i) = h_{0k}(t)\exp\{\boldsymbol{\gamma}_k^T\boldsymbol{w}_i + f(\boldsymbol{M}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}_k)\} \;\; k = 1, 2.$$

Cause-specific parameters $(\boldsymbol{\gamma}_k^T, \boldsymbol{\alpha}_k)$ and the baseline hazard function (for each event) need to be estimated, extending the same principles as presented in **Chapter 2**. The likelihood can be calculated straightforwardly with software by transforming the analysis data set in a long format, such that each patient has k rows with an indicator whether that row outcome has been met [67]. The competing risk joint model is developed using the *JMbayes2* **R** package (v0.3-0) with the default settings and priors [136]. There are four parallel chains and 27,500 iterations per chain and a burn-in of 2,500 iterations with a thinning of including every 5th iteration.

6.2.2 Assessing predictive performance

Predictions for competing risk joint model can be obtained [69,179]. I wish to predict the probability that a new patient (indexed by l) experiences an event of type k by time u, given that the patient is event-free at $t < u$, and given their baseline risk factor values $\boldsymbol{w}_l$ (such as age, tumoural severities, and treatment received), and provided a set of longitudinal PSA biomarker values up to time t ($\boldsymbol{y}_l(t)$):

$$\pi_{lk}(u|t) = \Pr(T_l^* \leq u\,, \delta_l = k|\, T_l^* > t,\;\; \boldsymbol{y}_l(t), \boldsymbol{w}_l, \boldsymbol{D}_n\,),$$

where $\boldsymbol{D}_n = \{T_i, \delta_i, \boldsymbol{y}_{in}; i = 1, \ldots, n\}$ is the observed sample data that the competing risk joint model was fitted on. As the Bayesian framework will be utilised, the expectation of $\pi_{lk}(u, t)$ can be estimated by using the corresponding posterior predictive distribution,

$$\pi_{lk}(u|t) = \int \underbrace{\Pr(T_l^* \leq u, \delta_l = k | T_l^* > t, \ \boldsymbol{y}_l(t), \boldsymbol{w}_l; \boldsymbol{\theta})}_{A} \underbrace{p(\boldsymbol{\theta}|\boldsymbol{D}_n)}_{B} \, d\boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ is a vector of parameters for the entire joint model. The second term (B) of the integrand, $p(\boldsymbol{\theta}|\boldsymbol{D}_n)$, is the posterior distribution of the parameters, given the observed data. The first component (A) in the integrand can be expanded as,

$$A = \int \Pr(T_l^* \leq u, \delta_l = k | T_l^* > t, \boldsymbol{y}_l(t), \boldsymbol{w}_l \, \boldsymbol{b}_l; \boldsymbol{\theta}) \, p(\boldsymbol{b}_l | T_l^* > t, \boldsymbol{y}_l(t), \boldsymbol{w}_l; \boldsymbol{\theta}) \, d\boldsymbol{b}$$

$$= \int \Pr(T_l^* \leq u, \delta_l = k | T_l^* > t, \boldsymbol{w}_l, \boldsymbol{b}_l; \boldsymbol{\theta}) \, p(\boldsymbol{b}_l | T_l^* > t, \boldsymbol{y}_l(t), \boldsymbol{w}_l; \boldsymbol{\theta}) \, d\boldsymbol{b}.$$

The transition to the final line is due to the full conditional independence assumption as described in **Chapter 2.3.1**. Furthermore,

$$A = \int \frac{\Pr(T_l^* \leq u, \delta_l = k, \ T_l^* > t | \ \boldsymbol{w}_l, \boldsymbol{b}_l; \boldsymbol{\theta})}{\Pr(\ T_l^* > t | \ \boldsymbol{w}_l, \boldsymbol{b}_l; \boldsymbol{\theta})} \, p(\boldsymbol{b}_l | T_l^* > t, \boldsymbol{y}_l(t), \boldsymbol{w}_l; \boldsymbol{\theta}) \, d\boldsymbol{b}$$

$$= \int \frac{\Pr\left(t < T_l^* \leq u, \delta_l = k\right)}{S(t)} p(\boldsymbol{b}_l, \delta_l = k | T_l^* > t, \boldsymbol{y}_l(t), \boldsymbol{w}_l; \boldsymbol{\theta}) \, d\boldsymbol{b}$$

An estimate of $\pi_{lk}(u|t)$ can be extracted using a Monte Carlo simulation scheme by drawing from the posterior distribution of (B) and the random effects $p(\boldsymbol{b}_l|.)$; given those two draws $\pi_{lk}(u|t, \boldsymbol{b}_l^*, \boldsymbol{\theta}^*) = \frac{\Pr\left(t<T_l^* \leq u, \delta_l = k\right)}{S(t)}$. This is repeated M times to generate the posterior distribution for π_{lk} and hence the mean and α% credible intervals can be calculated from the Monte Carlo sample [179].

Predictive performance of the competing risk joint model to predict recurrence of disease is performed, using the predictions obtained in the presence of the competing event. The model-based performance metrics that were introduced for the CDPJM in **Chapters 2**, **4**, **5** are updated by using the predictions obtained from the competing risk joint model.

Blanche and colleagues proposed alternative estimators of predictive performance of joint models in the presence of censoring and competing risks [87,88]. Nonparametric inverse probability censoring weighting (IPCW) is used to estimate the dynamic Brier score and AUC, with the advantage that they are model-free with no assumption of the exactness of the specification of the competing risk joint model. I compare these IPCW metrics with the model-

based ones and also compare to the standard model-based validation metrics given in **Chapter 4**.

The AUC & Brier IPCW estimators are similar to those defined in **Chapter 2** but with a weighting applied,

$$\widehat{AUC_{k=1}}(u|t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{I}_{\pi_{ik}(u|t)>\pi_{jk}(u|t)}\, \widetilde{D}_{ik}(u|t)(1-\widetilde{D}_{jk}(u|t))\widehat{w}_i(u|t)\widehat{w}_j(u|t)}{\sum_{i=1}^{n}\sum_{j=1}^{n} \widetilde{D}_{ik}(u|t)(1-\widetilde{D}_{ik}(u|t))\widehat{w}_i(u|t)\widehat{w}_j(u|t)}$$

$$\widehat{\mathrm{BS}_{k=1}}(u|t) = \frac{1}{\sum_{i=1}^{n}\mathbb{I}_{T_i>t}} \sum_{i=1}^{n} \widehat{w}_i(u|t)\left(\widetilde{D}_i(u|t) - \pi_{ik}(u|t)\right)^2.$$

In the above, the estimator $\widetilde{D}_{ik}(u|t) = \mathbb{I}_{t<T_i\leq u,\delta_i=k}$ equals 1 when patient(s) i, j is known to have experienced an event of type $k = 1$ between time t and time u; or 0 when either the subject(s) i, j experiences a competing event within the time interval or is event-free in the interval. To account for censoring, the weighting is defined to be,

$$\widehat{w}_i(u|t) = \frac{\mathbb{I}_{T_i\ >u}}{\widehat{G}(u|t)} + \frac{\mathbb{I}_{t<T_i\ \leq u}\mathbb{I}_{\delta_i\ \neq 0}}{\widehat{G}(T_i\ |t)}.$$

Where $\widehat{G}$ is the Kaplan–Meier estimator of the survival function of the censoring time; $\widehat{G}(u|t) = \widehat{G}(u)/\widehat{G}(t)$ estimates the conditional probability of being uncensored at time u given not being censored at landmark time t. $\mathbb{I}_E$ is an indicator function (1 when expression E is true, 0 otherwise). Corresponding pointwise confidence intervals can be extracted by the asymptotic normality assumption, using the central limit theorem [88].

6.3 Results

6.3.1 Cumulative incidences

In **Figure 6-2**, the non-parametric estimates of the cumulative incidences for each type of event are presented under a competing risks framework, and these are compared to the estimates from the Kaplan-Meier estimator, censoring the competing event at the time it occurs. For each outcome, both $1-\mathrm{KM}$ and cumulative incidence CR estimators are superimposed up until around 5—6 years, after which there is some slight divergence, indicating that the effect of these unrelated deaths may have a greater impact on predictions for times after 5 years. For

the cumulative incidence of recurrence (red) at 10 years, the 1 − KM is 0.23 (95% CI=0.21–0.25) *vs* the cumulative incidence CR of 0.22 (95% CI=0.20–0.23).

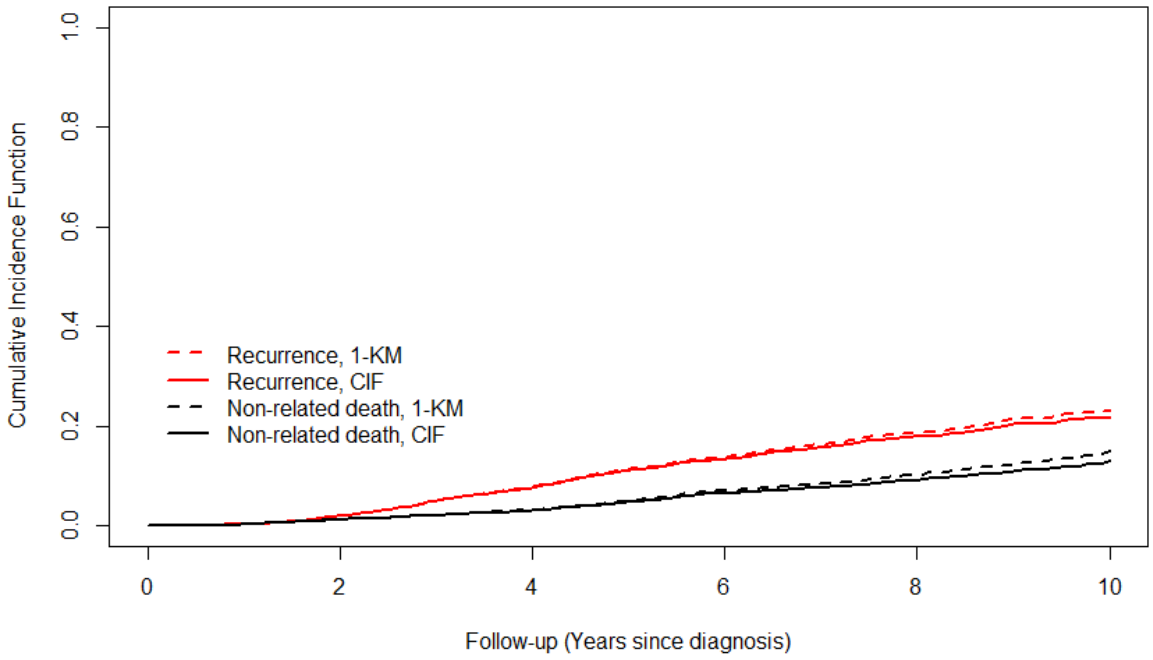


Figure 6-2 compares the 1-KM and cumulative incidence function estimators of each outcome (recurrence or death unrelated to prostate cancer).

6.3.2 Competing risk joint model parameter estimates

The joint model parameter estimates for the time-to-event submodels, comparing the standard CDPJM and the competing risk joint model, are presented in **Table 6-1**. In the longitudinal submodel, as expected there is very little difference in the parameter estimates, as this contains the exact same parameterisation from the two joint models (see **Table 4-3**). For both the competing risk joint model and CPDJM, the parameter estimates for the recurrence time-to-event submodels are very similar**.** For the unrelated deaths time-to-event submodel, it is noted how the parameter estimates for tumour features stage and grade are statistically significant and negative, indicating a 'protective' effect of more severe disease presentation on risk of unrelated death as a first event. This protective induced effect has been reported in the literature, and simply shows how patients with more severe stage and grade are more likely to experience recurrence first, rather than dying *first* of non-disease related causes [197,198]. A similar phenomenon is observed for bicalutamide, patients who received bicalutamide appeared to have lower risk of first event of recurrence (HR=0.70) compared to LHRHa (conditional on all other prognostic factors and PSA being fixed). Therefore, it could be more

likely to observe a non-prostate cancer death as the first event in this group of patients. However, as stated in **Chapter 4**, it is worth reminding that only 13% of patients received bicalutamide and they presented with generally better prognostic factors than those receiving LHRHa and hence likely confounds this results and explains why they are less likely to recur as a first event [17]. Further discussion and explanation of the bicalutamide findings can be found in section **6.4**.

The only independent prognostic factor of unrelated deaths from those considered is age. This is not unexpected. Each additional year above 69 years of age (mean in the population) significantly increases the risk of both recurrence and unrelated death, albeit with a small absolute increase in risk (4-5%).

I also quantified the association of the trajectory of PSA, i.e., using its value and rate-of-change (**Table 6-1**). The log-hazard ratio parameter estimates for each joint model for the value (log-$HR_{standard}$ = 4.52, log-HR_{CR} = 4.61, both $p<0.001$), and for the gradient (log-$HR_{standard}$ = 2.08, log-HR_{CR} = 2.16, both $p<0.001$) are remarkably similar for the impact on risk of recurrence. The competing risk joint model also quantifies the impact of PSA upon the competing risk of death. Specifically for the value of PSA, log-HR_{CR} = -0.39, $p<0.001$, i.e., a unit increase in log(PSA) reduces the risk of observing death unrelated to prostate cancer as a first event by $32\% = 1 - \exp(-0.39)$. Again, this seems to be a protective effect induced by PSA being a strong predictor for recurrence as a first event, as there is no biological rationale for PSAs to be a protective effect for the competing event.

Table 6-1 parameter estimates of the time-to-event outcomes, comparing the competing risk joint model specification (left) to the standard joint model (right, developed in chapter 4); ref = reference level, HT = hormone therapy, LHRHa = Luteinizing-hormone-releasing-hormone analogue, PCa = prostate cancer.

Survival submodel	**Competing risk JM**						**Standard JM (Chapter 4)**					
Outcome: PCa recurrence	**log-HR**	**SD**	**2.50%**	**97.50%**	**p-val**	$\widehat{R}$	**log-HR**	**SD**	**2.50%**	**97.50%**	**p-val**	$\widehat{R}$
Arm: 57gy/19f (ref 74gy/37f)	-0.023	0.21	-0.428	0.384	0.909	1.002	-0.013	0.213	-0.433	0.403	0.953	1.001
Arm: 60gy/20f	0.007	0.176	-0.334	0.351	0.982	1.003	0.009	0.18	-0.343	0.365	0.966	1.007
Gleason score: 3+4 (ref ≤6)	0.6	0.152	0.302	0.9	<0.001	1.004	0.595	0.161	0.287	0.91	<0.001	1.001
Gleason score: 4+3	1.025	0.174	0.686	1.371	<0.001	1.003	1.016	0.182	0.667	1.374	<0.001	1.003
Gleason score: 4+3	0.929	0.331	0.283	1.573	0.006	1.003	0.914	0.346	0.24	1.601	0.009	1.002
T-stage: T2 (ref T1)	0.382	0.141	0.1	0.652	0.009	1.008	0.382	0.149	0.098	0.679	0.006	1.002
T-stage: T3	0.884	0.217	0.46	1.31	<0.001	1.002	0.879	0.23	0.419	1.325	<0.001	1.004
HT: 150mg bicalutamide (ref LHRHa)	-0.359	0.181	-0.72	-0.006	0.046	1.002	-0.36	0.191	-0.738	0.003	0.053	1.003
(Age-69) yrs	0.052	0.01	0.032	0.072	<0.001	1.002	0.052	0.011	0.03	0.073	<0.001	1.001
log(PSA(t))	4.614	0.228	4.197	5.08	<0.001	1.015	4.521	0.233	4.072	4.987	<0.001	1.007
d log(PSA(t)) / dt	2.155	0.18	1.814	2.526	<0.001	1.044	2.075	0.176	1.737	2.43	<0.001	1.008
Outcome: death unrelated to PCa												
Arm: 57gy/19f (ref 74gy/37f)	-0.078	0.238	-0.544	0.381	0.744	1.002						
Arm: 60gy/20f	-0.169	0.223	-0.607	0.261	0.455	1.002						
Gleason score: 3+4 (ref ≤6)	-0.607	0.2	-1.001	-0.208	0.002	1.003						
Gleason score: 4+3	-1.203	0.242	-1.671	-0.727	<0.001	1.003						
Gleason score: 4+3	-1.144	0.477	-2.099	-0.223	0.016	1.002						
T-stage: T2 (ref T1)	-0.328	0.187	-0.692	0.046	0.082	1.008						
T-stage: T3	-0.706	0.303	-1.311	-0.126	0.016	1.003						
HT: 150mg bicalutamide (ref LHRHa)	0.446	0.249	-0.045	0.934	0.074	1.002						
(Age-69) yrs	0.042	0.015	0.013	0.07	0.004	1.003						
log(PSA(t))	-0.385	0.097	-0.579	-0.197	<0.001	1.005						
d log(PSA(t)) / dt	0.27	0.135	0.01	0.543	0.042	1.005						

6.3.3 Predictive performance

I appraise the predictive performance of the competing risk joint model for the main outcome recurrence using both model-based performance metrics and the IPCW approach, and compare the results with the standard CDPJM (obtained in **Chapter 4**). **Figure 6-3** evaluates and compares the apparent predictive performance (AUC & Brier score) for the predictions of recurrence by eight years.

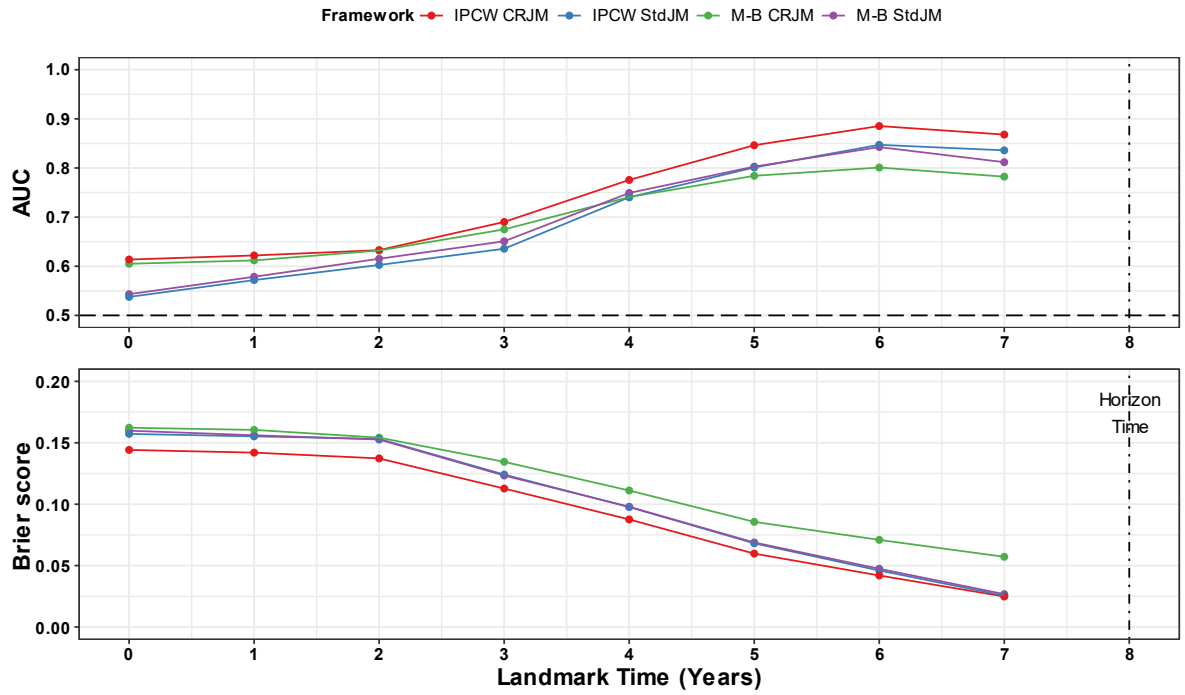


Figure 6-3 compares the IPCW and M-B methods applied to both the StdJM and CRJM in assessing each of its apparent validation metrics for the AUC (top) and the Brier scores (bottom) at each landmark time to predict up to a horizon time of eight years. AUC = area under the receiver operating characteristic curve; IPCW = inverse probability of censored weighting; M-B = model-based; CRJM = competing risk joint model; StdJM = standard joint model (chapter 4).

For the competing risk joint models, the AUC metrics are similar to one another in the first three landmark years, then the IPCW is superior to the model-based approach from three years onwards. These competing risks metrics supersede the standard model AUCs in the first three years, with the IPCW competing risk performing best across all landmarks, with a maximum AUC of 0.89 at landmark 6 years. The standard model-based AUCs are superior to the model-based competing risk from landmark time four years and onwards.

For the overall predictive performance, the Brier loss function score follows a similar pattern, with all estimates being very similar at all landmark times; a reduction in the prediction error

is seen from landmark time at three years and onwards. The lowest (best) Brier score is given by the IPCW competing risk framework across all landmarks, the model-based competing risk gives the highest scores, particularly at the latter landmarks. The standard (non-competing risk) joint model, both frameworks give near-identical scores.

6.4 Discussion

In this chapter, I have further extended the CDPJM to explicitly consider the competing risk of death unrelated to prostate cancer, rather than treat that outcome as censored. This is theoretically to correctly consider those deaths as a terminal event. However, I compared the cumulative incidences for each outcome using a standard one minus Kaplan-Meier estimator, to the competing risk cumulative incidences which found to be nearly identical until five years of follow-up and not to be considerably different at ten years. Though these are small, the differences are expected to rise given the ageing population of patients in the CHHiP trial. I produced examples of the dynamic predictions, similarly to **Chapter 4.4.4**, where predictions of the two competing outcomes are depicted, however the focus is not in predicting the competing risk outcome of an unrelated death but predicting risk of recurrence accounting for the competing event (see dynamic predictions presented in ***Appendix C***, ***Supplementary Figure C1***).

The apparent validation metrics were by and large similar between the standard and competing risks joint models. It was shown that the IPCW competing risk metrics had the best appraisals of the competing risk joint model. It is worth noting that the predictions of recurrence from the standard IPCW- and model-based metrics were remarkably similar across all landmarks.

The cause-specific hazard ratios between the standard and competing risks joint models for recurrence are expectedly similar, as the predictions for the primary outcome should not change much. The effect of bicalutamide appearing to reduce the risk of recurrence, compared to LHRHa, is as observed in **Chapter 4**, whilst conditioning on PSA and prognostic factors. The borderline significance of this result is somewhat surprising, as it is known to be less effective than LHRHa in advance-stage disease [199,200] and no known improvement in biochemical failure rates in a case-control matched study [201]. There are no known randomised trials comparing LHRHa and bicalutamide in the radiotherapy setting, however

a large meta-analysis of almost 10,000 men showed that bicalutamide in addition to the standard of care (radical prostatectomy, radiotherapy, or watchful waiting) was not beneficial compared to the addition of androgen deprivation therapy [202]. Selection bias is a factor in this nonrandomised comparison, which confounds this result. These were younger patients with proportionately less tumour in their biopsies and would have naturally better prognosis thus conditioning on covariates related to positive core biopsy data may remove the observed bicalutamide effect [17,203].

For the competing risk event, it is observed that worse prognostic factors for recurrence have a significant protective effect on death, i.e., for those patients who are more likely to experience recurrence first, they would be less likely to experience an unrelated prostate cancer death *before* recurrence. Conversely, those older patients with lower PSAs and better recurrence prognostic risk factors are more likely to experience non-prostate cancer related death first. In addition, it appears bicalutamide raises the risk of non-prostate cancer related deaths, compared to LHRHa. These results are likely explained by the better prognosis of the patients who received bicalutamide, therefore implying a greater probability of observing (not causing) the competing event deaths first, though this was not significant at the 5% level (p-val = 0.074). In this study, other prognostic factors such as smoking history, socioeconomic, cardiovascular risk profiles, treatment decision rationales, and other comorbidities were not captured which could be useful in explaining these competing events. It was noted in **Table 6-1** that bicalutamide appeared to increase risk of the competing event. One possible explanation for this result is that many patients who received bicalutamide were from Belfast, Northern Ireland. Where according to the Office of National Statistics (ONS), male life expectancy is slightly lower in there, compared to England where the majority of patients were treated, potentially confounding this result [204]. As expected, the only true prognostic factor associated with the competing event is age, which is also associated with recurrence.

I considered model and IPCW-based metrics to appraise the predictive performance of the joint models. There were some differences, particularly in the AUC for the performance metrics of the competing risk joint model between the modelling-based and IPCW approaches. There were some modest improvements in the overall predictive power of the competing risk model, with a lower Brier score in the earlier landmark times, exhibiting a

similar trend. The advantage with the IPCW approach is that the estimators are inherently model-free, i.e., the correctness of the specification of the joint model used to elicit the predicted probabilities $\pi_i(u|t)$ is not assumed and therefore leads to unbiased estimates of AUC & the Brier score [88]. In the examples provided by Blanche, they use a joint latent class model which has fewer parameters in the joint model, compared to the shared-parameter joint model (i.e., no association structure is required to be defined) and hence their proposed metrics may well be better suited. However Rizopoulos and colleagues argue that there are biases that are elucidated using IPCW, including censoring being dependent on the longitudinal biomarker response that is not accounted for [55]. This could go some way to explaining the disparities between the two approaches, i.e., the model-based accuracy measures correctly accounting for such dependence on the censoring distribution.

Some limitations are that it is not trivial (e.g. computationally time-consuming) to calculate the validation metrics, certainly for the model-based approach, for a competing risk parameterisation of the joint model. For this reason, I only considered the apparent metrics for AUC and the Brier score. There have been some recent approaches to assess model performance in the presence of competing risks using a fully specified modelling approach, such as calibration using the integrated calibration index by Austin and Geloven [205,206]. However, they focus on baseline Cox proportional hazards models exclusively; it is not straightforward to apply them to competing risk dynamic joint models. Another limitation of this work is the use of apparent metrics only. No resampling methods such as cross-validation or bootstrapping were used to evaluate the CR model (nor for the external validation presented in **Chapter 5**). However, given the small differences between the corrected bootstrap and apparent metrics of **Chapter 4** (**Table 4-7** & ***Appendix A Supplementary Table A2***), I suggest there will be perhaps little differences in the metrics if resampling methods were to be applied here and find that there is in fact minimal over-optimism.

For the predictive performance metrics (section **6.2.2**), the 'control' was defined as a patient who has not had recurrence in the interval of interest, but could be censored by the horizon time or could have experienced the competing event. There is an alternative approach of defining the control as only those patients who are truly censored by the horizon time u [87,207]. I opted for the former given there are more patients included in this definition and

therefore more patients contribute towards the weights and the numerator in the Brier score, and hence give more conservative estimators for the model-based approaches.

A frequentist competing risk joint modelling framework applied to prostate cancer was developed by Ferrer and colleagues [111], presented in one of the 12 review articles of **Chapter 3**. They developed estimators for the AUC and mean squared prediction error (equivalent to the Brier score), though appraisal of these estimators is inherently model-free in the simulation studies they performed, as no censoring was considered.

I assessed predictions at $\pi_i(u = 8|t = \{0, \dots, 7\})$ and did not explore other prediction window procedures. It is possible that in an ageing population, predictions of recurrence beyond that of a horizon time of eight years would be more appropriate. It is also worth noting that the competing risk approach might perform better at latter horizon times and performance may be negligible at shorter horizon times. Regardless of the estimators used, it is apparent that the improved performance of the model at 5 years (AUC=0.78—0.85; BS=0.06—0.09) and onwards, suggests an optimum lead-time of 2 years, in line with what was presented in **Chapter 4**. The CHHiP snapshot was taken in October 2019 with a median of ~8½ years of follow-up. I would suggest in future work assessing predictions with an updated snapshot of at least a horizon of ten years or beyond and investigating other prediction windows of interest, for example two or five years from present-day landmark.

The competing risk framework is a special case of the multi-state model. It would be possible to consider intermittent progressive events similarly to Ferrer [110] which were presented in **Chapter 3.6**. This would be an interesting approach to use PSA to model the progression of the disease beyond that of biochemical failure. However as previously mentioned, progression through the states to possible distant and metastatic end stages is not the focus of this thesis.

One could consider further joint modelling extensions, such as the inclusion of a cure submodel component, whereby a proportion, or fraction, of patients are assumed to be "cured" and will not experience any recurrence ever. Survival models for a single time to event, such as the Cox proportional hazards model, assume that all patients will eventually experience the event, $S(t \to \infty) = 0$, which may not be the case for the event of interest in a competing risk setting. Localised prostate cancer is by and largely a long-term disease. It is

true that many patients will never experience recurrence, but this can never truly be observed, i.e., $\mathrm{CIF}_k(t \to \infty) = \Pr(\delta = k)$. Fitting these types of models is non-trivial, due to the assumptions required of the proportion of the cured fraction, together with the complexity of adding a further third submodel component to the joint model. However, this will certainly be of interest to clinicians in the localised setting, particularly as stereotactic body radiotherapy (SBRT) becomes the norm, maximising the efficacy of SBRT in progression-free survival; this approach should be considered for future work.

To conclude, I demonstrated that the framework to extend the joint model when considering the competing event is feasible, and predictions can be extracted reasonably easily. Despite the additional model complexity, it may give more accurate discrimination and improved predictive performance, although its improvement from the standard model may be modest and may depend on the horizon time of prediction. Interestingly, the metrics are considerably improved using the IPCW approach. Regardless of the increased complexity of the competing risk CDPJM, this will be of interest to clinicians if predictions of non-cancer related deaths are of interest; certainly an increasing proportion of these competing events will be the case at later follow-up times in an ageing population, which need to be correctly accounted for.

Chapter 7 – Concluding remarks

7.1. Summary

The overarching aim of this thesis was to develop a dynamically updated predictive model, in order to characterise a patient's personalised prognosis after radical treatment of their localised prostate cancer. Historically, traditional clinical predictive models (e.g. implemented via Cox proportional hazard modelling) were based solely on baseline information. Here I maximise the use of PSA measurements routinely collected over time, to use all the available information in addition to known baseline prognostic factors, to improve precision and predictions of prognosis dynamically.

Joint modelling using longitudinal PSA to predict recurrence is well founded and has been used previously to inform predictions of prognosis in an updated or dynamic manner [84,108,111]. There are several novel contributions made in this thesis, which are as follows. It provides a prognostic model for patients treated with hypofractionated radiotherapy, using well-curated clinical trial data from a large-scale phase III RCT, CHHiP, where two thirds of patients received hypofractionation. In addition and compared to previous studies modelling recurrence in localised prostate cancer reviewed in **Chapter 3** – Literature Review**,** CHHiP patients also received neoadjuvant and concurrent hormone therapy, which changes the PSA trajectory. Additionally, the in-treatment phase is also modelled from patients presenting PSA as the time origin and fully captured their PSAs in this thesis. To this author's knowledge, this is the first dynamic predictive tool developed using outcomes following contemporary hypofractionated radiotherapy schedules with hormone therapy. This translational prospective analysis made use of data from two additional RCTs to maximise clinical utility over-and-above answering their trial-specific hypotheses for external validation. Finally, this thesis extended the use of validation metrics of the predictive joint model in a competing risk setting, comparing the IPCW and model-based approaches in a dynamic framework.

I started by synthesising and appraising the current and relevant literature in the use of the joint modelling methodology applied to prostate cancer clinical studies in **Chapter 3.** I focused my review on the model specification for the PSA trajectories over time and the different prostate-related event intended to predict, in addition to identifying appropriate predictive

performance tools in this setting. This literature review has been published in a methodology journal [41]. In conducting the review, I identified key differences between the clinical scenario I was aiming to model in this thesis, and in the previous studies. I utilised PSA measurements at presentation of prostate cancer and throughout treatment. I directly modelled these longitudinal PSAs with a flexible cubic splines model, while previous studies used a parametric parameterisation, including an exponential decay-growth model [208,209], and variations therein [63,194]. Moreover, while these studies would model PSA trajectories from the end of conventional EBRT, in this thesis I considered RCTs where patients had hormone therapy prior, and then concurrently to hypofractionated or conventional radiotherapy schemes, so PSA was captured and modelled throughout the entirety of their treatment (i.e., from start of hormone therapy) and follow-up.

The review formed the basis and rationale in the modelling undertaken in **Chapter 4**; particularly in predicting clinical failures, often associated with an increase in the PSA biomarker post-treatment, indicating (possible) cancer recurrence. The models developed and presented in this thesis address and quantify the improvement in predictions of recurrence when considering longitudinal PSA, compared to a model with baseline prognostic and treatment factors only. I have shown that longitudinal PSA trajectories are predictive, and certainly so from three years of follow-up and onwards. The time-dependent AUROC is improved, the predictive error loss function (Brier score) as an overall measure of predictive performance is minimised, as well as improving calibration (see **Table 4-7**). I also introduced dynamic predictions for individual patients, to extract and maximise clinical utility, and derived PSA thresholds that are indicative of good prognosis. For instance, PSA $\lessapprox$ 0.23ng/mL post-nadir is prognostically particularly good at 3 years, while PSAs less than 0.34 and 0.41ng/mL at years 4 & 5 respectively are associated with minimal (<5%) risk of recurrence by 8 years [42].

Chapter 5 set out to validate the developed model of **Chapter 4**, albeit reduced to exclude treatment so that the model was applied to more general hormone therapy and radiotherapy regimes, where differences in the exact antiandrogen analogues and radiotherapy schedules could vary from CHHiP. By and large, the model performed well externally; somewhat surprisingly in the external validation RADAR was markedly improved compared to the RT01

trial, as might not have been expected given RT01 is more similar to CHHiP than RADAR. The model did not appear to perform as well in the RADAR trial in the cohort of patients who received a 6-month short-course androgen suppression compared with the 18-month androgen suppression schedule. This was most likely due to improved outcomes in the 18-month schedule (and most similar to CHHiP's recurrence-free survival), compared to the 6-month schedule (see ***Appendix B Supplementary Figure B1***).

Chapter 6 sought to extend the joint model to account for a potential competing event of death unrelated to prostate cancer, as opposed to censoring those patients at their time of death, which may have an impact on the estimation of the cumulative incidence of recurrence. Dynamic predictions are demonstrated with the cumulative incidences of both outcomes (shown in ***Appendix C Supplementary Figure C1***). Predictive performance of the dynamic predictions using the competing risk joint model was conducted with two frameworks: the model-based weights, and the inverse probability of censoring weighting approaches, to account for censoring to derive valid estimates of predictive accuracy. The advantage of the IPCW method is that it is inherently model-free and does not assume correct specification of the model, unlike in the model-based method. However as discussed in the previous chapter, it is known that the model-based estimation method, given the joint model is well calibrated, correctly accounts for the fact that censoring can depend on the observed PSA concentrations, unlike IPCW [55,102].

7.2. Discussion and future work

7.2.1. Methodology used

The main focus in this thesis has been the use of the shared-parameter joint model [102,210,211], popularised by the work of the AIDS pandemic in directly modelling concentration of CD4 cells [212,213]. Although it is a natural way to model such data, it is a very computationally intensive process as it requires integrating over the random effects and all possible biomarker trajectories it can take from the landmark to horizon times. To model the nonlinearity of PSA, without imposing specific parametric assumptions, I opted for flexible natural cubic splines. In general, it should be robust to departures and misspecification as most patients exhibit a similar trajectory, given the treatment and starting follow-up similarities. There are some assumptions in correctly specifying the joint distribution between

PSA and recurrence and the distributional assumptions of PSA. A correctly specified joint distribution will lead to consistent predictions, $h_i\big(u|y_i(t)\big) = \mathbb{E}\big\{h_i\big(0|y_i(u)\big)|y_i(t)\big\}, u > t$, i.e., the expectation is the sample path of PSA from present landmark time t to horizon time u; implying the prediction made at u should be achieved at landmark time t by integreatng over the probability density function of PSA within the interval $[t, u]$ [55,214]. Joint modelling can also produce predictions at any landmark timepoint.

A Bayesian framework for estimation was implemented, which has several advantages compared to the frequentist framework. The Bayes paradigm is based on updating the posterior as more information becomes available, which is precisely what this thesis intended to achieve, i.e., dynamic predictions for an updated consideration of a patient's prognosis as added information (such as being alive to attend clinic, with a new PSA reading) becomes available. It has greater computational efficiency, easier interpretation of hypothesis testing, and the flexibility to amend and impose different informative priors to certain parameters (e.g. informative slab-or-spike priors) given information elicited from clinicians that may not necessarily be captured through the available covariates (e.g. information from scans, positive core biopsy proportions). The differences in implementing priors to the overall posterior has been explored by Fornacon-Wood [157].

In the literature review (**Chapter 3**), comparisons were made with other types of modelling (e.g. landmarking) and more recently landmarking 2.0 proposed by Putter & van Houwelingen to bridge the gap between the two techniques [60]. In this updated landmarking approach, the conditional survival function of the joint model is approximated by taking the expectation inside the integral and linear predictor. Landmarking violates the consistency condition above, as it can only make predictions at a specific timepoint (not at any time like joint modelling can). However, it does not need specification to the biomarker process and it is satisfactory for deriving dynamic predictions whilst not being as computationally intensive [59,62].

An alternative type of modelling that can bridge the gap between these two frameworks is the use of copulas. Gaussian copulas are typically used to model the joint distribution between the biomarker and time-to-event. They are advantageous as goodness-of-fit can be assessed, with quick and straightforward estimation. Copulas specify the marginal distribution of PSA

at each time without having to explicitly specify the longitudinal process [215]. Copulas have been applied to a prostate cancer study, but longitudinal clinical failure is used to predict death instead of longitudinal PSA [132].

At the start of this PhD there was much work on the formatting to the counting-process of the CHHiP dataset, the update of data snapshots, and centrally reviewed deaths that impacted on the event endpoints (e.g. whether some observed deaths were truly prostate cancer related, or otherwise). Similar formatting to the counting-process was substantially replicated for the RADAR and RT01 trials too, whereby the author replicated their respective primary analyses and substantive worked performed on formatting the data, in order that it could be used with the developed CHHiP CDPJM.

In this thesis, I also explored the JLCM, introduced in chapters **2** & **3**, as an alternative to the shared-parameter joint model. There are several advantages to modelling under this framework, like assuming further heterogeneous patient population, with each population having their own specific trajectory of the PSA biomarker; not assuming a specific association structure; or it being less computationally burdensome to estimate, given that to compute the log-likelihood it is only needed to integrate over the latent classes (rather than over the random-effect distribution in the shared-parameter framework) [66,84]. I explored and applied this framework to CHHiP early on in my PhD studies, though I found that there were no discernible classes that distinguished between risk groups beyond that of a non-recurrence/censored and recurrence trajectory. There was also the exploration of the frequentist joint modelling framework, using the **R** package *JM* [104]. Much of the provisional work was done within the predecessor package, *JMbayes* [71], where this author made open-source contributions. This author and supervisors plan to submit further publications featuring the work of **Chapters 0** & **0.** Code has been developed for this thesis, however currently embargoed whilst discussions were ongoing as to the appropriate dissemination of the potential of the model to be published as proprietary software. There are plans in future to publish the code to a public repository such as GitHub.

As discussed in **Chapter 3.7**, joint modelling has been dubbed an artificial intelligence (AI) method [146], though some commentators may dispute this as, fundamentally, the submodels are based on statistical likelihood. It stands to reason that AI and machine learning (ML)

approaches can be undertaken for dynamic predictions using longitudinal PSA biomarker data. These include deep learning, recurrent or convolutional neural networks (R/CNNs), which can handle longitudinal data and include survival information as an outcome [216]; a survival analysis model that is integrated with a machine learning model such as a random survival forest (RSF) [217].

Once such recent method in the literature is the use of *Dynamic-DeepHit*, a deep neural network [218,219]. It is a flexible ML method that can handle competing risks, that 'learns', using the longitudinal measurements, a data-driven time-to-event distribution. This approach removes the need to explicitly model the functional forms, nor make any parametric assumptions on either of the underlying outcomes and learns the complex relationship between the PSA and recurrence. This model has been used for prostate cancer patients on active surveillance to predict the risk of an upgrade to ≥ CPG3 using PSA, MRI and biopsy both at baseline and updated over follow-up; they also created and demonstrated the model in a practical application (Lee et al., 2022).

A pre-treatment localised prostate cancer study by Dai et al (2022) focuses on predicting a composite endpoint of PSA > 50ng/mL, metastasis or prostate cancer-related mortality. In particular, they use deep learning models, a recurrent deep survival machine (RDSM) and compare with RSF and a gradient boosting machine (GBM) [221]. For their composite endpoint, PSA is far beyond a typical localised biochemical failure definition, and they justify this by evaluating prognosis over a shorter timescale and before any treatment; as in clinical practice, one would rarely wait for a PSA to get that high before commencing treatment. They also include time-dependent age at each test and duration between each test from diagnosis, which could be redundant as this is inherently considered at baseline. Another study uses the XGBoost AI algorithm to predict 10-year prostate cancer mortality using 30 baseline prognostic factors [222].

There are other studies that use these AI/ML algorithms to predict prostate cancer recurrence. One such study uses CNNs on microscopic images of histopathological tissue to predict biochemical failure in a case-control study from radical prostatectomy patients; though not necessarily using PSA as a primary predictor [223]. In Toth et al., RSFs are used to detect genome-wide DNA methylation changes to identify patterns of expression corresponding to

prostate cancer progression between patients with good or poor prognosis, however they only use baseline and not longitudinal information to make those predictions [224].

ML and in particular, deep learning models lend themselves well to large datasets for them to be of use, but, though there are many ML/AI algorithms in development, which are growing in popularity, there is rarely any justification of their sample size [225]. Indeed, one would need at least a ten-fold increase in CHHiP's sample size for it to be considered worthwhile to use these deep learning approaches, particularly when the effective sample size is the number of events (rather than individual patients), which is comparatively low and gives rise to overfitting [226]. However in other medical domains, there are many papers comparing regression-based approaches to AI/ML, showing there is no superior performance of the latter over the former [227–230].

7.2.2. Specification of the model to predict prostate cancer recurrence

A snapshot of the CHHiP data taken on October 2019 (median 8½-year of follow-up) was used for the development of the clinical dynamic prediction model. In preparation for the updated 10-year analysis of CHHiP, a more updated snapshot is now available with additional data. This can be used to perform validation using data from those same patients who continued in follow-up from the October 2019 snapshot. Though these patients have been observed before and their random-effects known, this would be a pseudo-internal-external validation, which has been done similarly to Taylor and Yu [86,107].

Despite there being little (absolute) difference between the correctly specified cumulative incidence in the presence of competing risks and $1 - \mathrm{KM}$ estimator (censoring the competing event), due to the long-term follow-up of CHHiP, it will be expected that patients' overall survival will drop given the age of patients at recruitment (mean of 69 years old). This was evident in the 10-year updated CHHiP analysis recently presented (GU ASCO, Feb 2023), which showed that beyond 10 years the overall survival was ~80%, with only 15% of all-known causes of deaths related to prostate cancer; overall survival dropped to ~60-65% at 14 years after randomisation [34]. A competing risk analysis could allow for more accurate predictions if the goal were to predict at an extended time horizon of say ten years and beyond.

Competing risk joint models can provide further clinical utility when considering observational data, distinct from data curated from randomised clinical trials. Observational datasets often reflect real-world complexities that are not necessarily captured in clinical trials, e.g. patient diversity. Clinical trials have strict inclusion and exclusion criteria, leading to a more homogenous patient group. By contrast, observational studies capture a broader, more varied patient population, making the presence of competing risks more prevalent. Additionally, observational studies often have less frequent and structured follow-up compared to the stringent protocols of clinical trials. For instance, the CHHiP trial has a cut-off of ten years where these follow-up visits were at frequent prespecified intervals. However, many competing risk events may occur after such trial follow-up has concluded. This is particularly relevant in diseases affecting older populations, like prostate cancer. In these groups, observational data may have more competing risk events, since recurrence events might not be as rigorously recorded as they would be in a clinical trial setting.

The focus of the entire thesis has been using longitudinally collected PSA to predict clinical endpoints of recurrence only, there has been no consideration for clinician- or patient reported outcomes. These outcomes are recorded to report normal tissue effects to capture toxicity of (hypofractionated) radiotherapy. However, toxicities and side effects are not known to predict localised recurrence, and thus not accounted for in the model. There has been some recent work in using early reported toxicities to predict the likelihood of later side effects [231]. Joint models could also be used to extend current work to investigate the dosimetric determinants of radiotherapy toxicity, i.e., to predict the likelihood of experiencing late-onset adverse events.

I have only considered PSA over time as a potential predictor for recurrence, given its availability in CHHiP pre-planned data collection. However, it is well established that androgens (e.g. testosterone) are a key driver of prostate cancer growth and that PSA expression is regulated by androgen receptor activity [232–234]. Lower testosterone levels after radical prostatectomy treatment have been shown to be associated with unfavourable outcomes and increased risk of biochemical failure [235,236] and for radiotherapy [237], but in another study testosterone change was not predictive [238]. In the RADAR external cohort, longitudinal testosterone was also collected. The use of testosterone in the presence of PSA

could serve as a useful indicator of recurrence, or false positives of biochemical failure (e.g. simultaneous PSA and testosterone recovery could indicate flares or bounces). There are scenarios where testosterone could be helpful to ascertain the type of failure (or indeed cure). When PSA levels remain low, and testosterone has recovered back to normal pre-treatment levels indicates healthy post-treatment recovery. Where PSA and testosterone continually increase together post-treatment (where PSA does not plateau) could represent either clinical failure or a PSA 'bounce' [239]. Clinical failure with rising PSA may be controlled with intermittent salvage androgen suppression [240]. In this scenario, one can imagine PSA and testosterone drop after intermittent androgen suppression.

When salvage treatment ceases to be effective, PSA would be expected to increase whilst testosterone remains very low, indicating that the disease has become castrate resistant prostate cancer (CRPC). Accumulating evidence has shown that prostate cancers develop adaptive mechanisms for maintaining androgen receptor signalling to allow for survival and further evolution [241]. The mechanism of androgen receptor pathway modification and types of further therapeutic interventions (e.g. abiraterone or anti-androgens such as enzalutamide [242]) would be key to evaluating the testosterone and PSA relationship. In the case of testosterone synthesis inhibition, LHRH agonists and abiraterone, both testosterone and PSA would reduce proportionally, in the case of androgen receptor antagonists, such as the antiandrogens bicalutamide or enzalutamide, testosterone and PSA would not reduce proportionally, and therefore would need to be accounted for. Though the metastatic pathway is beyond the scope of this thesis, it is possible to model it with joint models in with a multivariate-multistate process (reviewed in **Chapter 3.6.3)** and incorporate testosterone with the reported outcomes already mentioned, in future work; this may have some additional benefit to reflect symptoms of recurrent disease [110,131].

Although CHHiP provided the evidence to implement hypofractionated EBRT regimes in clinical practice, there is a trend for even shorter treatments, achieved using stereotactic body radiotherapy (SBRT). Within the PACE study (NCT01584258), there are three independent trials currently being conducted by the ICR-CTSU. PACE-A evaluates SBRT to radical prostatectomy, PACE-B and PACE-C, which will combinedly recruit over 2,000 patients to compare (hypofractionated) conventional radiotherapy to SBRT (36.25Gy in just 5 fractions).

The PACE-B trial is for lower risk patients who have not received hormone therapy; PACE-C is for higher risk patients who are receiving 6 months of neoadjuvant hormonal therapy. When these trials and data have sufficiently matured in their follow-up, future work can include applying the developed CDPJM to ascertain whether it is generalisable to patients who have undergone SBRT and those who have not received short-course hormones. Like in the external validation performed in **Chapter 5**, it may be the case that baseline hazard adjustment will need to be made, particularly for PACE-B who have a lower NCCN risk profile compared to CHHiP patients.

The baseline covariates adjusted for in the model used known factors to be predictive of recurrence. Clinical T-stage was used instead of MRI defined stage due to better data completeness. Other known prognostic factors could not be implemented due to missing data, such as proportion of positive core biopsies [243,244], missing in a third of CHHiP patients. In **Chapter 4**, it was noted that 145 (5%) of CHHiP patients were not included due to missing covariates, a complete-case analysis was undertaken. There are known biases with complete-case analysis and that multiple imputation is recommended [245]. However, 90 patients had no hormone therapy and 5 received maximal androgen blockade, and therefore not eligible for imputation. Imputation could have been performed on the remaining 26 patients with missing baseline covariates. A further 24 patients had no longitudinal PSA. Imputation for this setting is more complicated but can be performed using a flexible fully conditional specification from the *JointAI* **R** package [246]. Given so few patients eligible for imputation were excluded, they are unlikely to have influenced any of the modelling procedure.

In a recent translational sub-study of CHHiP, immunohistochemistry (IHC) biomarkers were assessed on the diagnostic tissue sample for their predictive ability to radiotherapy fractionation response and prognosis [182,247,248]. Ki67 is a marker of cellular proliferation; it was found to be a strong independent predictor for recurrence [182]. Further work has been done to ascertain whether other IHC markers (e.g. HIF1α, Bcl-2, Ki67, Geminin, p16, p53, p-chk1 and PTEN), in the presence of established prognostic factors, are independently predictive of recurrence. Geminin (a proliferative marker), Ki67, and PTEN were prognostic [247]. Other markers are known to exist such as PCA3 [249], though it is unclear if it can be

used to predict recurrence after radical treatment. Another prostate cancer marker, TMPRSS2-ERG, was not found to be predictive of biochemical failure in a meta-analysis [250].

PSMA (prostate-specific membrane antigen) is another well founded biomarker and target, discovered in the 1980s [251]. It is an established target for molecular imaging techniques, such as PET scans, which allow for non-invasive visualisation of prostate cancer and assessment of its progression. It is possible to detect recurrence using prostatectomy specimens IHC staining for PSMA, particularly at low PSA levels after radical prostatectomy. Ross et al. found that PSMA overexpression was independently predictive of biochemical recurrence [252]. PSMA has been used for earlier detection of recurrence with patients being recharacterised with nodal and metastatic spread that would have historically been considered localised [253]. PSMA (68Ga-PSMA-11-PET) can also be used to confirm possible false-positives of biochemical recurrence, [254,255].

There exist consortia of radiogenomic studies to determine whether the inclusion of patient genomic data such as single nucleotide polymorphisms (SNPs) can be used to aid the prognosis of prostate cancer. Such data could in theory be extracted from the UK Biobank, PRACTICAL and/or RAPPER translational studies [256–258]. Studies have sought to quantify the profiling of SNPs on predicting biochemical failure post-treatment. One such study by Morote et al. showed a significant improvement in predicting biochemical failure after radical prostatectomy; the three genotypes associated with recurrence were KLK2 (rs198977), SULT1A1 (rs9282861) and TLR4 (rs1536889) SNPs [259]. In another study in radical radiotherapy, two SNPs were associated with biochemical failure: ERCC2 (rs1799793) & EXO1 (rs4149963); MSH6 (rs3136228) was associated with poorer overall survival [260].

Despite the prognostic evidence available for these additional biomarkers, they are not all yet routinely collected. This thesis provided strong predictive evidence of recurrence (in particular, biochemical failure) using only established baseline clinical variables, treatment, and repeatedly collected PSAs, without the need for these additional markers, which is advantageous as many patients may not have had these captured. There have been previous studies using joint modelling of PSAs and previous biopsy results to predict a bespoke cumulative risk of Gleason score upgrade to ≥ 7, to guide and optimise when the next pre-treatment biopsy should take place, balancing between patient burden and progression delay

and to reduce potentially unnecessary biopsies and corresponding complications [194]. It is possible to use the joint models developed in this thesis in a similar manner to guide personalised PSA scheduling when biochemical failure has yet to happen; or it could also be used as a Bayesian decision tool to trigger PSMA PET scans. For example, directing additional scans if PSAs surpass an unacceptable risk threshold, or guiding salvage therapies as appropriate, occurring within a clinically relevant prediction window of interest. Consequently, this can direct the frequency of follow-ups, further appointments/additional PSA readings taken. Conversely, if the patient is deemed to have good prognosis, then recommending fewer clinical exposures (e.g. reduced PSA readings taken) reduces patient burden. Further extensions can include making use of multi-state models, or intermediate events featuring progression, for example, (likely) biochemical failure → PSA-driven imaging i.e., PSMA PET → local failure → salvage therapies recommended to prevent any further clinical progressions.

It may be the case that these IHC & SNP markers become used in standard practice as costs reduce and become more widely available, so they can be incorporated to update these predictive models. As I demonstrated, the real power of these tools is incorporating the use of longitudinal information; other repeatedly collected information, such as testosterone, PET/MRI, and biopsies can be incorporated into the model to improve prediction of recurrence and therefore instigate salvage intervention earlier.

7.2.3. Implementation of prediction models in clinical practice

Regardless of which framework is used (from standard baseline CPMs, landmarking, joint modelling, or machine learning), there is hope that these developed and validated models can be translated into an interactive web application, to facilitate their use and in order to maximise clinical utility. There have been some attempts to do this in previous work: Ferrer and colleagues have made some of their code available https://github.com/LoicFerrer[4]. The 'Prostate Cancer Calculator' can be found at https://psacalc.sph.umich.edu/[5], which is aimed at those patients who have received monotherapy EBRT *without* short-course hormones. These tools are considered research tools and not intended to direct treatment but intended to help

[4] Accessed in March 2023
[5] Accessed in March 2023

the patient and physician decide on what action to take. In order for these tools to be implemented into clinical practice, appropriate evidence of their performance and qualification for clinical use must be obtained.

In recent years, there has been much progress in the regulatory framework to enable and disseminate software and algorithms for medicinal use. Regulatory considerations and systems are required to be put in place so that safe and kind treatments can be delivered effectively; therefore, the same should apply to software/algorithm-based tools used for clinical decisions. In order to use software for clinical use, a CE or UKCA (UK Conformity Assessed) mark is required for market launch, i.e., commercialised use. The UK now has its own regulatory infrastructure post-Brexit (UKCA), where governance of medicinal devices is given by the MHRA (Medicines and Healthcare products Regulatory Agency). Regardless of commercialisation, medical device software to put to use on patients requires an effective Quality Management System (QMS) and the definition of intended use and appropriate risk class.

There are three risk classes: IIa, IIb, and III [261]; this is summarised in **Table 7-1**. The aim of an app implementing the prediction model developed in this thesis would be to primarily *drive clinical management*. Diseases pertaining to cancer are classed as *serious* (other diseases, such as strokes, that require a response within an acute timeframe are considered *critical)*. Therefore, the proposed app class risk rating is IIa.

Table 7-1 medical device classification guidance, taken from MDCG 2019-11 Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR, 2019 [261].

		What is the exact application use?		
What is the risk of harming a patient?		**Treat/diagnose**	**Drive clinical management**	**Inform treatment**
	Critical	Class III	Class IIb	Class IIa
	Serious	Class IIb	Class IIa	Class IIa
	Other	Class IIa	Class IIa	Class IIa

With the class risk rating defined, the *intended use statement* can be written, referring to the legislation, and communicating to the regulators the rationale behind my assumption of risk class IIa.

Another key component in receiving regulatory approval for a medical device is rigorous use of a QMS. This is set out by the international organisation for standardisation (ISO) 13485, which is a platform to store documentation and processes [262]. Other such standards are needed such as security management (ISO 27001), and risk management (ISO 14971). Of course, all treatments and devices have a risk / harm-to-benefit profile, software being no different. There could be some risk associated with the medical device, however it is vital that there is, by and large, an overall benefit demonstrated, and that quality assurance is maintained for audits.

Another requirement is to establish clinical validation; this is different from the already mentioned validation of the predictive joint model itself. Clinical investigation is required for all class III devices and class II devices that have a new intended use, for which there is no equivalent tool on the market, or that are developed on innovative technology such as AI/ML. This involves producing a clinical evaluation report (CER), which documents evidence and assessment of the safety, performance, and clinical effectiveness of the medical device. It is necessary for regulatory agencies to demonstrate compliance with essential requirements and to support the marketing authorisation of a medical device. A CER typically includes a review of relevant scientific literature, conducting clinical studies to demonstrate the tool's clinical performance and its conformity to relevant standards. The purpose of the CER is to provide assurance to regulators that the device is safe and effective for its intended use, and that it meets regulatory requirements pertaining to quality, safety, and performance. One could perform an observational study to assess the tool, by allocating (possibly randomising) patients after treatment to follow either a fixed schedule, or a personalised one provided by the CDPJM, similarly suggested by Tomer and colleagues [134].

There is also a requirement to place emphasis on updating documentation and the product in post-filing surveillance and follow-up (the equivalent of phase IV clinical trial monitoring), as the tool may be subject to biases not known during model development. There are many components in post-marketing surveillance, in particular under ISO 14971, where there is a requirement to write a periodic safety update report biennially for class IIa and annually for classes IIb & III, with post-market clinical follow-up.

Finally, once the tool has had regulatory approval, the device can be subject to changes and amendments as required, i.e., it is not frozen and brand-new approvals are not required to update the tool, provided that there are no substantial changes. A change (or version) control provides a framework to address this and is an important component of quality and regulatory management, which needs to be defined before going to market. It is a process in which changes to the device/tool are systematically evaluated, approved, or rejected, and managed. This process is implemented to ensure that changes to the device do not negatively impact its performance, quality, or compliance with regulations. The process itself may stem from a change request, in the form of an error or bug. This leads to initiation of the change control *process*. Human notification is key, i.e., involvement from a project manager and the solution delivery team to decide on the change to be made, documentation of said change (e.g. change control record), and how one defines the metric to quantify and show how the change was correct. There are several ranging tiers depending on the type, or significance of the change. For example, a tier zero change is considerable and needs to be discussed with the regulatory bodies and it is possible the tool may need to be recertified. A tier three / four change constitutes a very minor fix for a bug not normally noticeable by users.

Here I merely touch upon what is involved and required to launch a predictive tool to market, or at least put it to wider use. If it is to have commercialisation, i.e., placed onto the market, then CE/UKCA marking is required, which needs approval from the relevant regulatory bodies. However, if the intention is to put into use without the commercialisation of intellectual property, then regulatory approvals are not necessarily required. However, it is important to ensure that these models are validated and that their use is guided by ethical principles.

Recently there has been further focus by the MHRA that Large Language Models (LLMs) such as ChatGPT when used for medicinal purposes will qualify as software as a medical device and hence need to be regulated under the Medical Device Regulation (2002), as well as its intended use to be defined [263,264].

The intended use and who is to have access are important factors that require clarification. The intention of this device would be for clinicians to use in the first instance. Conveying and communicating risk to the patient is a hugely pertinent topic that needs due diligence and

care, in the presence of a clinical professional. Misconceptions of risk and its communication – to the wider public diagnosed with cancer – was eloquently explored by Prof Hannah Fry (who is a British professor of mathematics at UCL) in a recent BBC Horizon documentary *Making Sense of Cancer* (first broadcast in June 2022) after her own diagnosis of cervical cancer, aged 36. This highlighted the impact of misinterpretation of risk, cancer overdiagnosis and overtreatment [265].

Clinical prediction models have the potential to revolutionise the way healthcare is delivered by providing personalised, data-driven recommendations for diagnosis, treatment, and disease management. In the future, there are several trends that are likely to shape the future development and use of clinical prediction models:

- The increased use of electronic health records (EHRs) and other data sources, such as genomics, imaging, self-reported measures, provide a wealth of information that can be used to train and validate clinical prediction models. This data could be used to create better predictive models, improve the generalisability of the models (e.g. diversity & inclusion), and collect necessary data more easily via primary care (e.g. GPs), without the need to visit out-patient clinics.

- Development of artificial intelligence and machine learning models: as discussed earlier, ML models, such as deep learning and neural networks, can process large amounts of data and can be trained to identify complex patterns in the data, particularly if large genomic data is available. These models have the potential to improve the accuracy and generalisability of clinical prediction models. They should be developed in line with reporting guidance (e.g. TRIPOD-AI / PROBAST-AI [154,266]).

• The use of explainable AI (XAI): as the use of ML models increases, there is a growing need to understand how these models make predictions, as AI/ML tools have traditionally been viewed as 'black boxes' that provide output only. XAI techniques, such as feature importance and model interpretability, can be used to provide insight into the underlying mechanisms of the models and to identify potential biases.

• Clinical prediction models can be integrated into clinical decision support systems to provide real-time recommendations to healthcare providers. This can help to improve the efficiency and effectiveness of care delivery, in addition to collaboration with other professionals such as researchers, data scientists and engineers, policy makers, and patients, which is essential for building accurate and dependable prediction models that can be integrated into the clinical workflow.

7.3 Conclusions

To conclude, the work I presented in this thesis has focused on the development of Bayesian predictive tools to characterise the prognosis of localised prostate cancer patients, after short-course hormones and (hypofractionated) radiotherapy. PSA trajectories are predictive of impending (biochemical) recurrence; I suggested broader clinical PSA cut-offs that are indicative of prolonged event-free survival (or cure), PSA $\lessapprox$ 0.23ng/mL three years follow-up indicates very good prognosis.

These tools are not known to have been developed under this treatment modality, which is now the standard of care. There will be many more patients who have (or will) undergo this treatment pathway in clinic external from any trial, I hope this work will provide clinical utility of their prognosis for many years to come. The model is generalisable in other healthcare settings, in locally advanced patient populations, those who have received long-course hormone therapy, as demonstrated through the external validation, and can be extended to incorporate the presence of competing risks.

I recommend future research that explores the potential of using explainable deep learning methods when combining IHC and genomic data, along with longitudinally collected information, so long as the effective sample size permits its use to prevent overfitting. Furthermore, I introduced the regulatory framework required to commercialise software as a medical device to put into clinical use.

The future of clinical prediction models appears encouraging as they have the potential to enhance patient outcomes by facilitating earlier intervention, resulting in reduced costs through proactive prevention, and improved quality of life for patients. This can contribute significantly to advancing the frontiers of science and improving outcomes for future patients.

8 Bibliography

[1] Cancer today, (2020). http://gco.iarc.fr/today/home (accessed May 4, 2022).

[2] Cancer registration statistics, England: final release, 2018, GOV.UK. (2020). https://www.gov.uk/government/statistics/cancer-registration-statistics-england-2018-final-release/cancer-registration-statistics-england-final-release-2018 (accessed May 4, 2022).

[3] C.E. Lovegrove, O. Musbahi, N. Ranasinha, A. Omer, F. Lopez, A. Campbell, R.J. Bryant, T. Leslie, R. Bell, S. Brewster, F.C. Hamdy, B. Wright, A.D. Lamb, Implications of celebrity endorsement of prostate cancer awareness in a tertiary referral unit - the 'Fry-Turnbull' effect, BJU Int. 125 (2020) 484–486. https://doi.org/10.1111/bju.14992.

[4] Prostate cancer statistics, Cancer Research UK. (2015). https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer (accessed September 6, 2020).

[5] H. Lilja, Biology of prostate-specific antigen, Urology. 62 (2003) 27–33. https://doi.org/10.1016/S0090-4295(03)00775-1.

[6] A.R. Rao, H.G. Motiwala, O.M.A. Karim, The discovery of prostate-specific antigen, BJU International. 101 (2008) 5–10. https://doi.org/10.1111/j.1464-410X.2007.07138.x.

[7] J. Xiang, H. Yan, J. Li, X. Wang, H. Chen, X. Zheng, Transperineal versus transrectal prostate biopsy in the diagnosis of prostate cancer: a systematic review and meta-analysis, World Journal of Surgical Oncology. 17 (2019) 31. https://doi.org/10.1186/s12957-019-1573-0.

[8] Early Diagnosis, (2022). https://crukcancerintelligence.shinyapps.io/EarlyDiagnosis/_w_f7c9dcf1/_w_eae61c40/#shiny-tab-incidence1 (accessed February 15, 2023).

[9] NPCA Annual Report 2022, National Prostate Cancer Audit. (2023). https://www.npca.org.uk/reports/npca-annual-report-2022/ (accessed February 1, 2023).

[10] Survival by stage, (2018). http://www.ncin.org.uk/publications/survival_by_stage (accessed November 23, 2020).

[11] S. McPhail, S. Johnson, D. Greenberg, M. Peake, B. Rous, Stage at diagnosis and early mortality from cancer in England, British Journal of Cancer. 112 (2015) 108–115. https://doi.org/10.1038/bjc.2015.49.

[12] C.N. Catton, H. Lukka, C.-S. Gu, J.M. Martin, S. Supiot, P.W.M. Chung, G.S. Bauman, J.-P. Bahary, S. Ahmed, P. Cheung, K.H. Tai, J.S. Wu, M.B. Parliament, T. Tsakiridis, T.B. Corbett, C. Tang, I.S. Dayes, P. Warde, T.K. Craig, J.A. Julian, M.N. Levine, Randomized Trial of a Hypofractionated Radiation Regimen for the Treatment of Localized Prostate Cancer, JCO. 35 (2017) 1884–1890. https://doi.org/10.1200/JCO.2016.71.7397.

[13] D. Dearnaley, I. Syndikus, H. Mossop, V. Khoo, A. Birtle, D. Bloomfield, J. Graham, P. Kirkbride, J. Logue, Z. Malik, J. Money-Kyrle, J.M. O'Sullivan, M. Panades, C. Parker, H. Patterson, C. Scrase, J. Staffurth, A. Stockdale, J. Tremlett, M. Bidmead, H. Mayles, O. Naismith, C. South, A. Gao, C. Cruickshank, S. Hassan, J. Pugh, C. Griffin, E. Hall, Conventional versus hypofractionated high-dose intensity-modulated radiotherapy for prostate cancer: 5-year outcomes of the randomised, non-inferiority, phase 3

CHHiP trial, The Lancet Oncology. 17 (2016) 1047–1060. https://doi.org/10.1016/S1470-2045(16)30102-4.

[14] W.R. Lee, J.J. Dignam, M.B. Amin, D.W. Bruner, D. Low, G.P. Swanson, A.B. Shah, D.P. D'Souza, J.M. Michalski, I.S. Dayes, S.A. Seaward, W.A. Hall, P.L. Nguyen, T.M. Pisansky, S.L. Faria, Y. Chen, B.F. Koontz, R. Paulus, H.M. Sandler, Randomized Phase III Noninferiority Study Comparing Two Radiotherapy Fractionation Schedules in Patients With Low-Risk Prostate Cancer, JCO. 34 (2016) 2325–2332. https://doi.org/10.1200/JCO.2016.67.0448.

[15] K.J. Ray, N.R. Sibson, A.E. Kiltie, Treatment of Breast and Prostate Cancer by Hypofractionated Radiotherapy: Potential Risks and Benefits, Clin Oncol (R Coll Radiol). 27 (2015) 420–426. https://doi.org/10.1016/j.clon.2015.02.008.

[16] M. Bolla, A. Neven, P. Maingon, C. Carrie, A. Boladeras, D. Andreopoulos, A. Engelen, S. Sundar, E.M. van der Steen-Banasik, J. Armstrong, K. Peignaux-Casasnovas, J. Boustani, F.G. Herrera, B.R. Pieters, A. Slot, A. Bahl, C.D. Scrase, D. Azria, J. Jansa, J.M. O'Sullivan, A.C.M. Van Den Bergh, L. Collette, EORTC Radiation Oncology Group, Short Androgen Suppression and Radiation Dose Escalation in Prostate Cancer: 12-Year Results of EORTC Trial 22991 in Patients With Localized Intermediate-Risk Disease, J Clin Oncol. 39 (2021) 3022–3033. https://doi.org/10.1200/JCO.21.00855.

[17] A. Tree, C. Griffin, I. Syndikus, A. Birtle, A. Choudhury, J. Graham, C. Ferguson, V. Khoo, Z. Malik, J. O'Sullivan, M. Panades, C. Parker, Y. Rimmer, C. Scrase, J. Staffurth, D. Dearnaley, E. Hall, Nonrandomized Comparison of Efficacy and Side Effects of Bicalutamide Compared With Luteinizing Hormone-Releasing Hormone (LHRH) Analogs in Combination With Radiation Therapy in the CHHiP Trial, International Journal of Radiation Oncology*Biology*Physics. (2022). https://doi.org/10.1016/j.ijrobp.2021.12.160.

[18] M. Roach, G. Hanks, H. Thames, P. Schellhammer, W.U. Shipley, G.H. Sokol, H. Sandler, Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: Recommendations of the RTOG-ASTRO Phoenix Consensus Conference, International Journal of Radiation Oncology*Biology*Physics. 65 (2006) 965–974. https://doi.org/10.1016/j.ijrobp.2006.04.029.

[19] M.B. Amin, S. Edge, F. Greene, D.R. Byrd, R.K. Brookland, M.K. Washington, J.E. Gershenwald, C.C. Compton, K.R. Hess, D.C. Sullivan, J.M. Jessup, J.D. Brierley, L.E. Gaspar, R.L. Schilsky, C.M. Balch, D.P. Winchester, E.A. Asare, M. Madera, D.M. Gress, L.R. Meyer, eds., AJCC Cancer Staging Manual, 8th ed., Springer International Publishing, 2017. https://www.springer.com/gp/book/9783319406176 (accessed November 23, 2020).

[20] J.L. Mohler, E.S. Antonarakis, A.J. Armstrong, A.V. D'Amico, B.J. Davis, T. Dorff, J.A. Eastham, C.A. Enke, T.A. Farrington, C.S. Higano, E.M. Horwitz, M. Hurwitz, J.E. Ippolito, C.J. Kane, M.R. Kuettel, J.M. Lang, J. McKenney, G. Netto, D.F. Penson, E.R. Plimack, J.M. Pow-Sang, T.J. Pugh, S. Richey, M. Roach, S. Rosenfeld, E. Schaeffer, A. Shabsigh, E.J. Small, D.E. Spratt, S. Srinivas, J. Tward, D.A. Shead, D.A. Freedman-Cass, Prostate Cancer, Version 2.2019, NCCN Clinical Practice Guidelines in Oncology, Journal of the National Comprehensive Cancer Network. 17 (2019) 479–505. https://doi.org/10.6004/jnccn.2019.0023.

[21] W. Xie, M.M. Regan, M. Buyse, S. Halabi, P.W. Kantoff, O. Sartor, H. Soule, N.W. Clarke, L. Collette, J.J. Dignam, K. Fizazi, W.R. Paruleker, H.M. Sandler, M.R. Sydes, B. Tombal, S.G. Williams, C.J. Sweeney, Metastasis-Free Survival Is a Strong Surrogate of Overall Survival in Localized Prostate Cancer, JCO. 35 (2017) 3097–3104. https://doi.org/10.1200/JCO.2017.73.9987.

[22] A. Widmark, A. Gunnlaugsson, L. Beckman, C. Thellenberg-Karlsson, M. Hoyer, M. Lagerlund, J. Kindblom, C. Ginman, B. Johansson, K. Björnlinger, M. Seke, M. Agrup, P. Fransson, B. Tavelin, D. Norman, B. Zackrisson, H. Anderson, E. Kjellén, L. Franzén, P. Nilsson, Ultra-hypofractionated versus conventionally fractionated radiotherapy for prostate cancer: 5-year outcomes of the HYPO-RT-PC randomised, non-inferiority, phase 3 trial, The Lancet. 394 (2019) 385–395. https://doi.org/10.1016/S0140-6736(19)31131-6.

[23] Tools and resources | Prostate cancer: diagnosis and management | Guidance | NICE, (2019). https://www.nice.org.uk/guidance/ng131/resources (accessed February 2, 2023).

[24] M.G. Parry, T.E. Cowling, A. Sujenthiran, J. Nossiter, B. Berry, P. Cathcart, A. Aggarwal, H. Payne, J. van der Meulen, N.W. Clarke, V.J. Gnanapragasam, Risk stratification for prostate cancer management: value of the Cambridge Prognostic Group classification for assessing treatment allocation, BMC Medicine. 18 (2020) 114. https://doi.org/10.1186/s12916-020-01588-9.

[25] W.E. Barlow, E. White, R. Ballard-Barbash, P.M. Vacek, L. Titus-Ernstoff, P.A. Carney, J.A. Tice, D.S.M. Buist, B.M. Geller, R. Rosenberg, B.C. Yankaskas, K. Kerlikowske, Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography, JNCI: Journal of the National Cancer Institute. 98 (2006) 1204–1214. https://doi.org/10.1093/jnci/djj331.

[26] M. Aladwani, A. Lophatananon, W. Ollier, K. Muir, Prediction models for prostate cancer to be used in the primary care setting: a systematic review, BMJ Open. 10 (2020) 034661. https://doi.org/10.1136/bmjopen-2019-034661.

[27] D.R. Thurtle, D.C. Greenberg, L.S. Lee, H.H. Huang, P.D. Pharoah, V.J. Gnanapragasam, Individual prognosis at diagnosis in nonmetastatic prostate cancer: Development and external validation of the PREDICT Prostate multivariable model, PLOS Medicine. 16 (2019) e1002758. https://doi.org/10.1371/journal.pmed.1002758.

[28] G.C. Wishart, E.M. Azzato, D.C. Greenberg, J. Rashbass, O. Kearins, G. Lawrence, C. Caldas, P.D. Pharoah, PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer, Breast Cancer Research. 12 (2010) R1. https://doi.org/10.1186/bcr2464.

[29] S.F. Shariat, P.I. Karakiewicz, C.G. Roehrborn, M.W. Kattan, An updated catalog of prostate cancer predictive tools, Cancer. 113 (2008) 3075–3099. https://doi.org/10.1002/cncr.23908.

[30] S.F. Shariat, P.I. Karakiewicz, V. Margulis, M.W. Kattan, Inventory of prostate cancer predictive tools, Current Opinion in Urology. 18 (2008) 279–296. https://doi.org/10.1097/MOU.0b013e3282f9b3e5.

[31] E. Lalonde, A.S. Ishkanian, J. Sykes, M. Fraser, H. Ross-Adams, N. Erho, M.J. Dunning, S. Halim, A.D. Lamb, N.C. Moon, G. Zafarana, A.Y. Warren, X. Meng, J. Thoms, M.R. Grzadkowski, A. Berlin, C.L. Have, V.R. Ramnarine, C.Q. Yao, C.A. Malloff, L.L. Lam, H. Xie, N.J. Harding, D.Y.F. Mak, K.C. Chu, L.C. Chong, D.H. Sendorek, C. P'ng, C.C. Collins, J.A. Squire, I. Jurisica, C. Cooper, R. Eeles, M. Pintilie,

A.D. Pra, E. Davicioni, W.L. Lam, M. Milosevic, D.E. Neal, T. van der Kwast, P.C. Boutros, R.G. Bristow, Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study, The Lancet Oncology. 15 (2014) 1521–1532. https://doi.org/10.1016/S1470-2045(14)71021-6.

[32] D. Rizopoulos, J.M.G. Taylor, J. Van Rosmalen, E.W. Steyerberg, J.J.M. Takkenberg, Personalized screening intervals for biomarkers using joint models for longitudinal and survival data, Biostatistics. 17 (2016) 149–164. https://doi.org/10.1093/biostatistics/kxv031.

[33] J.G. Ibrahim, H. Chu, L.M. Chen, Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data, J Clin Oncol. 28 (2010) 2796–2801. https://doi.org/10.1200/JCO.2009.25.0654.

[34] I. Syndikus, C. Griffin, L. Philipps, A. Tree, V. Khoo, A.J. Birtle, A. Choudhury, C. Ferguson, J.M. O'Sullivan, M. Panades, Y.L. Rimmer, C.D. Scrase, J. Staffurth, C. Cruickshank, S. Hassan, J. Pugh, D.P. Dearnaley, E. Hall, 10-Year efficacy and co-morbidity outcomes of a phase III randomised trial of conventional vs. hypofractionated high dose intensity modulated radiotherapy for prostate cancer (CHHiP; CRUK/06/016)., JCO. 41 (2023) 304–304. https://doi.org/10.1200/JCO.2023.41.6_suppl.304.

[35] J. Nossiter, A. Sujenthiran, T.E. Cowling, M.G. Parry, S.C. Charman, P. Cathcart, N.W. Clarke, H. Payne, J. van der Meulen, A. Aggarwal, Patient-Reported Functional Outcomes After Hypofractionated or Conventionally Fractionated Radiation for Prostate Cancer: A National Cohort Study in England, JCO. 38 (2020) 744–752. https://doi.org/10.1200/JCO.19.01538.

[36] A. Sujenthiran, M. Parry, J. Nossiter, B. Berry, P.J. Cathcart, N.W. Clarke, H. Payne, J. van der Meulen, A. Aggarwal, Comparison of Treatment-Related Toxicity With Hypofractionated or Conventionally Fractionated Radiation Therapy for Prostate Cancer: A National Population-Based Study, Clinical Oncology. 32 (2020) 501–508. https://doi.org/10.1016/j.clon.2020.02.004.

[37] E.W. Steyerberg, K.G.M. Moons, D.A. van der Windt, J.A. Hayden, P. Perel, S. Schroter, R.D. Riley, H. Hemingway, D.G. Altman, Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research, PLoS Med. 10 (2013). https://doi.org/10.1371/journal.pmed.1001381.

[38] A.J. Vickers, A.M. Cronin, Everything You Always Wanted to Know About Evaluating Prediction Models (But Were Too Afraid to Ask), Urology. 76 (2010) 1298–1301. https://doi.org/10.1016/j.urology.2010.06.019.

[39] J.C. Wyatt, D.G. Altman, Commentary: Prognostic models: clinically useful or quickly forgotten?, BMJ. 311 (1995) 1539–1541. https://doi.org/10.1136/bmj.311.7019.1539.

[40] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement, BMC Medicine. 13 (2015) 1. https://doi.org/10.1186/s12916-014-0241-z.

[41] H. Parr, E. Hall, N. Porta, Joint models for dynamic prediction in localised prostate cancer: a literature review, BMC Medical Research Methodology. 22 (2022) 245. https://doi.org/10.1186/s12874-022-01709-3.

[42] H. Parr, N. Porta, A.C. Tree, D. Dearnaley, E. Hall, A Personalised Clinical Dynamic Prediction Model to Characterise Prognosis for Patients with Localised Prostate

Cancer: analysis of the CHHiP Phase III Trial, International Journal of Radiation Oncology, Biology, Physics. 0 (2023). https://doi.org/10.1016/j.ijrobp.2023.02.022.

[43] S. Halabi, C. Li, S. Luo, Developing and Validating Risk Assessment Models of Clinical Outcomes in Modern Oncology, JCO Precision Oncology. (2019) 1–12. https://doi.org/10.1200/PO.19.00068.

[44] P. Royston, D.G. Altman, External validation of a Cox prognostic model: principles and methods, BMC Medical Research Methodology. 13 (2013) 33. https://doi.org/10.1186/1471-2288-13-33.

[45] E.W. Steyerberg, F.E. Harrell, Prediction models need appropriate internal, internal-external, and external validation, J Clin Epidemiol. 69 (2016) 245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005.

[46] J.W. Denham, D. Joseph, D.S. Lamb, N.A. Spry, G. Duchesne, J. Matthews, C. Atkinson, K.-H. Tai, D. Christie, L. Kenny, S. Turner, N.K. Gogna, T. Diamond, B. Delahunt, C. Oldmeadow, J. Attia, A. Steigler, Short-term androgen suppression and radiotherapy versus intermediate-term androgen suppression and radiotherapy, with or without zoledronic acid, in men with locally advanced prostate cancer (TROG 03.04 RADAR): 10-year results from a randomised, phase 3, factorial trial, The Lancet Oncology. 20 (2019) 267–281. https://doi.org/10.1016/S1470-2045(18)30757-5.

[47] D.P. Dearnaley, G. Jovic, I. Syndikus, V. Khoo, R.A. Cowan, J.D. Graham, E.G. Aird, D. Bottomley, R.A. Huddart, C.C. Jose, J.H.L. Matthews, J.L. Millar, C. Murphy, J.M. Russell, C.D. Scrase, M.K.B. Parmar, M.R. Sydes, Escalated-dose versus control-dose conformal radiotherapy for prostate cancer: long-term results from the MRC RT01 randomised controlled trial, The Lancet Oncology. 15 (2014) 464–473. https://doi.org/10.1016/S1470-2045(14)70040-3.

[48] D.P. Ankerst, J. Straubinger, K. Selig, L. Guerrios, A. De Hoedt, J. Hernandez, M.A. Liss, R.J. Leach, S.J. Freedland, M.W. Kattan, R. Nam, A. Haese, F. Montorsi, S.A. Boorjian, M.R. Cooperberg, C. Poyet, E. Vertosick, A.J. Vickers, A Contemporary Prostate Biopsy Risk Calculator Based on Multiple Heterogeneous Cohorts, Eur. Urol. 74 (2018) 197–203. https://doi.org/10.1016/j.eururo.2018.05.003.

[49] J.A. Brockman, S. Alanee, A.J. Vickers, P.T. Scardino, D.P. Wood, A.S. Kibel, D.W. Lin, F.J. Bianco, D.M. Rabah, E.A. Klein, J.P. Ciezki, T. Gao, M.W. Kattan, A.J. Stephenson, Nomogram Predicting Prostate Cancer-specific Mortality for Men with Biochemical Recurrence After Radical Prostatectomy, Eur. Urol. 67 (2015) 1160–1167. https://doi.org/10.1016/j.eururo.2014.09.019.

[50] C.R. Pound, A.W. Partin, M.A. Eisenberger, D.W. Chan, J.D. Pearson, P.C. Walsh, Natural history of progression after PSA elevation following radical prostatectomy, JAMA. 281 (1999) 1591–1597. https://doi.org/10.1001/jama.281.17.1591.

[51] Prostate Cancer Nomograms, Memorial Sloan Kettering Cancer Center. (2019). https://www.mskcc.org/nomograms/prostate (accessed September 16, 2019).

[52] S.F. Shariat, K. Mw, V. Aj, K. Pi, S. Pt, Critical review of prostate cancer predictive tools., Future Oncol. 5 (2009) 1555–1584. https://doi.org/10.2217/fon.09.121.

[53] P.K. Andersen, R.D. Gill, Cox's Regression Model for Counting Processes: A Large Sample Study, Ann. Statist. 10 (1982) 1100–1120. https://doi.org/10.1214/aos/1176345976.

[54] T.M. Therneau, P.M. Grambsch, Modeling Survival Data: Extending the Cox Model, Springer Science & Business Media, 2000.

[55] D. Rizopoulos, G. Molenberghs, E.M.E.H. Lesaffre, Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking, Biometrical Journal. 59 (2017) 1261–1276. https://doi.org/10.1002/bimj.201600238.

[56] G. Papageorgiou, K. Mauff, A. Tomer, D. Rizopoulos, An Overview of Joint Modeling of Time-to-Event and Longitudinal Outcomes, Annual Review of Statistics and Its Application. 6 (2019). https://doi.org/10.1146/annurev-statistics-030718-105048.

[57] J.D. Kalbfleisch, R.L. Prentice, The statistical analysis of failure time data, 2. ed, Wiley, Hoboken, NJ, 2002.

[58] J.R. Anderson, K.C. Cain, R.D. Gelber, Analysis of survival by tumor response., JCO. 1 (1983) 710–719. https://doi.org/10.1200/JCO.1983.1.11.710.

[59] H. van Houwelingen, H. Putter, Dynamic Prediction in Clinical Survival Analysis, CRC Press, 2011.

[60] H. Putter, H.C. van Houwelingen, Landmarking 2.0: Bridging the gap between joint models and landmarking, Statistics in Medicine. 41 (2022) 1901–1917. https://doi.org/10.1002/sim.9336.

[61] K. Suresh, J.M.G. Taylor, D.E. Spratt, S. Daignault, A. Tsodikov, Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model, Biom J. 59 (2017) 1277–1300. https://doi.org/10.1002/bimj.201600235.

[62] H. van Houwelingen, Dynamic Prediction by Landmarking in Event History Analysis, Scandinavian Journal of Statistics. 34 (2007) 70–85. https://doi.org/10.1111/j.1467-9469.2006.00529.x.

[63] C. Proust-Lima, J.M.G. Taylor, S.G. Williams, D.P. Ankerst, N. Liu, L.L. Kestin, K. Bae, H.M. Sandler, Determinants of Change in Prostate-Specific Antigen Over Time and Its Association With Recurrence After External Beam Radiation Therapy for Prostate Cancer in Five Large Cohorts, International Journal of Radiation Oncology*Biology*Physics. 72 (2008) 782–791. https://doi.org/10.1016/j.ijrobp.2008.01.056.

[64] M.J. Sweeting, S.G. Thompson, Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture, Biom J. 53 (2011) 750–763. https://doi.org/10.1002/bimj.201100052.

[65] N.M. Laird, J.H. Ware, Random-Effects Models for Longitudinal Data, Biometrics. 38 (1982) 963–974. https://doi.org/10.2307/2529876.

[66] C. Proust-Lima, M. Séne, J.M. Taylor, H. Jacqmin-Gadda, Joint latent class models for longitudinal and time-to-event data: A review, Statistical Methods in Medical Research. 23 (2014) 74. https://doi.org/10.1177/0962280212445839.

[67] D. Rizopoulos, Joint Models for Longitudinal and Time-to-Event Data: With Applications in R, Chapman and Hall/CRC, Boca Raton, UNITED KINGDOM, 2012. http://ebookcentral.proquest.com/lib/icruk/detail.action?docID=952042 (accessed November 7, 2018).

[68] A.L. Gould, M.E. Boye, M.J. Crowther, J.G. Ibrahim, G. Quartey, S. Micallef, F.Y. Bois, Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group, Stat Med. 34 (2015) 2181–2195. https://doi.org/10.1002/sim.6141.

[69] E.-R. Andrinopoulou, D. Rizopoulos, J.J.M. Takkenberg, E. Lesaffre, Joint modeling of two longitudinal outcomes and competing risk data, Statistics in Medicine. 33 (2014) 3167–3178. https://doi.org/10.1002/sim.6158.

[70] F. Hsieh, Y.-K. Tseng, J.-L. Wang, Joint modeling of survival and longitudinal data: likelihood approach revisited, Biometrics. 62 (2006) 1037–1043. https://doi.org/10.1111/j.1541-0420.2006.00570.x.
[71] D. Rizopoulos, The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC, Journal of Statistical Software. 72 (2016) 1–46. https://doi.org/10.18637/jss.v072.i07.
[72] J.G. Ibrahim, M.-H. Chen, D. Sinha, Bayesian Survival Analysis, Springer Science & Business Media, 2001.
[73] M. Yu, N.J. Law, J.M.G. Taylor, H.M. Sandler, Joint Longitudinal-survival-cure Models and Their Application to Prostate Cancer, Statistica Sinica. (2004) 28.
[74] A. Gelman, D.B. Rubin, Inference from Iterative Simulation Using Multiple Sequences, Statistical Science. 7 (1992) 457–472. https://doi.org/10.1214/ss/1177011136.
[75] M. Stephens, Dealing with Label Switching in Mixture Models, Journal of the Royal Statistical Society. Series B (Statistical Methodology). 62 (2000) 795–809. https://www.jstor.org/stable/2680622 (accessed November 23, 2022).
[76] Y. Li, J. Lord-Bessen, M. Shiyko, R. Loeb, Bayesian Latent Class Analysis Tutorial, Multivariate Behav Res. 53 (2018) 430–451. https://doi.org/10.1080/00273171.2018.1428892.
[77] H. Jacqmin-Gadda, C. Proust-Lima, J.M.G. Taylor, D. Commenges, Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model, Biometrics. 66 (2010) 11–19. https://doi.org/10.1111/j.1541-0420.2009.01234.x.
[78] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A.V.D. Linde, Bayesian measures of model complexity and fit, Journal of the Royal Statistical Society: Series B (Statistical Methodology). 64 (2002) 583–639. https://doi.org/10.1111/1467-9868.00353.
[79] E.R. Brown, J.G. Ibrahim, V. DeGruttola, A Flexible B-Spline Model for Multiple Longitudinal Biomarkers and Survival, Biometrics. 61 (2005) 64–73. https://doi.org/10.1111/j.0006-341X.2005.030929.x.
[80] A.E. Gelfand, D.K. Dey, H. Chang, Model Determination Using Predictive Distributions with Implementation via Sampling-Based Methods, (1992) 45.
[81] D. Zhang, M.-H. Chen, J.G. Ibrahim, M.E. Boye, W. Shen, Bayesian Model Assessment in Joint Modeling of Longitudinal and Survival Data with Applications to Cancer Clinical Trials, J Comput Graph Stat. 26 (2017) 121–133. https://doi.org/10.1080/10618600.2015.1117472.
[82] S. Watanabe, Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory, Journal of Machine Learning Research. 11 (2010) 3571–3594. http://jmlr.org/papers/v11/watanabe10a.html (accessed November 21, 2022).
[83] T. Hanson, STAT 740: Testing & Model Selection, (2017) 34.
[84] C. Proust-Lima, J.M.G. Taylor, Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach, Biostatistics. 10 (2009) 535–549. https://doi.org/10.1093/biostatistics/kxp009.
[85] D. Rizopoulos, Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data, Biometrics. 67 (2011) 819–829. https://doi.org/10.1111/j.1541-0420.2010.01546.x.

[86] M. Yu, J.M.G. Taylor, H.M. Sandler, Individual Prediction in Prostate Cancer Studies Using a Joint Longitudinal Survival–Cure Model, Journal of the American Statistical Association. 103 (2008) 178–187. https://doi.org/10.1198/016214507000000400.

[87] P. Blanche, J.-F. Dartigues, H. Jacqmin-Gadda, Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks, Statistics in Medicine. 32 (2013) 5381–5397. https://doi.org/10.1002/sim.5958.

[88] P. Blanche, C. Proust-Lima, L. Loubère, C. Berr, J.-F. Dartigues, H. Jacqmin-Gadda, Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks, Biometrics. 71 (2015) 102–113. https://doi.org/10.1111/biom.12232.

[89] Y. Zheng, P.J. Heagerty, Prospective Accuracy for Longitudinal Markers, Biometrics. 63 (2007) 332–341. https://doi.org/10.1111/j.1541-0420.2006.00726.x.

[90] P.J. Heagerty, Y. Zheng, Survival Model Predictive Accuracy and ROC Curves, Biometrics. 61 (2005) 92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x.

[91] P.C. Austin, F.E. Harrell, D. van Klaveren, Graphical calibration curves and the integrated calibration index (ICI) for survival models, Statistics in Medicine. 39 (2020) 2714–2742. https://doi.org/10.1002/sim.8570.

[92] C.S. Crowson, E.J. Atkinson, T.M. Therneau, Assessing calibration of prognostic risk scores, Stat Methods Med Res. 25 (2016) 1692–1706. https://doi.org/10.1177/0962280213497434.

[93] B. Van Calster, D.J. McLernon, M. van Smeden, L. Wynants, E.W. Steyerberg, P. Bossuyt, G.S. Collins, P. Macaskill, D.J. McLernon, K.G.M. Moons, E.W. Steyerberg, B. Van Calster, M. van Smeden, A.J. Vickers, On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative, Calibration: the Achilles heel of predictive analytics, BMC Medicine. 17 (2019) 230. https://doi.org/10.1186/s12916-019-1466-7.

[94] M.-C. Fournier, E. Dantan, P. Blanche, An R2-curve for evaluating the accuracy of dynamic predictions, Statistics in Medicine. 37 (2018) 1125–1133. https://doi.org/10.1002/sim.7571.

[95] T.A. Gerds, M. Schumacher, Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times, Biom. J. 48 (2006) 1029–1040. https://doi.org/10.1002/bimj.200610301.

[96] E. Graf, C. Schmoor, W. Sauerbrei, M. Schumacher, Assessment and comparison of prognostic classification schemes for survival data, Statistics in Medicine. 18 (1999) 2529–2545. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5.

[97] M. Schemper, R. Henderson, Predictive Accuracy and Explained Variation in Cox Regression, Biometrics. 56 (2000) 249–255. https://doi.org/10.1111/j.0006-341X.2000.00249.x.

[98] R. Henderson, P. Diggle, A. Dobson, Identification and efficacy of longitudinal markers for survival, Biostatistics. 3 (2002) 33–50. https://doi.org/10.1093/biostatistics/3.1.33.

[99] R. Schoop, E. Graf, M. Schumacher, Quantifying the Predictive Performance of Prognostic Models for Censored Survival Data with Time-Dependent Covariates, Biometrics. 64 (2008) 603–610. https://doi.org/10.1111/j.1541-0420.2007.00889.x.

[100] D. Commenges, B. Liquet, C. Proust-Lima, Choice of Prognostic Estimators in Joint Models by Estimating Differences of Expected Conditional Kullback–Leibler Risks, Biometrics. 68 (2012) 380–387. https://doi.org/10.1111/j.1541-0420.2012.01753.x.

[101] D. Commenges, C. Proust-Lima, C. Samieri, B. Liquet, A Universal Approximate Cross-Validation Criterion for Regular Risk Functions, The International Journal of Biostatistics. 11 (2015) 51–67. https://doi.org/10.1515/ijb-2015-0004.

[102] A.A. Tsiatis, M. Davidian, Joint Modeling of Longitudinal and Time-to-Event Data: An Overview, Statistica Sinica. 14 (2004) 809–834. https://www.jstor.org/stable/24307417 (accessed December 3, 2018).

[103] C. Proust-Lima, V. Philipps, B. Liquet, Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm, Journal of Statistical Software. 78 (2017) 1–56. https://doi.org/10.18637/jss.v078.i02.

[104] D. Rizopoulos, JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data, Journal of Statistical Software. 35 (2010) 1–33. https://doi.org/10.18637/jss.v035.i09.

[105] D.K. Pauler, D.M. Finkelstein, Predicting time to prostate cancer recurrence based on joint models for non-linear longitudinal biomarkers and event time outcomes, Statistics in Medicine. 21 (2002) 3897–3911. https://doi.org/10.1002/sim.1392.

[106] N.J. Law, J.M.G. Taylor, H. Sandler, The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure, Biostatistics. 3 (2002) 547–563. https://doi.org/10.1093/biostatistics/3.4.547.

[107] J.M.G. Taylor, M. Yu, H.M. Sandler, Individualized Predictions of Disease Progression Following Radiation Therapy for Prostate Cancer, JCO. 23 (2005) 816–825. https://doi.org/10.1200/JCO.2005.12.156.

[108] J.M.G. Taylor, Y. Park, D.P. Ankerst, C. Proust-Lima, S. Williams, L. Kestin, K. Bae, T. Pickles, H. Sandler, Real-Time Individual Predictions of Prostate Cancer Recurrence Using Joint Models, Biometrics. 69 (2013) 206–213. https://doi.org/10.1111/j.1541-0420.2012.01823.x.

[109] M. Sene, J.M.G. Taylor, J.J. Dignam, H. Jacqmin-Gadda, C. Proust-Lima, Individualized dynamic prediction of prostate cancer recurrence with and without the initiation of a second treatment: development and validation, Stat Methods Med Res. 25 (2016) 2972–2991. https://doi.org/10.1177/0962280214535763.

[110] L. Ferrer, V. Rondeau, J.J. Dignam, T. Pickles, H. Jacqmin-Gadda, C. Proust-Lima, Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer, Stat Med. 35 (2016) 3933–3948. https://doi.org/10.1002/sim.6972.

[111] L. Ferrer, H. Putter, C. Proust-Lima, Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment, Stat Methods Med Res. (2018) 0962280218811837. https://doi.org/10.1177/0962280218811837.

[112] C. Proust-Lima, J.M.G. Taylor, S. Sécher, H. Sandler, L. Kestin, T. Pickles, K. Bae, R. Allison, S. Williams, Confirmation of a Low α/β Ratio for Prostate Cancer Treated by External Beam Radiation Therapy Alone Using a Post-Treatment Repeated-Measures Model for PSA Dynamics, International Journal of Radiation Oncology*Biology*Physics. 79 (2011) 195–201. https://doi.org/10.1016/j.ijrobp.2009.10.008.

[113] S.D. Collins, N. Peek, R.D. Riley, G.P. Martin, Sample sizes of prediction model studies in prostate cancer were rarely justified and often insufficient, Journal of Clinical Epidemiology. 133 (2021) 53–60. https://doi.org/10.1016/j.jclinepi.2020.12.011.
[114] P.K. Andersen, R.B. Geskus, T. de Witte, H. Putter, Competing risks in epidemiology: possibilities and pitfalls, Int J Epidemiol. 41 (2012) 861–870. https://doi.org/10.1093/ije/dyr213.
[115] C. van Walraven, F.A. McAlister, Competing risk bias was common in Kaplan–Meier risk estimates published in prominent medical journals, Journal of Clinical Epidemiology. 69 (2016) 170-173.e8. https://doi.org/10.1016/j.jclinepi.2015.07.006.
[116] L.C. de Wreede, M. Fiocco, H. Putter, The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models, Computer Methods and Programs in Biomedicine. 99 (2010) 261–274. https://doi.org/10.1016/j.cmpb.2010.01.001.
[117] R.A. Mohammadpour, A. Alizadeh, M. Barzegartahamtan, A. Akbarzadeh Pasha, Association between prostate specific antigen change over time and prostate cancer recurrence risk: a joint model, Caspian Journal of Internal Medicine. 11 (2020) 324–328. https://doi.org/10.22088/cjim.11.3.324.
[118] X. Lin, J.M.G. Taylor, W. Ye, A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data, Stat. Interface. 1 (2008) 33–45. https://doi.org/10.4310/SII.2008.v1.n1.a4.
[119] S. Kim, D. Zeng, J.M.G. Taylor, Joint partially linear model for longitudinal data with informative drop-outs, Biometrics. 73 (2017) 72–82. https://doi.org/10.1111/biom.12566.
[120] H. Li, C. Gatsonis, Dynamic optimal strategy for monitoring disease recurrence, Sci. China Math. 55 (2012) 1565–1582. https://doi.org/10.1007/s11425-012-4475-y.
[121] M. Sène, C.A. Bellera, C. Proust-Lima, Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence, Journal de la société française de statistique. 155 (2014) 134–155. http://www.numdam.org/item/JSFS_2014__155_1_134_0/ (accessed January 8, 2022).
[122] A. Tomer, D. Nieboer, M.J. Roobol, A. Bjartell, E.W. Steyerberg, D. Rizopoulos, Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium, Personalized Biopsy Schedules Based on Risk of Gleason Upgrading for Low-Risk Prostate Cancer Active Surveillance Patients, BJU Int. (2020). https://doi.org/10.1111/bju.15136.
[123] A. Tomer, D. Nieboer, M.J. Roobol, E.W. Steyerberg, D. Rizopoulos, Personalized schedules for surveillance of low-risk prostate cancer patients, Biometrics. 75 (2019) 153–162. https://doi.org/10.1111/biom.12940.
[124] A. Tomer, D. Rizopoulos, D. Nieboer, F.-J. Drost, M.J. Roobol, E.W. Steyerberg, Personalized Decision Making for Biopsies in Prostate Cancer Active Surveillance Programs, Med Decis Making. (2019) 0272989X19861963. https://doi.org/10.1177/0272989X19861963.
[125] C. Serrat, M. Rué, C. Armero, X. Piulachs, H. Perpiñán, A. Forte, Á. Páez, G. Gómez, Frequentist and Bayesian approaches for a joint model for prostate cancer risk and longitudinal prostate-specific antigen data, Journal of Applied Statistics. 42 (2015) 1223–1239. https://doi.org/10.1080/02664763.2014.999032.
[126] H. Lin, B.W. Turnbull, C.E. McCulloch, E.H. Slate, Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal

prostate-specific antigen readings and prostate cancer, Journal of the American Statistical Association; Alexandria. 97 (2002) 53–65. https://search.proquest.com/docview/274779131/abstract/560B3B4920FB49F7PQ/1 (accessed August 7, 2019).

[127] R.Y. Coley, A.J. Fisher, M. Mamawala, H.B. Carter, K.J. Pienta, S.L. Zeger, A Bayesian hierarchical model for prediction of latent health states from multiple data sources with application to active surveillance of prostate cancer, Biometrics. 73 (2017) 625–634. https://doi.org/10.1111/biom.12577.

[128] S. Desmée, F. Mentré, C. Veyrat-Follet, J. Guedj, Nonlinear Mixed-Effect Models for Prostate-Specific Antigen Kinetics and Link with Survival in the Context of Metastatic Prostate Cancer: a Comparison by Simulation of Two-Stage and Joint Approaches, AAPS J. 17 (2015) 691–699. https://doi.org/10.1208/s12248-015-9745-5.

[129] S. Desmée, F. Mentré, C. Veyrat-Follet, B. Sébastien, J. Guedj, Using the SAEM algorithm for mechanistic joint models characterizing the relationship between nonlinear PSA kinetics and survival in prostate cancer patients, Biometrics. 73 (2016) 305–312. https://doi.org/10.1111/biom.12537.

[130] S. Desmée, F. Mentré, C. Veyrat-Follet, B. Sébastien, J. Guedj, Nonlinear joint models for individual dynamic prediction of risk of death using Hamiltonian Monte Carlo: application to metastatic prostate cancer, BMC Med Res Methodol. 17 (2017). https://doi.org/10.1186/s12874-017-0382-9.

[131] A. Finelli, T.M. Beer, S. Chowdhury, C.P. Evans, K. Fizazi, C.S. Higano, J. Kim, L. Martin, F. Saad, O. Saarela, Comparison of Joint and Landmark Modeling for Predicting Cancer Progression in Men With Castration-Resistant Prostate Cancer: A Secondary Post Hoc Analysis of the PREVAIL Randomized Clinical Trial, JAMA Network Open. 4 (2021) e2112426–e2112426. https://doi.org/10.1001/jamanetworkopen.2021.12426.

[132] K. Suresh, J.M.G. Taylor, A. Tsodikov, A copula-based approach for dynamic prediction of survival with a binary time-dependent covariate, Statistics in Medicine. 40 (2021) 4931–4946. https://doi.org/10.1002/sim.9102.

[133] G.L. Hickey, P. Philipson, A. Jorgensen, R. Kolamunnage-Dona, Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues, BMC Medical Research Methodology. 16 (2016) 117. https://doi.org/10.1186/s12874-016-0212-5.

[134] A. Tomer, D. Nieboer, M.J. Roobol, E.W. Steyerberg, D. Rizopoulos, Shared decision making of burdensome surveillance tests using personalized schedules and their burden and benefit, Statistics in Medicine. 41 (2022) 2115–2131. https://doi.org/10.1002/sim.9347.

[135] D.H. Brand, A.C. Tree, P. Ostler, H. van der Voet, A. Loblaw, W. Chu, D. Ford, S. Tolan, S. Jain, A. Martin, J. Staffurth, P. Camilleri, K. Kancherla, J. Frew, A. Chan, I.S. Dayes, D. Henderson, S. Brown, C. Cruickshank, S. Burnett, A. Duffton, C. Griffin, V. Hinder, K. Morrison, O. Naismith, E. Hall, N. van As, D. Dodds, E. Lartigau, S. Patton, A. Thompson, M. Winkler, P. Wells, T. Lymberiou, D. Saunders, M. Vilarino-Varela, P. Vavassis, T. Tsakiridis, R. Carlson, G. Rodrigues, J. Tanguay, S. Iqbal, M. Winkler, S. Morgan, A. Mihai, A. Li, O. Din, M. Panades, R. Wade, Y. Rimmer, J. Armstrong, M. Panades, N. Oommen, Intensity-modulated fractionated radiotherapy versus stereotactic body radiotherapy for prostate cancer (PACE-B): acute toxicity findings

from an international, randomised, open-label, phase 3, non-inferiority trial, The Lancet Oncology. 20 (2019) 1531–1543. https://doi.org/10.1016/S1470-2045(19)30569-8.

[136] D. Rizopoulos, G. Papageorgiou, P.M. Afonso, JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data, 2022.

[137] D. Rizopoulos, L.A. Hatfield, B.P. Carlin, J.J.M. Takkenberg, Combining Dynamic Predictions From Joint Models for Longitudinal and Time-to-Event Data Using Bayesian Model Averaging, Journal of the American Statistical Association. 109 (2014) 1385–1397. https://doi.org/10.1080/01621459.2014.931236.

[138] D. Tilki, A.V. D'Amico, Timing of radiotherapy after radical prostatectomy, The Lancet. 396 (2020) 1374–1375. https://doi.org/10.1016/S0140-6736(20)31957-7.

[139] F. Zattoni, I. Heidegger, V. Kasivisvanathan, A. Kretschmer, G. Marra, A. Magli, F. Preisser, D. Tilki, I. Tsaur, M. Valerio, R. van den Bergh, C. Kesch, F. Ceci, C. Fankhauser, G. Gandaglia, Radiation Therapy After Radical Prostatectomy: What Has Changed Over Time?, Frontiers in Surgery. 8 (2021) 245. https://doi.org/10.3389/fsurg.2021.691473.

[140] H.A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S.H. AlDubayan, K. Salari, S. Kregel, C. Richter, T.E. Arnoff, J. Park, W.C. Hahn, E.M. Van Allen, Biologically informed deep neural network for prostate cancer discovery, Nature. 598 (2021) 348–352. https://doi.org/10.1038/s41586-021-03922-4.

[141] S. Bernatz, J. Ackermann, P. Mandel, B. Kaltenbach, Y. Zhdanovich, P.N. Harter, C. Döring, R. Hammerstingl, B. Bodelle, K. Smith, A. Bucher, M. Albrecht, N. Rosbach, L. Basten, I. Yel, M. Wenzel, K. Bankov, I. Koch, F.K.-H. Chun, J. Köllermann, P.J. Wild, T.J. Vogl, Comparison of machine learning algorithms to predict clinically significant prostate cancer of the peripheral zone with multiparametric MRI using clinical assessment categories and radiomic features, Eur Radiol. 30 (2020) 6757–6769. https://doi.org/10.1007/s00330-020-07064-5.

[142] S. Yu, J. Tao, B. Dong, Y. Fan, H. Du, H. Deng, J. Cui, G. Hong, X. Zhang, Development and head-to-head comparison of machine-learning models to identify patients requiring prostate biopsy, BMC Urol. 21 (2021) 80. https://doi.org/10.1186/s12894-021-00849-w.

[143] S.B. Ginsburg, M. Rusu, J. Kurhanewicz, A. Madabhushi, Computer extracted texture features on T2w MRI to predict biochemical recurrence following radiation therapy for prostate cancer, in: Medical Imaging 2014: Computer-Aided Diagnosis, SPIE, 2014: pp. 69–81. https://doi.org/10.1117/12.2043937.

[144] N. Momenzadeh, H. Hafezalseheh, M.R. Nayebpour, M. Fathian, R. Noorossana, A hybrid machine learning approach for predicting survival of patients with prostate cancer: A SEER-based population study, Informatics in Medicine Unlocked. 27 (2021) 100763. https://doi.org/10.1016/j.imu.2021.100763.

[145] O.S. Tătaru, M.D. Vartolomei, J.J. Rassweiler, O. Virgil, G. Lucarelli, F. Porpiglia, D. Amparore, M. Manfredi, G. Carrieri, U. Falagario, D. Terracciano, O. de Cobelli, G.M. Busetto, F.D. Giudice, M. Ferro, Artificial Intelligence and Machine Learning in Prostate Cancer Patient Management—Current Trends and Future Perspectives, Diagnostics (Basel). 11 (2021) 354. https://doi.org/10.3390/diagnostics11020354.

[146] M. Raynaud, O. Aubert, G. Divard, P.P. Reese, N. Kamar, D. Yoo, C.-S. Chin, É. Bailly, M. Buchler, M. Ladrière, M. Le Quintrec, M. Delahousse, I. Juric, N. Basic-Jukic, M. Crespo, H.T. Silva, K. Linhares, M.C. Ribeiro de Castro, G. Soler Pujol, J.-P. Empana, C. Ulloa, E. Akalin, G. Böhmig, E. Huang, M.D. Stegall, A.J. Bentall, R.A. Montgomery,

S.C. Jordan, R. Oberbauer, D.L. Segev, J.J. Friedewald, X. Jouven, C. Legendre, C. Lefaucheur, A. Loupy, Dynamic prediction of renal survival among deeply phenotyped kidney transplant recipients using artificial intelligence: an observational, international, multicohort study, The Lancet Digital Health. 3 (2021) e795–e805. https://doi.org/10.1016/S2589-7500(21)00209-0.

[147] E. Waldmann, D. Taylor-Robinson, N. Klein, T. Kneib, T. Pressler, M. Schmid, A. Mayr, Boosting joint models for longitudinal and time-to-event data, Biometrical Journal. 59 (2017) 1104–1121. https://doi.org/10.1002/bimj.201600158.

[148] P. Dhiman, J. Ma, C.A. Navarro, B. Speich, G. Bullock, J.A. Damen, S. Kirtley, L. Hooft, R.D. Riley, B. Van Calster, K.G.M. Moons, G.S. Collins, Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved, Journal of Clinical Epidemiology. 138 (2021) 60–72. https://doi.org/10.1016/j.jclinepi.2021.06.024.

[149] C.L. Andaur Navarro, J.A.A. Damen, T. Takada, S.W.J. Nijman, P. Dhiman, J. Ma, G.S. Collins, R. Bajpai, R.D. Riley, K.G.M. Moons, L. Hooft, Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review, BMC Medical Research Methodology. 22 (2022) 12. https://doi.org/10.1186/s12874-021-01469-6.

[150] A.J. Vickers, E.B. Elkin, Decision Curve Analysis: A Novel Method for Evaluating Prediction Models, Med Decis Making. 26 (2006) 565–574. https://doi.org/10.1177/0272989X06295361.

[151] E.W. Steyerberg, A.J. Vickers, N.R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M.J. Pencina, M.W. Kattan, Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures, Epidemiology. 21 (2010) 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2.

[152] X. Liu, L. Faes, M.J. Calvert, A.K. Denniston, Extension of the CONSORT and SPIRIT statements, The Lancet. 394 (2019) 1225. https://doi.org/10.1016/S0140-6736(19)31819-7.

[153] The Lancet Digital Health, Walking the tightrope of artificial intelligence guidelines in clinical practice, The Lancet Digital Health. 1 (2019) e100. https://doi.org/10.1016/S2589-7500(19)30063-9.

[154] G.S. Collins, P. Dhiman, C.L.A. Navarro, J. Ma, L. Hooft, J.B. Reitsma, P. Logullo, A.L. Beam, L. Peng, B.V. Calster, M. van Smeden, R.D. Riley, K.G. Moons, Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence, BMJ Open. 11 (2021) e048008. https://doi.org/10.1136/bmjopen-2020-048008.

[155] G.S. Collins, K.G.M. Moons, Reporting of artificial intelligence prediction models, The Lancet. 393 (2019) 1577–1579. https://doi.org/10.1016/S0140-6736(19)30037-6.

[156] H. Parr, N. Porta, A.C. Tree, D. Dearnaley, E. Hall, Dynamic Prediction Models to Characterise the Prognosis of Post-Radiotherapy Prostate Cancer Patients in the CHHiP Clinical Trial, (2019).

[157] I. Fornacon-Wood, H. Mistry, C. Johnson-Hart, C. Faivre-Finn, J.P.B. O'Connor, G.J. Price, Understanding the Differences Between Bayesian and Frequentist Statistics, International Journal of Radiation Oncology, Biology, Physics. 112 (2022) 1076–1082. https://doi.org/10.1016/j.ijrobp.2021.12.011.

[158] D. Rizopoulos, P. Ghosh, A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event, Statistics in Medicine. 30 (2011) 1366–1380. https://doi.org/10.1002/sim.4205.
[159] F. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd ed., Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-19425-7.
[160] Z. Wang, Y. Ni, J. Chen, G. Sun, X. Zhang, J. Zhao, X. Zhu, H. Zhang, S. Zhu, J. Dai, P. Shen, H. Zeng, The efficacy and safety of radical prostatectomy and radiotherapy in high-risk prostate cancer: a systematic review and meta-analysis, World Journal of Surgical Oncology. 18 (2020) 42. https://doi.org/10.1186/s12957-020-01824-9.
[161] C.G. Mazariego, S. Egger, M.T. King, I. Juraskova, H. Woo, M. Berry, B.K. Armstrong, D.P. Smith, Fifteen year quality of life outcomes in men with localised prostate cancer: population based Australian prospective study, BMJ. 371 (2020) m3503. https://doi.org/10.1136/bmj.m3503.
[162] J.L. Donovan, F.C. Hamdy, J.A. Lane, M. Mason, C. Metcalfe, E. Walsh, J.M. Blazeby, T.J. Peters, P. Holding, S. Bonnington, T. Lennon, L. Bradshaw, D. Cooper, P. Herbert, J. Howson, A. Jones, N. Lyons, E. Salter, P. Thompson, S. Tidball, J. Blaikie, C. Gray, P. Bollina, J. Catto, A. Doble, A. Doherty, D. Gillatt, R. Kockelbergh, H. Kynaston, A. Paul, P. Powell, S. Prescott, D.J. Rosario, E. Rowe, M. Davis, E.L. Turner, R.M. Martin, D.E. Neal, Patient-Reported Outcomes after Monitoring, Surgery, or Radiotherapy for Prostate Cancer, N Engl J Med. 375 (2016) 1425–1437. https://doi.org/10.1056/NEJMoa1606221.
[163] Recommendations | Prostate cancer: diagnosis and management | Guidance | NICE, (2021). https://www.nice.org.uk/guidance/ng131/chapter/Recommendations.
[164] D. Rodin, B. Tawk, O. Mohamad, S. Grover, F.Y. Moraes, M.L. Yap, E. Zubizarreta, Y. Lievens, Hypofractionated radiotherapy in the real-world setting: An international ESTRO-GIRO survey, Radiotherapy and Oncology. 157 (2021) 32–39. https://doi.org/10.1016/j.radonc.2021.01.003.
[165] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, R Core Team, nlme: Linear and Nonlinear Mixed Effects Models, 2021. https://CRAN.R-project.org/package=nlme.
[166] T.M. Therneau, A Package for Survival Analysis in R, 2021. https://CRAN.R-project.org/package=survival.
[167] H. Putter, M. Fiocco, R.B. Geskus, Tutorial in biostatistics: competing risks and multi-state models, Statistics in Medicine. 26 (2007) 2389–2430. https://doi.org/10.1002/sim.2712.
[168] C.J. Stone, [Generalized Additive Models]: Comment, Statist. Sci. 1 (1986) 312–314. https://doi.org/10.1214/ss/1177013607.
[169] F.B. Geara, M. Bulbul, R.B. Khauli, T.Y. Andraos, M. Abboud, A. Al Mousa, N. Sarhan, A. Salem, H. Ghatasheh, A. Alnsour, Z. Ayoub, I.A. Gheida, M. Charafeddine, M. Shahait, A. Shamseddine, R.A. Gheida, J. Khader, Nadir PSA is a strong predictor of treatment outcome in intermediate and high risk localized prostate cancer patients treated by definitive external beam radiotherapy and androgen deprivation, Radiat Oncol. 12 (2017). https://doi.org/10.1186/s13014-017-0884-y.
[170] J. Herbers, R. Miller, A. Walther, L. Schindler, K. Schmidt, W. Gao, F. Rupprecht, How to deal with non-detectable and outlying values in biomarker research: Best practices and recommendations for univariate imputation approaches, Comprehensive

Psychoneuroendocrinology. 7 (2021) 100052. https://doi.org/10.1016/j.cpnec.2021.100052.

[171] J.W. Davis, P. Kolm, G.L. Wright, D. Kuban, -Mahdi Anas El, P.F. Schellhammer, The durability of external beam radiation therapy for prostate cancer: can it be identified?, Journal of Urology. 162 (1999) 758–761. https://doi.org/10.1097/00005392-199909010-00036.

[172] T.I. Yock, A.L. Zietman, W.U. Shipley, H.K. Thakral, J.J. Coen, Long-term durability of PSA failure-free survival after radiotherapy for localized prostate cancer, International Journal of Radiation Oncology, Biology, Physics. 54 (2002) 420–426. https://doi.org/10.1016/S0360-3016(02)02957-7.

[173] J.M. Crook, C. Tang, H. Thames, P. Blanchard, J. Sanders, J. Ciezki, M. Keyes, W.J. Morris, G. Merrick, C. Catton, H. Raziee, R. Stock, F. Sullivan, M. Anscher, J. Millar, S. Frank, A biochemical definition of cure after brachytherapy for prostate cancer, Radiotherapy and Oncology. 149 (2020) 64–69. https://doi.org/10.1016/j.radonc.2020.04.038.

[174] R.D. Riley, K.I. Snell, J. Ensor, D.L. Burke, F.E. Harrell Jr, K.G. Moons, G.S. Collins, Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes, Statistics in Medicine. 38 (2019) 1276–1296. https://doi.org/10.1002/sim.7992.

[175] R.D. Riley, J. Ensor, K.I.E. Snell, F.E. Harrell, G.P. Martin, J.B. Reitsma, K.G.M. Moons, G. Collins, M. van Smeden, Calculating the sample size required for developing a clinical prediction model, BMJ. 368 (2020). https://doi.org/10.1136/bmj.m441.

[176] E.-R. Andrinopoulou, P.H.C. Eilers, J.J.M. Takkenberg, D. Rizopoulos, Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using P-splines, Biometrics. 74 (2018) 685–693. https://doi.org/10.1111/biom.12814.

[177] E.-R. Andrinopoulou, D. Rizopoulos, Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures: Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures, Statistics in Medicine. 35 (2016) 4813–4823. https://doi.org/10.1002/sim.7027.

[178] Prostate cancer survival statistics, Cancer Research UK. (2015). https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer/survival (accessed September 5, 2022).

[179] E.-R. Andrinopoulou, D. Rizopoulos, J.J. Takkenberg, E. Lesaffre, Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data, Stat Methods Med Res. 26 (2015) 1787–1801. https://doi.org/10.1177/0962280215588340.

[180] L. Klotz, C. O'Callaghan, K. Ding, P. Toren, D. Dearnaley, C.S. Higano, E. Horwitz, S. Malone, L. Goldenberg, M. Gospodarowicz, J.M. Crook, Nadir Testosterone Within First Year of Androgen-Deprivation Therapy (ADT) Predicts for Time to Castration-Resistant Progression: A Secondary Analysis of the PR-7 Trial of Intermittent Versus Continuous ADT, JCO. 33 (2015) 1151–1156. https://doi.org/10.1200/JCO.2014.58.2973.

[181] S. Sumanasuriya, G. Seed, H. Parr, R. Christova, L. Pope, C. Bertan, D. Bianchini, P. Rescigno, I. Figueiredo, J. Goodall, G. Fowler, P. Flohr, N. Mehra, A. Neeb, J. Rekowski, M. Eisenberger, O. Sartor, S. Oudard, C. Geffriaud-Ricouard, A. Ozatilgan, M. Chadjaa, S. Macé, C. Lord, J. Baxter, S. Pettitt, M. Lambros, A. Sharp, J. Mateo, S. Carreira, W. Yuan, J.S. de Bono, Elucidating Prostate Cancer Behaviour During

Treatment via Low-pass Whole-genome Sequencing of Circulating Tumour DNA, European Urology. (2021). https://doi.org/10.1016/j.eururo.2021.05.030.

[182] A. Wilkins, B. Gusterson, Z. Szijgyarto, J. Haviland, C. Griffin, C. Stuttle, F. Daley, C.M. Corbishley, D.P. Dearnaley, E. Hall, N. Somaiah, Ki67 Is an Independent Predictor of Recurrence in the Largest Randomized Trial of 3 Radiation Fractionation Schedules in Localized Prostate Cancer, International Journal of Radiation Oncology • Biology • Physics. 101 (2018) 309–315. https://doi.org/10.1016/j.ijrobp.2018.01.072.

[183] A.J. Vickers, B. van Calster, E.W. Steyerberg, A simple, step-by-step guide to interpreting decision curve analysis, Diagnostic and Prognostic Research. 3 (2019) 18. https://doi.org/10.1186/s41512-019-0064-7.

[184] D.P. Dearnaley, M.R. Sydes, J.D. Graham, E.G. Aird, D. Bottomley, R.A. Cowan, R.A. Huddart, C.C. Jose, J.H. Matthews, J. Millar, A.R. Moore, R.C. Morgan, J.M. Russell, C.D. Scrase, R.J. Stephens, I. Syndikus, M.K. Parmar, Escalated-dose versus standard-dose conformal radiotherapy in prostate cancer: first results from the MRC RT01 randomised controlled trial, The Lancet Oncology. 8 (2007) 475–487. https://doi.org/10.1016/S1470-2045(07)70143-2.

[185] R Core Team, R: A Language and Environment for Statistical Computing, (2021). https://www.R-project.org.

[186] J. Epstein, L. Egevad, M. Amin, B. Delahunt, J. Srigley, P. Humphrey, T. Al Hussain, A. F., M. Aron, B. D., B. D., B. F., C. D., J. Cheville, D. Clouston, M. Colecchia, E. Compérat, I. Cunha, de A., R. G., The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma, The American Journal of Surgical Pathology. 40 (2015). https://doi.org/10.1097/PAS.0000000000000530.

[187] F. Brimo, R. Montironi, L. Egevad, A. Erbersdobler, D.W. Lin, J.B. Nelson, M.A. Rubin, T. van der Kwast, M. Amin, J.I. Epstein, Contemporary Grading for Prostate Cancer: Implications for Patient Care, European Urology. 63 (2013) 892–901. https://doi.org/10.1016/j.eururo.2012.10.015.

[188] A. Freeman, Prognostic Gleason grade grouping: data based on the modified Gleason scoring system, BJU International. 111 (2013) 691–692. https://doi.org/10.1111/j.1464-410X.2012.11743.x.

[189] P.A. Humphrey, Gleason grading and prognostic factors in carcinoma of the prostate, Mod Pathol. 17 (2004) 292–306. https://doi.org/10.1038/modpathol.3800054.

[190] S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software. 45 (2011) 1–67. https://doi.org/10.18637/jss.v045.i03.

[191] D.L. Fitch, S. McGrath, A.A. Martinez, F.A. Vicini, L.L. Kestin, Unification of a common biochemical failure definition for prostate cancer treated with brachytherapy or external beam radiotherapy with or without androgen deprivation, International Journal of Radiation Oncology, Biology, Physics. 66 (2006) 1430–1439. https://doi.org/10.1016/j.ijrobp.2006.03.024.

[192] R.D. Riley, G.S. Collins, J. Ensor, L. Archer, S. Booth, S.I. Mozumder, M.J. Rutherford, M. van Smeden, P.C. Lambert, K.I.E. Snell, Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome, Statistics in Medicine. n/a (2021). https://doi.org/10.1002/sim.9275.

[193] L.M. Chen, J.G. Ibrahim, H. Chu, Sample size and power determination in joint modeling of longitudinal and survival data, Stat Med. 30 (2011) 2295–2309. https://doi.org/10.1002/sim.4263.
[194] A. Tomer, D. Nieboer, M.J. Roobol, A. Bjartell, E.W. Steyerberg, D. Rizopoulos, Personalised biopsy schedules based on risk of Gleason upgrading for patients with low-risk prostate cancer on active surveillance, BJU International. 127 (2021) 96–107. https://doi.org/10.1111/bju.15136.
[195] P.C. Austin, D.S. Lee, J.P. Fine, Introduction to the Analysis of Survival Data in the Presence of Competing Risks, Circulation. 133 (2016) 601–609. https://doi.org/10.1161/CIRCULATIONAHA.115.017719.
[196] T. Therneau, C. Crowson, E. Atkinson, Multi-state models and competing risks, (2023) 29. https://www.vps.fmvz.usp.br/CRAN/web/packages/survival/vignettes/compete.pdf.
[197] C.D. Serio, The Protective Impact of a Covariate on Competing Failures with an Example from a Bone Marrow Transplantation Study, Lifetime Data Anal. 3 (1997) 99–122. https://doi.org/10.1023/A:1009672300875.
[198] C.A. Thompson, Z.-F. Zhang, O.A. Arah, Competing risk bias to explain the inverse relationship between smoking and malignant melanoma, Eur J Epidemiol. 28 (2013) 557–567. https://doi.org/10.1007/s10654-013-9812-0.
[199] F. Kunath, H.R. Grobe, G. Rücker, E. Motschall, G. Antes, P. Dahm, B. Wullich, J.J. Meerpohl, Non-steroidal antiandrogen monotherapy compared with luteinising hormone–releasing hormone agonists or surgical castration monotherapy for advanced prostate cancer, Cochrane Database of Systematic Reviews. (2014). https://doi.org/10.1002/14651858.CD009266.pub2.
[200] J. Seidenfeld, D. J. Samson, V. Hasselblad, N. Aronson, P. C. Albertsen, C. L. Bennett, T. J. Wilt, Single-Therapy Androgen Suppression in Men with Advanced Prostate Cancer, Annals of Internal Medicine. (2000). https://www.acpjournals.org/doi/10.7326/0003-4819-132-7-200004040-00009 (accessed March 8, 2023).
[201] U. McGivern, D.M. Mitchell, C. McDowell, J. O'Hare, G. Corey, J.M. O'Sullivan, Neoadjuvant Hormone Therapy for Radical Prostate Radiotherapy: Bicalutamide Monotherapy vs. Luteinizing Hormone–Releasing Hormone Agonist Monotherapy: A Single-Institution Matched-Pair Analysis, Clinical Genitourinary Cancer. 10 (2012) 190–195. https://doi.org/10.1016/j.clgc.2012.04.003.
[202] A. Baydoun, Y. Sun, H.M. Sandler, M. Bolla, A. Nabid, J.W. Denham, A.Y. Jia, N.G. Zaorsky, J. Garcia, J. Brown, W.C. Jackson, R.T. Dess, J.A. Efstathiou, F.Y. Feng, P. Maingon, A. Steigler, L. Souhami, A. Berlin, A.U. Kishan, D.E. Spratt, Efficacy of Bicalutamide Monotherapy in Prostate Cancer: A Network Meta-Analysis of 10 Randomized Trials, International Journal of Radiation Oncology*Biology*Physics. 114 (2022) e211–e212. https://doi.org/10.1016/j.ijrobp.2022.07.1146.
[203] A.Y. Jia, D.E. Spratt, Bicalutamide Monotherapy With Radiation Therapy for Localized Prostate Cancer: A Non-Evidence-Based Alternative, International Journal of Radiation Oncology*Biology*Physics. 113 (2022) 316–319. https://doi.org/10.1016/j.ijrobp.2022.01.037.
[204] National life tables – life expectancy in the UK - Office for National Statistics, (2021). https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/

lifeexpectancies/bulletins/nationallifetablesunitedkingdom/2018to2020 (accessed March 22, 2023).

[205] P.C. Austin, H. Putter, D. Giardiello, D. van Klaveren, Graphical calibration curves and the integrated calibration index (ICI) for competing risk models, Diagnostic and Prognostic Research. 6 (2022) 2. https://doi.org/10.1186/s41512-021-00114-6.

[206] N. van Geloven, D. Giardiello, E.F. Bonneville, L. Teece, C.L. Ramspek, M. van Smeden, K.I.E. Snell, B. van Calster, M. Pohar-Perme, R.D. Riley, H. Putter, E. Steyerberg, Validation of prediction models in the presence of competing risks: a guide through modern methods, BMJ. 377 (2022) e069249. https://doi.org/10.1136/bmj-2021-069249.

[207] Y. Zheng, T. Cai, Y. Jin, Z. Feng, Evaluating Prognostic Accuracy of Biomarkers under Competing Risk, Biometrics. 68 (2012) 388–396. https://doi.org/10.1111/j.1541-0420.2011.01671.x.

[208] I.D. Kaplan, R.S. Cox, M.A. Bagshaw, A model of prostatic carcinoma tumor kinetics based on prostate specific antigen levels after radiation therapy, Cancer. 68 (1991) 400–405. https://doi.org/10.1002/1097-0142(19910715)68:2<400::AID-CNCR2820680231>3.0.CO;2-Z.

[209] G.K. Zagars, A. Pollack, The fall and rise of prostate-specific antigen: Kinetics of serum prostate-specific antigen levels after radiation therapy for prostate cancer, Cancer. 72 (1993) 832–842. https://doi.org/10.1002/1097-0142(19930801)72:3<832::AID-CNCR2820720332>3.0.CO;2-6.

[210] C.L. Faucett, D.C. Thomas, Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach, Statistics in Medicine. 15 (1996) 1663–1685. https://doi.org/10.1002/(SICI)1097-0258(19960815)15:15<1663::AID-SIM294>3.0.CO;2-1.

[211] M.S. Wulfsohn, A.A. Tsiatis, A Joint Model for Survival and Longitudinal Data Measured with Error, Biometrics. 53 (1997) 330–339. https://doi.org/10.2307/2533118.

[212] S. Self, Y. Pawitan, Modeling a Marker of Disease Progression and Onset of Disease, in: N.P. Jewell, K. Dietz, V.T. Farewell (Eds.), AIDS Epidemiology: Methodological Issues, Birkhäuser Boston, Boston, MA, 1992: pp. 231–255. https://doi.org/10.1007/978-1-4757-1229-2_11.

[213] V. De Gruttola, X.M. Tu, Modelling Progression of CD4-Lymphocyte Count and Its Relationship to Survival Time, Biometrics. 50 (1994) 1003–1014. https://doi.org/10.2307/2533439.

[214] N.P. Jewell, J.P. Nielsen, A framework for consistent prediction rules based on markers, Biometrika. 80 (1993) 153–164. https://doi.org/10.1093/biomet/80.1.153.

[215] K. Suresh, J.M.G. Taylor, A. Tsodikov, A Gaussian copula approach for dynamic prediction of survival with a longitudinal biomarker, Biostatistics. 22 (2019) 504–521. https://doi.org/10.1093/biostatistics/kxz049.

[216] J. Lin, S. Luo, Deep learning for the dynamic prediction of multivariate longitudinal and survival data, Statistics in Medicine. 41 (2022) 2894–2907. https://doi.org/10.1002/sim.9392.

[217] K.L. Pickett, K. Suresh, K.R. Campbell, S. Davis, E. Juarez-Colunga, Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker, BMC Medical Research Methodology. 21 (2021) 216. https://doi.org/10.1186/s12874-021-01375-x.

[218] C. Lee, W. Zame, J. Yoon, M. Van der Schaar, DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks, AAAI. 32 (2018). https://doi.org/10.1609/aaai.v32i1.11842.

[219] C. Lee, J. Yoon, M. van der Schaar, Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data, IEEE Transactions on Biomedical Engineering. 67 (2020) 122–133. https://doi.org/10.1109/TBME.2019.2909027.

[220] C. Lee, A. Light, E.S. Saveliev, M. van der Schaar, V.J. Gnanapragasam, Developing machine learning algorithms for dynamic estimation of progression during active surveillance for prostate cancer, Npj Digit. Med. 5 (2022) 1–7. https://doi.org/10.1038/s41746-022-00659-w.

[221] X. Dai, J.H. Park, S. Yoo, N. D'Imperio, B.H. McMahon, C.T. Rentsch, J.P. Tate, A.C. Justice, Survival analysis of localized prostate cancer with deep learning, Sci Rep. 12 (2022) 17821. https://doi.org/10.1038/s41598-022-22118-y.

[222] J.-E. Bibault, S. Hancock, M.K. Buyyounouski, H. Bagshaw, J.T. Leppert, J.C. Liao, L. Xing, Development and Validation of an Interpretable Artificial Intelligence Model to Predict 10-Year Prostate Cancer Mortality, Cancers. 13 (2021) 3064. https://doi.org/10.3390/cancers13123064.

[223] H. Pinckaers, J. van Ipenburg, J. Melamed, A. De Marzo, E.A. Platz, B. van Ginneken, J. van der Laak, G. Litjens, Predicting biochemical recurrence of prostate cancer with artificial intelligence, Commun Med. 2 (2022) 1–9. https://doi.org/10.1038/s43856-022-00126-3.

[224] R. Toth, H. Schiffmann, C. Hube-Magg, F. Büscheck, D. Höflmayer, S. Weidemann, P. Lebok, C. Fraune, S. Minner, T. Schlomm, G. Sauter, C. Plass, Y. Assenov, R. Simon, J. Meiners, C. Gerhäuser, Random forest-based modelling to detect biomarkers for prostate cancer progression, Clinical Epigenetics. 11 (2019) 148. https://doi.org/10.1186/s13148-019-0736-8.

[225] P. Dhiman, J. Ma, C.L. Andaur Navarro, B. Speich, G. Bullock, J.A.A. Damen, L. Hooft, S. Kirtley, R.D. Riley, B. Van Calster, K.G.M. Moons, G.S. Collins, Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review, BMC Medical Research Methodology. 22 (2022) 101. https://doi.org/10.1186/s12874-022-01577-x.

[226] T. van der Ploeg, P.C. Austin, E.W. Steyerberg, Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints, BMC Medical Research Methodology. 14 (2014) 137. https://doi.org/10.1186/1471-2288-14-137.

[227] E. Christodoulou, J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, B. Van Calster, A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, J Clin Epidemiol. 110 (2019) 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.

[228] B.Y. Gravesteijn, D. Nieboer, A. Ercole, H.F. Lingsma, D. Nelson, B. van Calster, E.W. Steyerberg, C. Åkerlund, K. Amrein, N. Andelic, L. Andreassen, A. Anke, A. Antoni, G. Audibert, P. Azouvi, M.L. Azzolini, R. Bartels, P. Barzó, R. Beauvais, R. Beer, B.-M. Bellander, A. Belli, H. Benali, M. Berardino, L. Beretta, M. Blaabjerg, P. Bragge, A. Brazinova, V. Brinck, J. Brooker, C. Brorsson, A. Buki, M. Bullinger, M. Cabeleira, A. Caccioppola, E. Calappi, M.R. Calvi, P. Cameron, G.C. Lozano, M. Carbonara, G. Chevallard, A. Chieregato, G. Citerio, M. Cnossen, M. Coburn, J. Coles, D.J. Cooper, M. Correia, A. Čović, N. Curry, E. Czeiter, M. Czosnyka, C. Dahyot-Fizelier, H.

Dawes, V. De Keyser, V. Degos, F. Della Corte, H. den Boogert, B. Depreitere, Đ. Đilvesi, A. Dixit, E. Donoghue, J.D. Guy-Loup Dulière, A. Ercole, P. Esser, E.E. Martin Fabricius, K.F. Feigin Valery L., S. Frisvold, A. Furmanov, P. Gagliardo, D. Galanaud, D. Gantner, G. Gao, P. George, A. Ghuysen, L. Giga, B. Glocker, J. Golubovic, P.A. Gomez, J. Gratz, B. Gravesteijn, F. Grossi, R.L. Gruen, D. Gupta, J.A. Haagsma, I. Haitsma, R. Helbok, E. Helseth, L. Horton, J. Huijben, P.J. Hutchinson, B. Jacobs, S. Jankowski, M.J. Ji-yao Jiang, K. Jones, M. Karan, A.G. Kolias, E. Kompanje, D. Kondziella, E. Koraropoulos, L.-O. Koskinen, N. Kovács, A. Lagares, L. Lanyon, S. Laureys, F. Lecky, R. Lefering, V. Legrand, A. Lejeune, L. Levi, R. Lightfoot, H. Lingsma, A.I.R. Maas, A.M. Castaño-León, M. Maegele, M. Majdan, A. Manara, G. Manley, C. Martino, H. Maréchal, J. Mattern, C. McMahon, B. Melegh, D. Menon, T. Menovsky, D. Mulazzi, V. Muraleedharan, L. Murray, N. Nair, A. Negru, D. Nelson, V. Newcombe, D. Nieboer, Q. Noirhomme, J. Nyirádi, O. Olubukola, M. Oresic, F. Ortolano, A. Palotie, P.M. Parizel, J.-F. Payen, N. Perera, V. Perlbarg, P. Persona, W. Peul, A. Piippo-Karjalainen, M. Pirinen, H. Ples, S. Polinder, I. Pomposo, J.P. Posti, L. Puybasset, A. Radoi, A. Ragauskas, R. Raj, M. Rambadagalla, R. Real, J. Rhodes, S. Richardson, S. Richter, S. Ripatti, S. Rocka, C. Roe, O. Roise, J. Rosand, J.V. Rosenfeld, C. Rosenlund, G. Rosenthal, R. Rossaint, S. Rossi, D. Rueckert, M. Rusnák, J. Sahuquillo, O. Sakowitz, R. Sanchez-Porras, J. Sandor, N. Schäfer, S. Schmidt, H. Schoechl, G. Schoonman, R.F. Schou, E. Schwendenwein, C. Sewalt, T. Skandsen, P. Smielewski, A. Sorinola, E. Stamatakis, S. Stanworth, A. Kowark, R. Stevens, W. Stewart, E.W. Steyerberg, N. Stocchetti, N. Sundström, A. Synnot, R. Takala, V. Tamás, T. Tamosuitis, M.S. Taylor, B.T. Ao, O. Tenovuo, A. Theadom, M. Thomas, D. Tibboel, M. Timmers, C. Tolias, T. Trapani, C.M. Tudora, P. Vajkoczy, S. Vallance, E. Valeinis, Z. Vámos, G. Van der Steen, J. van der Naalt, J.T.J.M. van Dijck, T.A. van Essen, W. Van Hecke, C. van Heugten, D. Van Praag, T.V. Vyvere, A. Vanhaudenhuyse, R.P.J. van Wijk, A. Vargiolu, E. Vega, K. Velt, J. Verheyden, P.M. Vespa, A. Vik, R. Vilcinis, V. Volovici, N. von Steinbüchel, D. Voormolen, P. Vulekovic, K.K.W. Wang, E. Wiegers, G. Williams, L. Wilson, S. Winzeck, S. Wolf, Z. Yang, P. Ylén, A. Younsi, F.A. Zeiler, V. Zelinkova, A. Ziverte, T. Zoerle, Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury, Journal of Clinical Epidemiology. 122 (2020) 95–107. https://doi.org/10.1016/j.jclinepi.2020.03.005.

[229] A.L. Lynam, J.M. Dennis, K.R. Owen, R.A. Oram, A.G. Jones, B.M. Shields, L.A. Ferrat, Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults, Diagnostic and Prognostic Research. 4 (2020) 6. https://doi.org/10.1186/s41512-020-00075-2.

[230] S. Nusinovici, Y.C. Tham, M.Y.C. Yan, D.S.W. Ting, J. Li, C. Sabanayagam, T.Y. Wong, C.-Y. Cheng, Logistic regression was as good as machine learning for predicting major chronic diseases, Journal of Clinical Epidemiology. 122 (2020) 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002.

[231] R. Ratnakumaran, V. Hinder, D. Brand, J. Staffurth, E. Hall, N. van As, A. Tree, The Association between Acute and Late Genitourinary and Gastrointestinal Toxicities: An Analysis of the PACE B Study, Cancers. 15 (2023) 1288. https://doi.org/10.3390/cancers15041288.

[232] C. Huggins, C.V. Hodges, Studies on Prostatic Cancer. I. The Effect of Castration, of Estrogen and of Androgen Injection on Serum Phosphatases in Metastatic Carcinoma of the Prostate*, Cancer Research. 1 (1941) 293–297.
[233] J. Kim, G.A. Coetzee, Prostate specific antigen gene regulation by androgen receptor, Journal of Cellular Biochemistry. 93 (2004) 233–241. https://doi.org/10.1002/jcb.20228.
[234] R.T. Prehn, On the prevention and therapy of prostate cancer by androgen administration, Cancer Res. 59 (1999) 4161–4164.
[235] M. Ferro, G. Lucarelli, O. de Cobelli, M.D. Vartolomei, R. Damiano, F. Cantiello, F. Crocerossa, S. Perdonà, P. Del Prete, G. Cordima, G. Musi, F. Del Giudice, G.M. Busetto, B.I. Chung, A. Porreca, P. Ditonno, M. Battaglia, D. Terracciano, Circulating preoperative testosterone level predicts unfavourable disease at radical prostatectomy in men with International Society of Urological Pathology Grade Group 1 prostate cancer diagnosed with systematic biopsies, World J Urol. 39 (2021) 1861–1867. https://doi.org/10.1007/s00345-020-03368-9.
[236] M.A. Røder, I.J. Christensen, K.D. Berg, L. Gruschy, K. Brasso, P. Iversen, Serum testosterone level as a predictor of biochemical failure after radical prostatectomy for localized prostate cancer, BJU International. 109 (2012) 520–524. https://doi.org/10.1111/j.1464-410X.2011.10335.x.
[237] J.M. Martin, D.M. Sopka, K.J. Ruth, M.K. Buyyounouski, A. Kutikov, M.L. Sobczak, D.Y. Chen, E.M. Horwitz, Do Testosterone Kinetics After Radiation Therapy (RT) Predict Biochemical Failure (BCF) for Low- and Intermediate-risk Prostate Cancer (CaP)?, International Journal of Radiation Oncology, Biology, Physics. 84 (2012) S185–S186. https://doi.org/10.1016/j.ijrobp.2012.07.479.
[238] R.S. Pompe, P.I. Karakiewicz, E. Zaffuto, A. Smith, M. Bandini, M. Marchioni, Z. Tian, S.-R. Leyh-Bannurah, J. Schiffmann, G. Delouya, C. Lambert, J.-P. Bahary, M.C. Beauchemin, M. Barkati, C. Ménard, M. Graefen, F. Saad, D. Tilki, D. Taussky, External Beam Radiotherapy Affects Serum Testosterone in Patients with Localized Prostate Cancer, The Journal of Sexual Medicine. 14 (2017) 876–882. https://doi.org/10.1016/j.jsxm.2017.04.675.
[239] F. Akyol, G. Ozyigit, U. Selek, E. Karabulut, PSA Bouncing after Short Term Androgen Deprivation and 3D-Conformal Radiotherapy for Localized Prostate Adenocarcinoma and the Relationship with the Kinetics of Testosterone, European Urology. 48 (2005) 40–45. https://doi.org/10.1016/j.eururo.2005.04.007.
[240] J.M. Crook, C.J. O'Callaghan, G. Duncan, D.P. Dearnaley, C.S. Higano, E.M. Horwitz, E. Frymire, S. Malone, J. Chin, A. Nabid, P. Warde, T. Corbett, S. Angyalfi, S.L. Goldenberg, M.K. Gospodarowicz, F. Saad, J.P. Logue, E. Hall, P.F. Schellhammer, K. Ding, L. Klotz, Intermittent Androgen Suppression for Rising PSA Level after Radiotherapy, New England Journal of Medicine. 367 (2012) 895–903. https://doi.org/10.1056/NEJMoa1201546.
[241] R. Ferraldeschi, J. Welti, J. Luo, G. Attard, J.S. de Bono, Targeting the androgen receptor pathway in castration-resistant prostate cancer: progresses and prospects, Oncogene. 34 (2015) 1745–1757. https://doi.org/10.1038/onc.2014.115.
[242] O. Sartor, J.S. de Bono, Metastatic Prostate Cancer, New England Journal of Medicine. 378 (2018) 645–657. https://doi.org/10.1056/NEJMra1701695.
[243] A. Briganti, A. Larcher, F. Abdollah, U. Capitanio, A. Gallina, N. Suardi, M. Bianchi, M. Sun, M. Freschi, A. Salonia, P.I. Karakiewicz, P. Rigatti, F. Montorsi, Updated Nomogram Predicting Lymph Node Invasion in Patients with Prostate Cancer

Undergoing Extended Pelvic Lymph Node Dissection: The Essential Importance of Percentage of Positive Cores, European Urology. 61 (2012) 480–487. https://doi.org/10.1016/j.eururo.2011.10.044.

[244] L.L. Kestin, N.S. Goldstein, F.A. Vicini, A.A. Martinez, Percentage of positive biopsy cores as predictor of clinical outcome in prostate cancer treated with radiotherapy, J Urol. 168 (2002) 1994–1999. https://doi.org/10.1097/01.ju.0000033329.22922.b9.

[245] I.R. White, J.B. Carlin, Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values, Statistics in Medicine. 29 (2010) 2920–2931. https://doi.org/10.1002/sim.3944.

[246] N.S. Erler, D. Rizopoulos, E.M.E.H. Lesaffre, JointAI: Joint Analysis and Imputation of Incomplete Data in R, Journal of Statistical Software. 100 (2021) 1–56. https://doi.org/10.18637/jss.v100.i20.

[247] A. Wilkins, B. Gusterson, H. Tovey, C. Griffin, C. Stuttle, F. Daley, C.M. Corbishley, D. Dearnaley, E. Hall, N. Somaiah, Multi-candidate immunohistochemical markers to assess radiation response and prognosis in prostate cancer: results from the CHHiP trial of radiotherapy fractionation, EBioMedicine. 88 (2023) 104436. https://doi.org/10.1016/j.ebiom.2023.104436.

[248] A. Wilkins, C. Stuttle, S. Hassan, C. Blanchard, C. Cruickshank, C. Griffin, J. Probert, C.M. Corbishley, C. Parker, D. Dearnaley, E. Hall, Methodology for tissue sample collection within a translational sub-study of the CHHiP trial (CRUK/06/016), a large randomised phase III trial in localised prostate cancer, CtRO. 10 (2018) 1–6. https://doi.org/10.1016/j.ctro.2018.02.002.

[249] M.J.G. Bussemakers, A. van Bokhoven, G.W. Verhaegh, F.P. Smit, H.F.M. Karthaus, J.A. Schalken, F.M.J. Debruyne, N. Ru, W.B. Isaacs, DD3::A New Prostate-specific Gene, Highly Overexpressed in Prostate Cancer1, Cancer Research. 59 (1999) 5975–5979.

[250] C. Song, H. Chen, Predictive significance of TMRPSS2-ERG fusion in prostate cancer: a meta-analysis, Cancer Cell International. 18 (2018) 177. https://doi.org/10.1186/s12935-018-0672-2.

[251] D.S. O'Keefe, D.J. Bacich, S.S. Huang, W.D.W. Heston, A Perspective on the Evolving Story of PSMA Biology, PSMA-Based Imaging, and Endoradiotherapeutic Strategies, Journal of Nuclear Medicine. 59 (2018) 1007–1013. https://doi.org/10.2967/jnumed.117.203877.

[252] J.S. Ross, C.E. Sheehan, H.A.G. Fisher, R.P. Kaufman Jr., P. Kaur, K. Gray, I. Webb, G.S. Gray, R. Mosher, B.V.S. Kallakury, Correlation of Primary Tumor Prostate-Specific Membrane Antigen Expression with Disease Recurrence in Prostate Cancer, Clinical Cancer Research. 9 (2003) 6357–6362.

[253] N.I. Simon, C. Parker, T.A. Hope, C.J. Paller, Best Approaches and Updates for Prostate Cancer Biochemical Recurrence, American Society of Clinical Oncology Educational Book. (2022) 352–359. https://doi.org/10.1200/EDBK_351033.

[254] D.A. Ferraro, J.H. Rüschoff, U.J. Muehlematter, B. Kranzbühler, J. Müller, M. Messerli, L. Husmann, T. Hermanns, D. Eberli, N.J. Rupp, I.A. Burger, Immunohistochemical PSMA expression patterns of primary prostate cancer tissue are associated with the detection rate of biochemical recurrence with 68Ga-PSMA-11-PET, Theranostics. 10 (2020) 6082–6094. https://doi.org/10.7150/thno.44584.

[255] C. Lawhn-Heath, R.R. Flavell, S.C. Behr, T. Yohannan, K.L. Greene, F. Feng, P.R. Carroll, T.A. Hope, Single-Center Prospective Evaluation of 68Ga-PSMA-11 PET in

Biochemical Recurrence of Prostate Cancer, American Journal of Roentgenology. 213 (2019) 266–274. https://doi.org/10.2214/AJR.18.20699.

[256] N.G. Burnet, G.C. Barnett, H.R. Summersgill, A.M. Dunning, C.M.L. West, RAPPER — A Success Story for Collaborative Translational Radiotherapy Research, Clinical Oncology. 31 (2019) 416–419. https://doi.org/10.1016/j.clon.2019.04.013.

[257] H.D. Green, S.W.D. Merriel, R.A. Oram, K.S. Ruth, J. Tyrrell, S.E. Jones, C. Thirlwell, M.N. Weedon, S.E.R. Bailey, Applying a genetic risk score for prostate cancer to men with lower urinary tract symptoms in primary care to predict prostate cancer diagnosis: a cohort study in the UK Biobank, Br J Cancer. 127 (2022) 1534–1539. https://doi.org/10.1038/s41416-022-01918-z.

[258] E.J. Saunders, T. Dadaev, D.A. Leongamornlert, A.A.A. Olama, S. Benlloch, G.G. Giles, F. Wiklund, H. Grönberg, C.A. Haiman, J. Schleutker, B.G. Nordestgaard, R.C. Travis, D. Neal, N. Pasayan, K.-T. Khaw, J.L. Stanford, W.J. Blot, S.N. Thibodeau, C. Maier, A.S. Kibel, C. Cybulski, L. Cannon-Albright, H. Brenner, J.Y. Park, R. Kaneva, J. Batra, M.R. Teixeira, H. Pandha, K. Govindasami, K. Muir, D.F. Easton, R.A. Eeles, Z. Kote-Jarai, Gene and pathway level analyses of germline DNA-repair gene variants and prostate cancer susceptibility using the iCOGS-genotyping array, Br J Cancer. 114 (2016) 945–952. https://doi.org/10.1038/bjc.2016.50.

[259] J. Morote, J. del Amo, A. Borque, E. Ars, C. Hernández, F. Herranz, A. Arruza, R. Llarena, J. Planas, M.J. Viso, J. Palou, C.X. Raventós, D. Tejedor, M. Artieda, L. Simón, A. Martínez, L.A. Rioja, Improved Prediction of Biochemical Recurrence After Radical Prostatectomy by Genetic Polymorphisms, The Journal of Urology. 184 (2010) 506–511. https://doi.org/10.1016/j.juro.2010.03.144.

[260] C. Zanusso, R. Bortolus, E. Dreussi, J. Polesel, M. Montico, E. Cecchin, S. Gagno, F. Rizzolio, M. Arcicasa, G. Novara, G. Toffoli, Impact of DNA repair gene polymorphisms on the risk of biochemical recurrence after radiotherapy and overall survival in prostate cancer, Oncotarget. 8 (2017) 22863–22875. https://doi.org/10.18632/oncotarget.15282.

[261] Medical Device Coordination Group, MDCG 2019-11 Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR, (2019). https://health.ec.europa.eu/system/files/2020-09/md_mdcg_2019_11_guidance_qualification_classification_software_en_0.pdf (accessed February 6, 2023).

[262] ISO/TC 210, ISO 13485:2016, ISO. (2021). https://www.iso.org/standard/59752.html (accessed February 6, 2023).

[263] J. Ordish, Large Language Models and software as a medical device - MedRegs, (2023). https://medregs.blog.gov.uk/2023/03/03/large-language-models-and-software-as-a-medical-device/ (accessed March 10, 2023).

[264] MHRA, Crafting an intended purpose in the context of software as a medical device (SaMD), GOV.UK. (2023). https://www.gov.uk/government/publications/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd (accessed March 23, 2023).

[265] Horizon: Making Sense of Cancer with Hannah Fry, 2022. https://www.bbc.co.uk/iplayer/episode/m0017wzq/horizon-2022-making-sense-of-cancer-with-hannah-fry (accessed February 14, 2023).

[266] W. en S. Ministerie van Volksgezondheid, Guideline for high-quality diagnostic and prognostic applications of AI in healthcare - Publicatie - Data voor gezondheid, (2021). https://www.datavoorgezondheid.nl/documenten/publicaties/2021/12/17/guideline-for-high-quality-diagnostic-and-prognostic-applications-of-ai-in-healthcare (accessed March 10, 2023).

[267] H. Uno, T. Cai, L. Tian, L.J. Wei, Evaluating Prediction Rules for t-Year Survivors with Censored Regression Models, Journal of the American Statistical Association. 102 (2007) 527–537. https://www.jstor.org/stable/27639883 (accessed March 20, 2021).

[268] J. Ensor, E.C. Martin, R.D. Riley, pmsampsize: Calculates the Minimum Sample Size Required for Developing a Multivariable Prediction Model, (2022). https://CRAN.R-project.org/package=pmsampsize (accessed February 17, 2023).

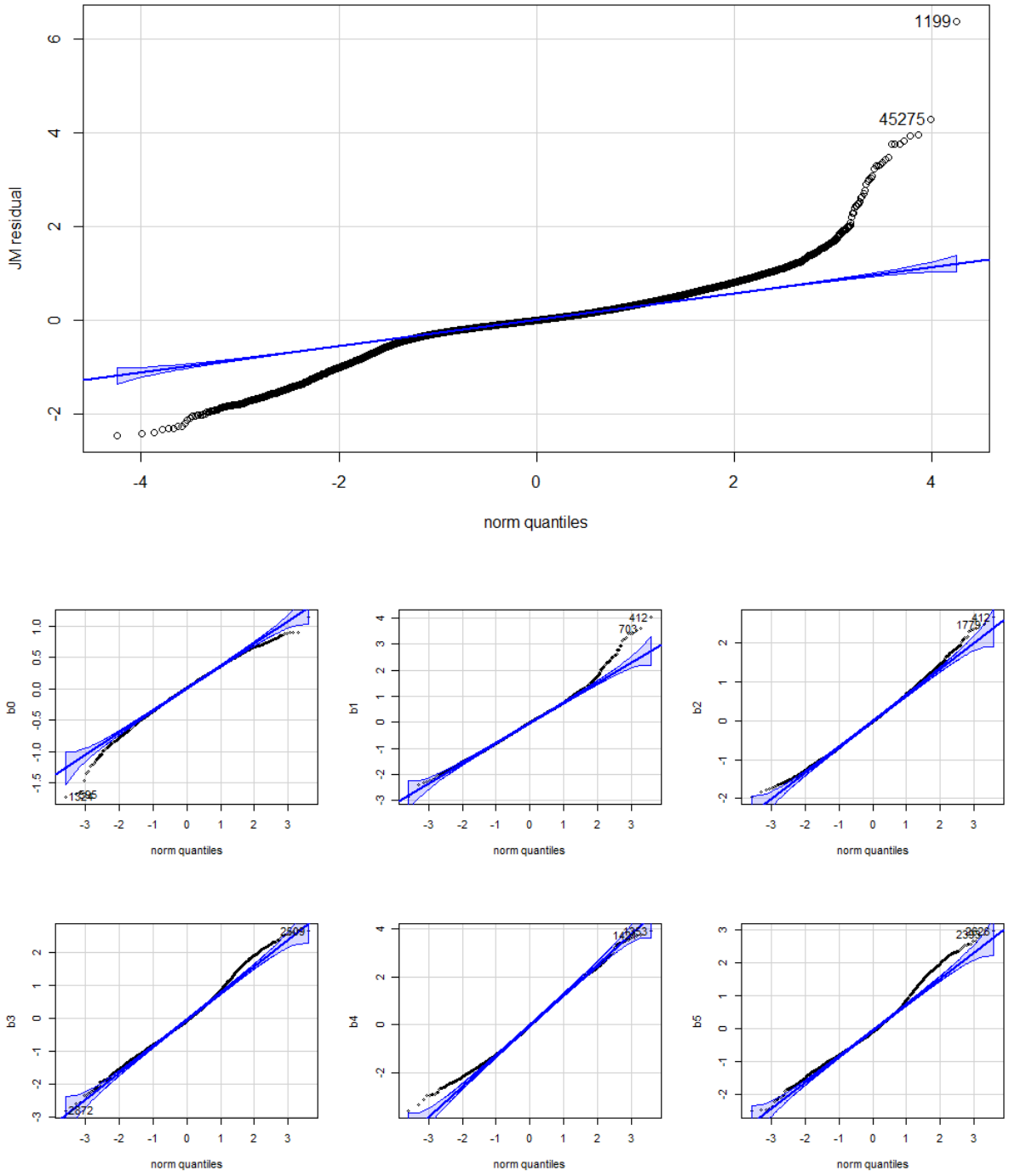
Appendices

Appendix A (chapter 4)

*Supplementary Table A1 TRIPOD checklist for Chapter 4. *Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D; V. TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis.*

Section/Topic	Items*		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	74
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	NA
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	74-75 & Ch2
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	74-75
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	75-76
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	29, 76
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	NA
	5b	D;V	Describe eligibility criteria for participants.	75
	5c	D;V	Give details of treatments received, if relevant.	75
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	7
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	7
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	81
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	76
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	77-79
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	76-79
	10c	V	For validation, describe how the predictions were calculated.	80-81
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	80-81
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	NA
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	NA
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	30, 81
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	82
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	81-82
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	86-88
	15b	D	Explain how to the use the prediction model.	89,93-94
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	92-93
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	96,98-99
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	100-101
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	Discussion
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	100-103
Other Information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	Appendix A
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	NA

Diagnostics of the joint model longitudinal residuals and random effects are shown in ***Supplementary Figure A1***.



Supplementary Figure A1 Quantile-Quantile plots of the residuals from the longitudinal joint model (top panel) and random effects (bottom panel).

Apparent validation metrics of both discrimination and calibration components of the developed joint model in **Chapter 4** – Development of a Personalised Clinical Dynamic Predictive Joint Model to Characterise Prognosis for Patients with Localised Prostate Cancer Patients: Analysis of the CHHiP Phase III Trialare assessed using the framework described in [85,90,267]. Similarly, the Brier score is used as an overall measure of predictive accuracy [88,99]. Calibration accuracy measures given by Austin et al. [91], specifically the ICI, i.e., the weighted difference between the predicted probabilities and smoothed observed proportions, are considered. These were extracted using the *tvAUC*, *calibration_metrics*, *tvBrier* functions from the ***JMbayes2*** package, respectively. These are shown in ***Supplementary Table A2***.

Supplementary Table A2 development apparent/internal prediction accuracy metrics at varying landmark times either with a fixed horizon time of t=8 years (first panel), or a fixed prediction window Dt=2 or Dt=5 years (second and third pane respectively). Time-dependent metrics: AUROC, Brier, ICI. Ns indicate the number of patients still at risk by the landmark time t.

Ns	[t, u]	AUC	ICI	Brier
3071	[0, 8]	0.543	0.046	0.16
3039	[1, 8]	0.579	0.067	0.156
2947	[2, 8]	0.615	0.072	0.153
2823	[3, 8]	0.651	0.063	0.123
2705	[4, 8]	0.749	0.047	0.098
2528	[5, 8]	0.803	0.038	0.069
2357	[6, 8]	0.843	0.023	0.047
2176	[7, 8]	0.812	0.016	0.027
3071	[0, 2]	0.58	0.01	0.02
3039	[1, 3]	0.621	0.029	0.045
2947	[2, 4]	0.705	0.056	0.059
2823	[3, 5]	0.704	0.047	0.063
2705	[4, 6]	0.785	0.04	0.058
2528	[5, 7]	0.824	0.032	0.048
2357	[6, 8]	*See above*		
2176	[7, 9]	0.808	0.024	0.051
1846	[8, 10]	0.811	0.021	0.043
3071	[0, 5]	0.549	0.031	0.103
3039	[1, 6]	0.593	0.059	0.119
2947	[2, 7]	0.624	0.072	0.134
2823	[3, 8]	*See above*		
2705	[4, 9]	0.728	0.046	0.118
2528	[5, 10]	0.776	0.044	0.103

In addition to the fixed landmark approach taken in the main manuscript, alternative clinically relevant fixed prediction windows of two and five years are considered here. For either prediction window, discrimination increases, better distinguishing between patients who do and do not have recurrence. These are optimal after five-years' worth of PSA (AUC: $[t = 5, u = 7] = 0.82$, $[t = 5, u = 10] = 0.78$). The Brier scores show a generally expected decrease in prediction error given more longitudinal information, showing a moderately low prediction error by five years. The low Brier score at baseline $[t = 0, u = 2]$ is explained by the very homogenous decrease in PSA within the first two years of treatment (seen in **Figure 4-2**), and there are very few events within the first two years of treatment. The ICIs are reported. The model generally become better calibrated in the latter years, with an initially low ICI that increases in the first two years, then decreases i.e., becoming better calibrated in the latter years. External validation can be used to further correct calibration-in-the-large [93].

A minimum sample size calculation is performed using the methodology by Riley et al. [174]. There are several parameters required to inform of the minimum sample size, which are: the number of candidate predictor parameters of interest p, the corresponding Cox-Snell statistic $R^2_{CS_{\text{adjusted}}}$, Van Houwelingen's global shrinkage factor $S_{VH} \geq 0.9$, the time point of interest and the mean follow-up time, and event rate (the number of events per person-year). This calculation is done using the **R** package and function *pmsampsize* [268].

- $p = 11$, this is made up of 9 baseline prognostic factor levels and the two association structure parameters (value + slope).
- $R^2_{CS_{\text{adjusted}}} = R^2_{CS_{\text{apparent}}} \times S_{VH} = 0.038 \times 0.9 = 0.034$
- rate$= \frac{\text{events}}{\Sigma \text{ person-years}} = \frac{607}{24103.92} = 0.025$
- the time point of interest is a horizon time of 8 years
- the mean follow-up is 7.85 years.

This gives an output of

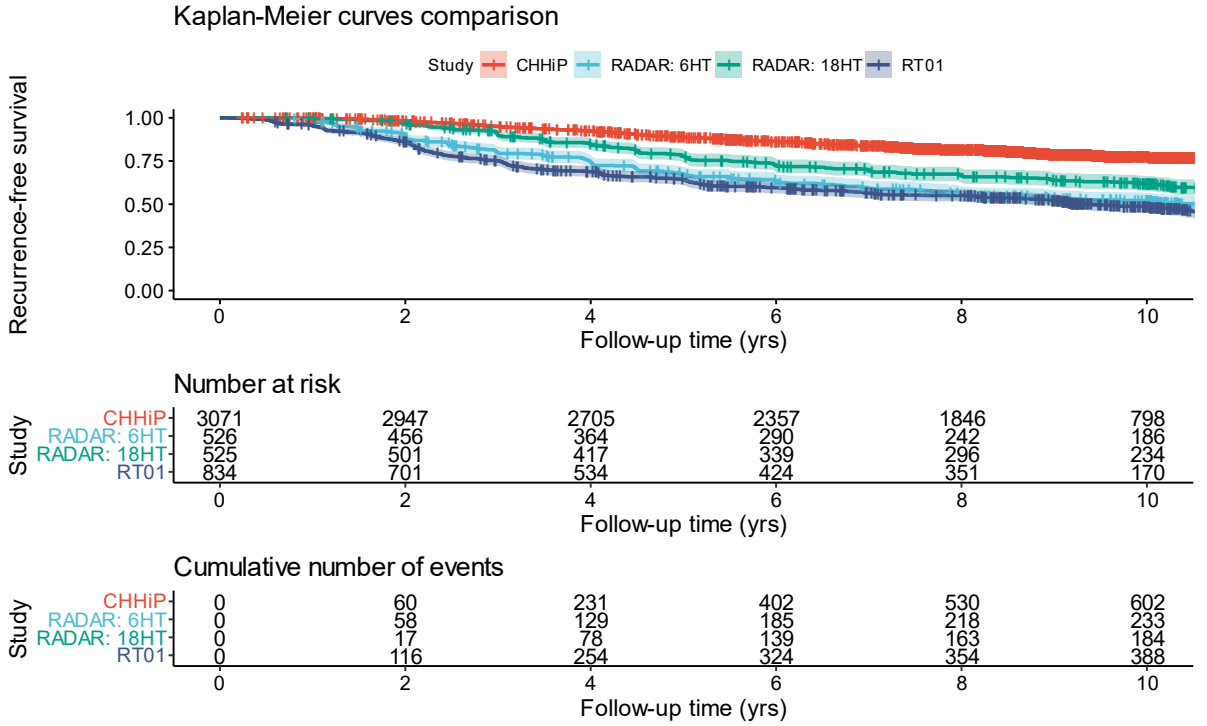
```
pmsampsize(type = "s", parameters = 11, rate = 0.02518262, rsquared =
0.03434196, timepoint = 8, meanfup = 7.8489)
NB: Assuming 0.05 acceptable difference in apparent & adjusted R-squared
NB: Assuming 0.05 margin of error in estimation of overall risk at time
point = 8
NB: Events per Predictor Parameter (EPP) assumes overall event rate =
0.02518262

             Samp_size Shrinkage Parameter     CS_Rsq Max_Rsq Nag_Rsq   EPP
Criteria 1        2828     0.900        11 0.03434196   0.645   0.053 50.82
Criteria 2         330     0.516        11 0.03434196   0.645   0.053  5.93
Criteria 3 *      2828     0.900        11 0.03434196   0.645   0.053 50.82
Final SS          2828     0.900        11 0.03434196   0.645   0.053 50.82

Minimum sample size required for new model development based on user inputs
= 2828, corresponding to 22196.7 person-time** of follow-up, with 559 outcome
events assuming an overall event rate = 0.02518262 and therefore an EPP =
50.82

 * 95% CI for overall risk = (0.169, 0.196), for true value of 0.182 and
sample size n = 2828
 **where time is in the units mean follow-up time was specified in
```

i.e., a sample size of 2828 with 559 events (events-per-parameter = 51).

Appendix B (chapter 5)

*Supplementary Table B1 TRIPOD checklist for Chapter 5. *Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D; V. TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis.*

Section/Topic	Items*		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	104
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	NA
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	104-105
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	104-105
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	105-106
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	105-106
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	105-106
	5b	D;V	Describe eligibility criteria for participants.	105-106
	5c	D;V	Give details of treatments received, if relevant.	105-106
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	106-107
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	107
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	109
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	108-109
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	109-110
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	107-108
	10c	V	For validation, describe how the predictions were calculated.	107-108
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	107-108
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	107-108
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	NA
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	109-110
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	106
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	1110
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	110-112
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	110-112
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	NA
	15b	D	Explain how to the use the prediction model.	Chapter 4
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	113-119
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	115-119
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	121-126
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	121-126
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	121-126
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	121-126
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	Appendix B
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	NA



*Supplementary Figure B1 Kaplan-Meier (similar to **Figure 5-2**) with a breakdown comparing outcomes of the RADAR 6- and 18-month hormone schedules.*

Supplementary Figure B1 demonstrates the improved outcomes for RADAR patients who receive the longer 18-month hormone schedule, compared to RT01 and the RADAR 6-month schedule.

Supplementary Table B2 assessing overall calibration using the integrated calibration index (ICI) of the external cohorts before and after recalibration. The calibration metrics at landmarks $t = 0, \ldots, 7$ to predict by a fixed horizon time of 8 years (first panel), and with varying time horizons, i.e., fixed prediction intervals of two and five years are presented in the second and third panels respectively (continued next page). The negative percentage difference indicates improvement in ICI after recalibration.

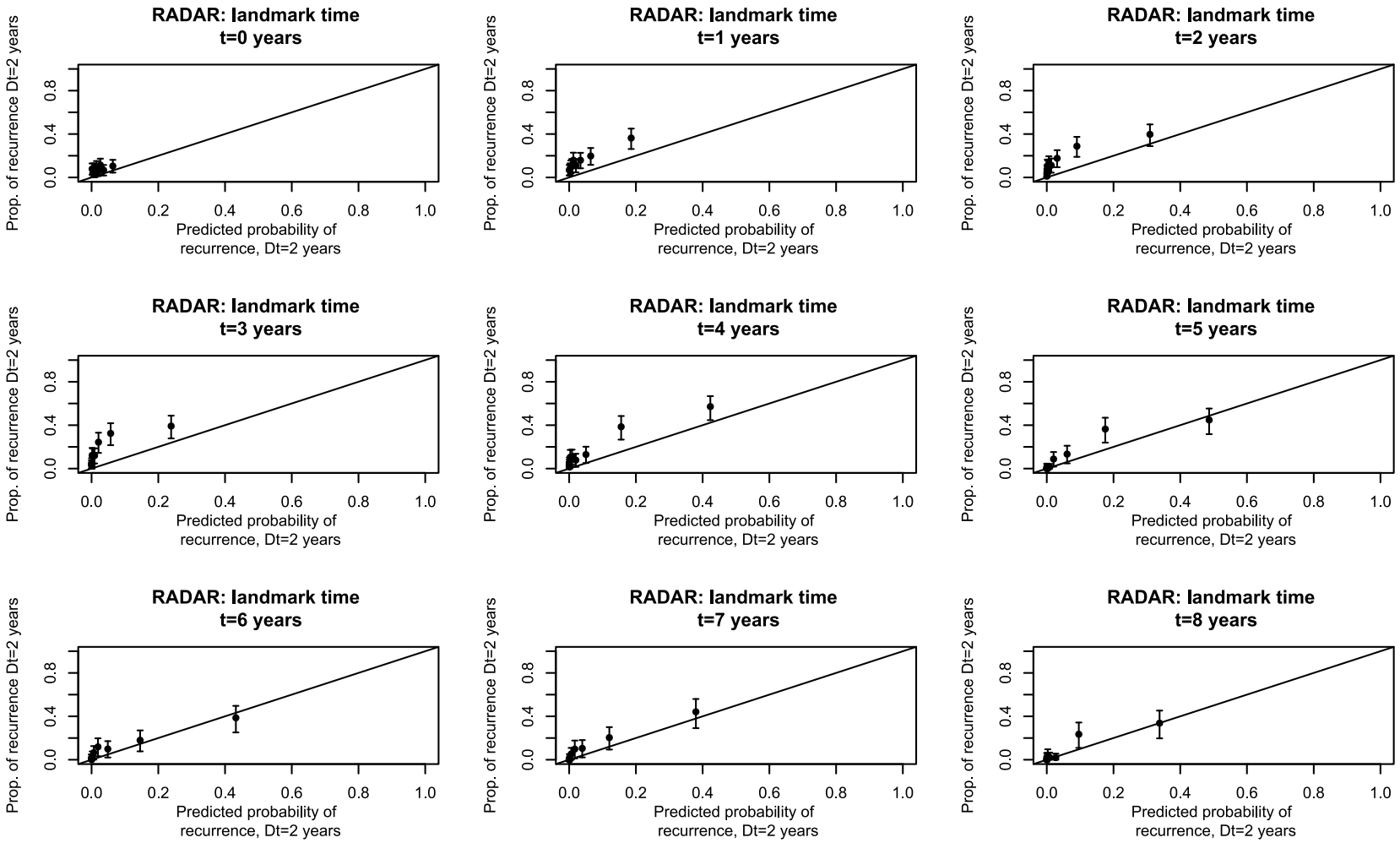
Trial [prediction interval]	N at risk	Prediction window	ICI (original)	ICI (recalibrated)	% Difference
RADAR [t, 8]	1051	[0, 8]	0.123	0.084	-32%
	1035	[1, 8]	0.216	0.181	-16%
	957	[2, 8]	0.205	0.17	-17%
	852	[3, 8]	0.181	0.151	-17%
	781	[4, 8]	0.127	0.097	-24%
	697	[5, 8]	0.072	0.061	-15%
	629	[6, 8]	0.047	0.051	9%
	587	[7, 8]	0.029	0.028	-3%
RT01 [t, 8]	815	[0, 8]	0.219	0.079	-64%
	788	[1, 8]	0.234	0.133	-43%
	701	[2, 8]	0.168	0.097	-42%
	600	[3, 8]	0.137	0.119	-13%
	534	[4, 8]	0.086	0.067	-22%
	481	[5, 8]	0.062	0.047	-24%
	424	[6, 8]	0.05	0.033	-34%
	378	[7, 8]	0.027	0.016	-41%

Trial [prediction interval]	N at risk	Prediction window	ICI (original)	ICI (recalibrated)	% Difference
	1051	**[0, 2]**	**0.054**	**0.046**	**-15%**
	1035	[1, 3]	0.113	0.103	-9%
	957	[2, 4]	0.108	0.104	-4%
	852	[3, 5]	0.117	0.115	-2%
RADAR [t, t+2]	781	[4, 6]	0.102	0.091	-11%
	697	[5, 7]	0.06	0.058	-3%
	629	[6, 8]		See above	
	587	[7, 9]	0.041	0.036	-12%
	538	[8, 10]	0.028	0.024	-14%
	815	[0, 2]	0.129	0.119	-8%
	788	[1, 3]	0.172	0.13	-24%
	701	[2, 4]	0.141	0.076	-46%
	600	[3, 5]	0.1	0.096	-4%
RT01 [t, t+2]	534	[4, 6]	0.079	0.069	-13%
	481	[5, 7]	0.063	0.046	-27%
	424	[6, 8]		See above	
	378	[7, 9]	0.048	0.031	-35%
	351	[8, 10]	0.069	0.073	6%
	1051	[0, 5]	0.129	0.115	-11%
	1035	[1, 6]	0.206	0.188	-9%
RADAR [t, t+5]	957	[2, 7]	0.196	0.166	-15%
	852	[3, 8]		See above	
	781	[4, 9]	0.128	0.099	-23%
	697	[5, 10]	0.069	0.056	-19%
	815	[0, 5]	0.241	0.076	-68%
	788	[1, 6]	0.245	0.137	-44%
RT01 [t, t+5]	701	[2, 7]	0.182	0.098	-46%
	600	[3, 8]		See above	
	534	[4, 9]	0.091	0.079	-13%
	481	[5, 10]	0.084	0.072	-14%

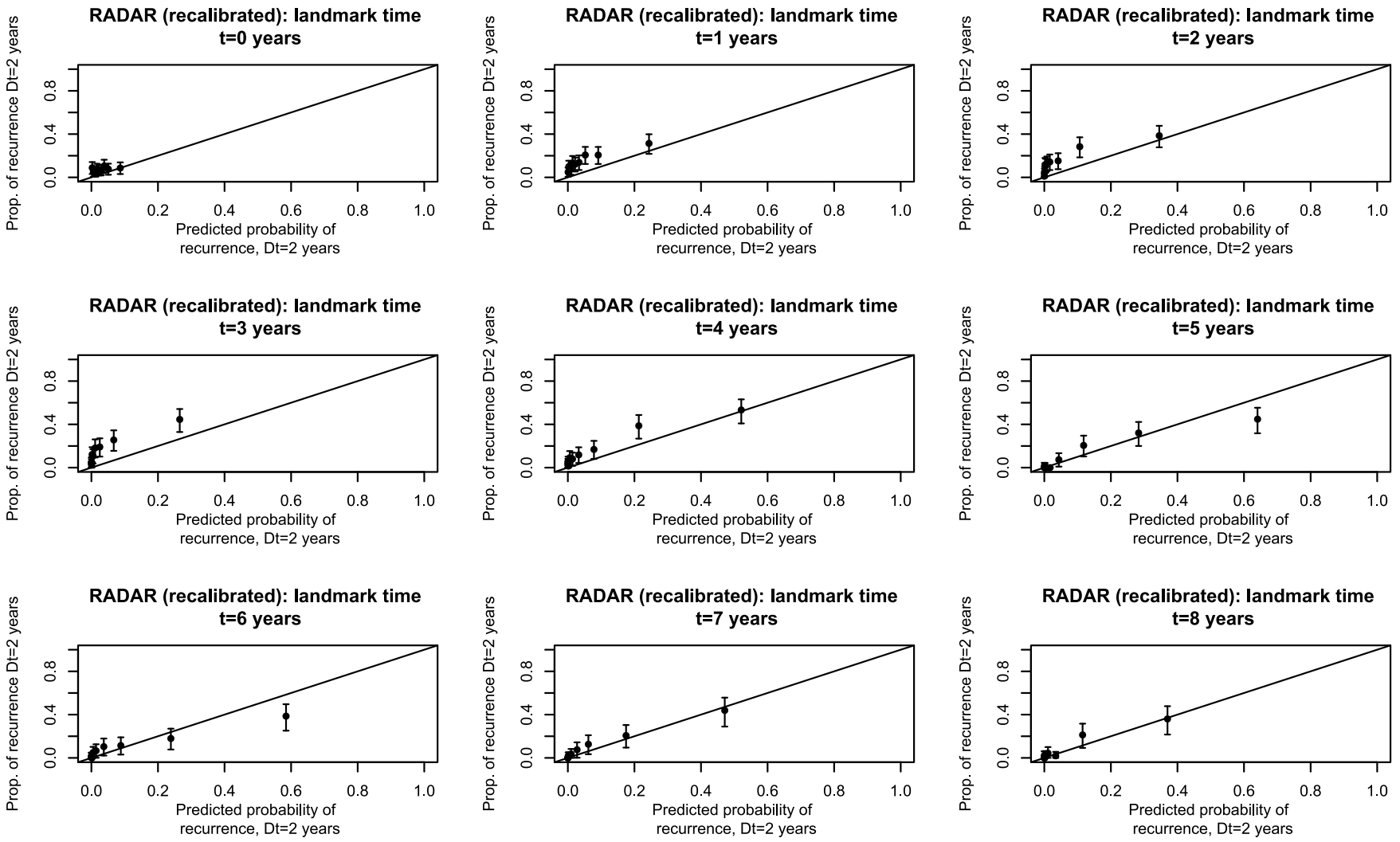
For a fixed horizon time of $u = 8$ years (***Supplementary Table B2***, first panel), there are systematic improvements by recalibrating the baseline hazard of the reduced CDPJM for the external cohorts, seen in the reduction of the ICIs and the negative percentage differences (other than for RADAR $[t = 6, u = 8]$). The recalibrated ICIs are similar for both cohorts for $t = 0$, with a sharp increase for $t = 1$, then ICIs generally decrease with more accrued longitudinal information, as expected given the horizon prediction time ($u = 8$) is closer to the landmark time t. RT01 has lower overall ICIs than RADAR, and an overall bigger improvement via recalibration.

For the fixed prediction windows (with varying landmark t times) of two years ($u = t + 2$, ***Supplementary Table B2*** second panel) there is a bigger disparity in the recalibrated ICIs at $t = 0$, with RT01 having a 2.6x increase in error predicting recurrence free rates at 2 years compared to RADAR (ICI 0.046 vs 0.12). The ICIs for both cohorts start to decrease after landmark $t = 3$, with RT01 producing lower ICI errors than RADAR from $t = 2$, except for $t = 8$ (RT01 ICI 0.07 *vs* RADAR ICI 0.02). Recalibration leads to larger improvements in the RT01 cohort, again. Overall, the ICIs for $\pi_i(t + 2|t)$ are better than $\pi_i(8|t)$, expected given the shorter horizon time. In these lines, ICIs for $\pi_i(t + 5|t)$ (***Supplementary Table B2***, third panel) are more similar to the performance for $\pi_i(8|t)$. Note however, how the ICIs for RT01 under the three scenarios ($\pi_i(8|t), \pi_i(t + 2|t)$, and $\pi_i(t + 5|t)$) are more closely aligned compared to RADAR.

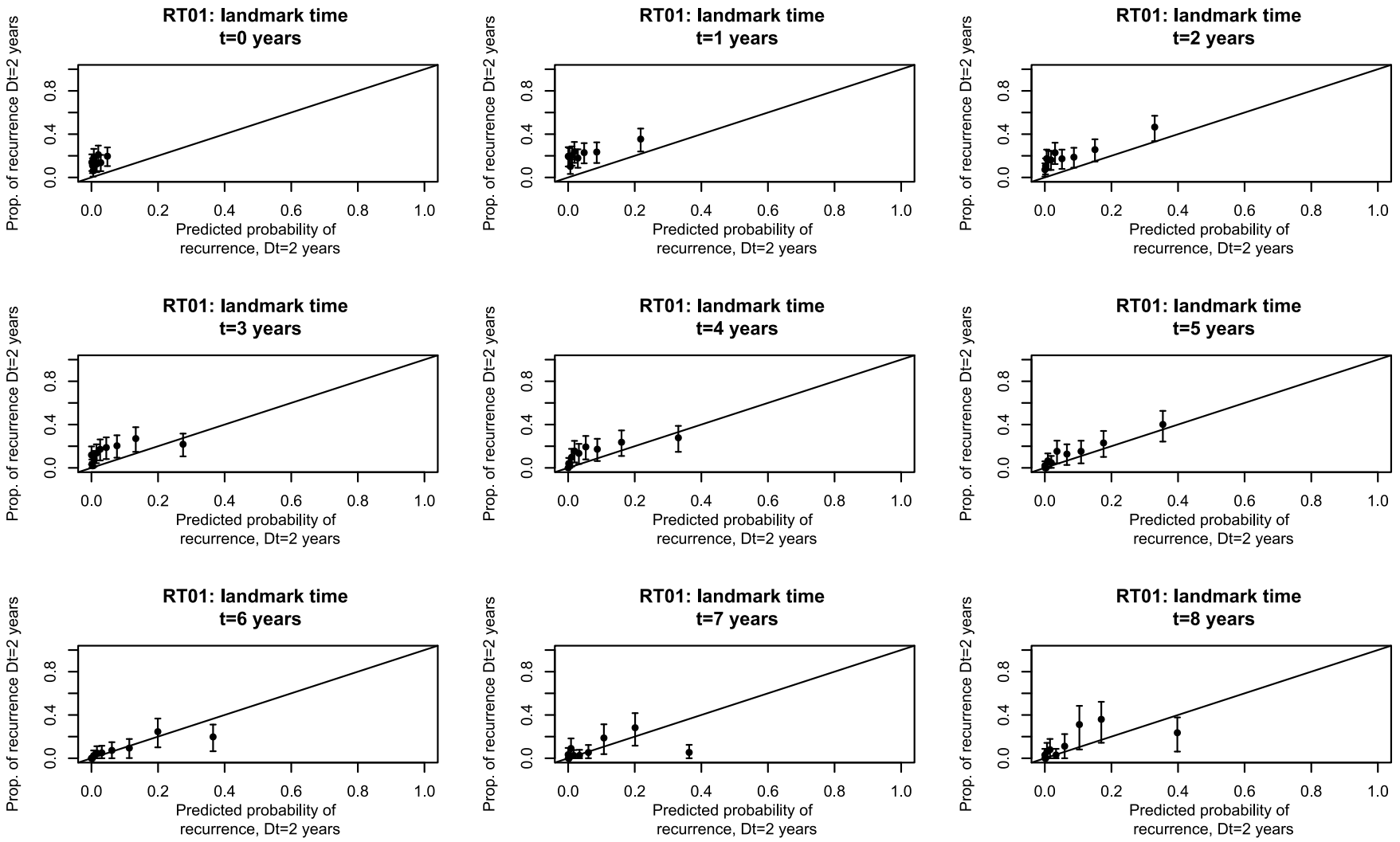
The calibration and recalibration plots for RADAR $\pi_i(t + 2|t)$ are shown in ***Supplementary Figure B2*** & ***Supplementary Figure B3***, respectively. Similarly, for RT01 in ***Supplementary Figure B4*** & ***Supplementary Figure B5***. The calibration/recalibration plots $\pi_i(t + 5|t)$ for RADAR are shown in ***Supplementary Figure B6*** & ***Supplementary Figure B7***, respectively; RT01 in ***Supplementary Figure B8*** & ***Supplementary Figure B9*** respectively. These are described in **Chapter 5.3.2**.



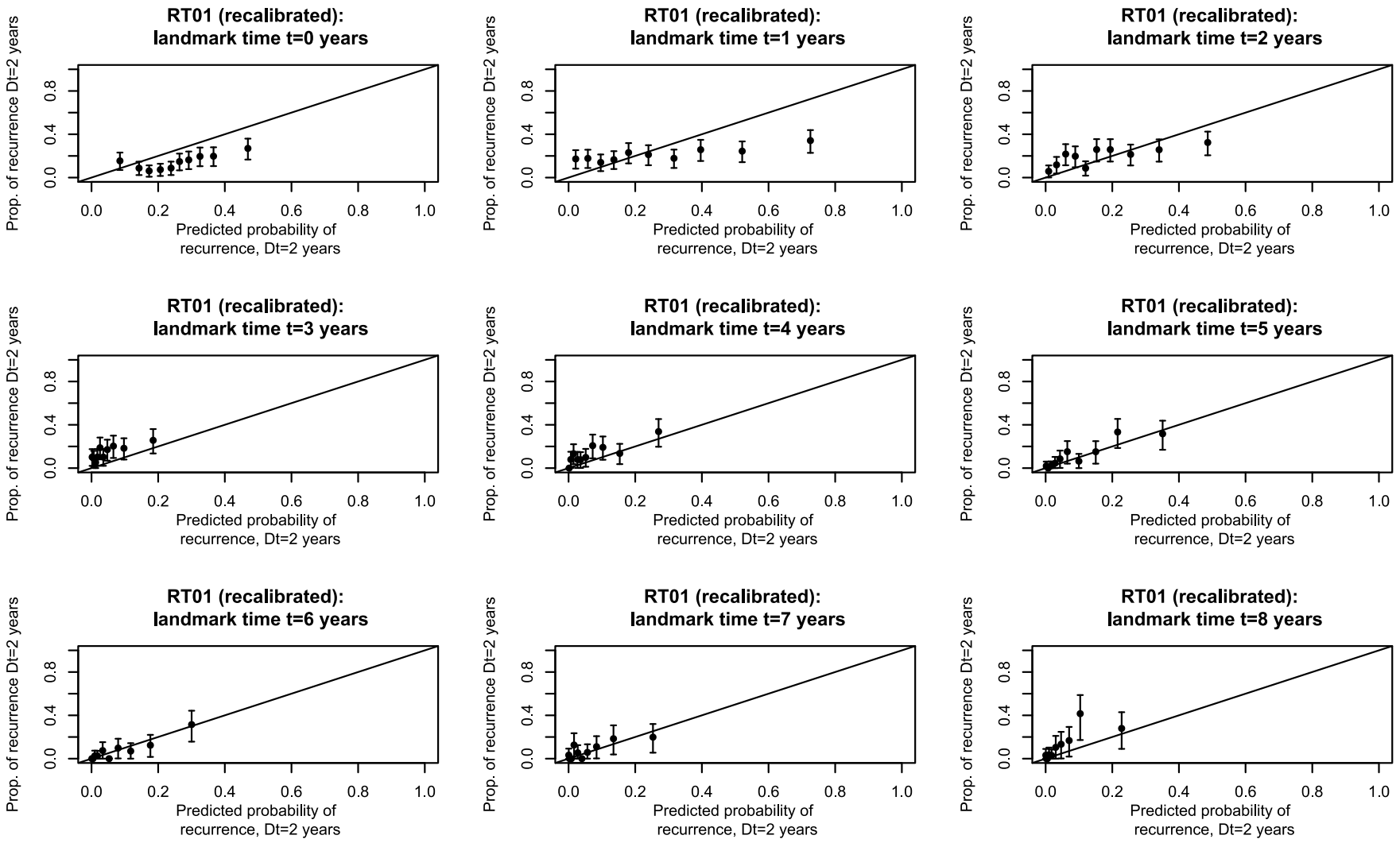
Supplementary Figure B2 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RADAR cohort, before recalibration, for a fixed prediction window of two years.



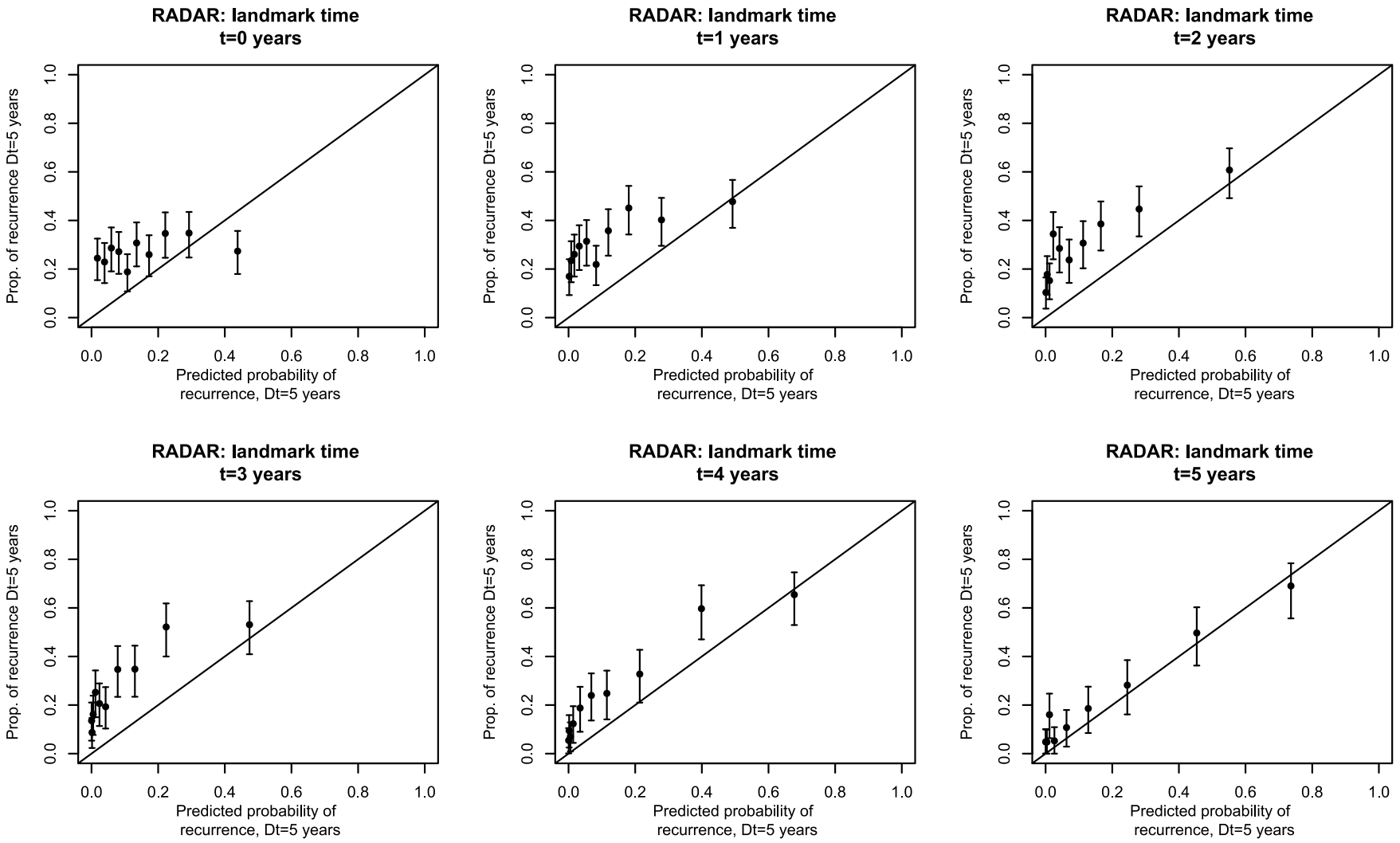
Supplementary Figure B3 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RADAR cohort, after recalibration, for a fixed prediction window of two years.



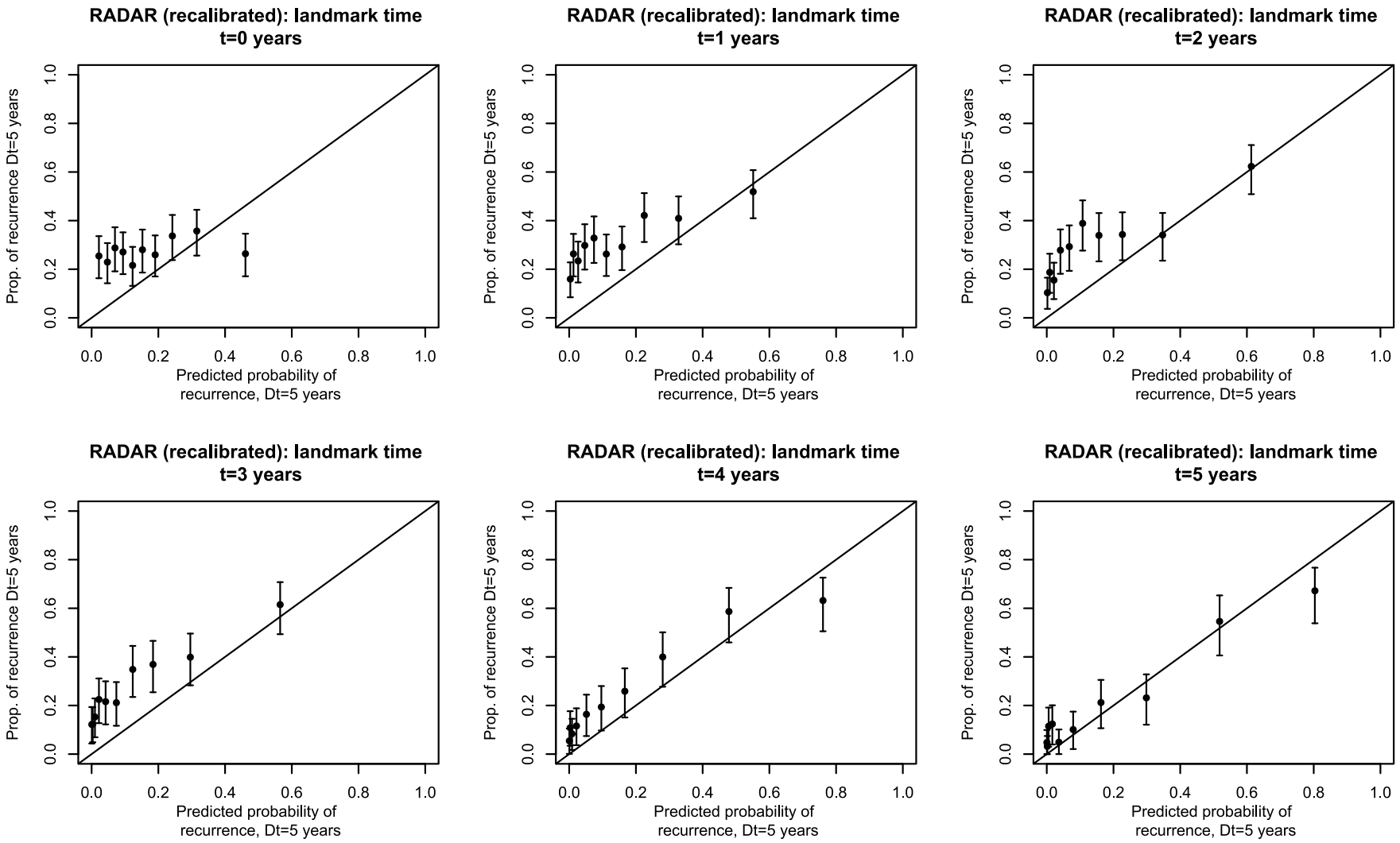
Supplementary Figure B4 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RT01 cohort, before recalibration, for a fixed prediction window of two years.



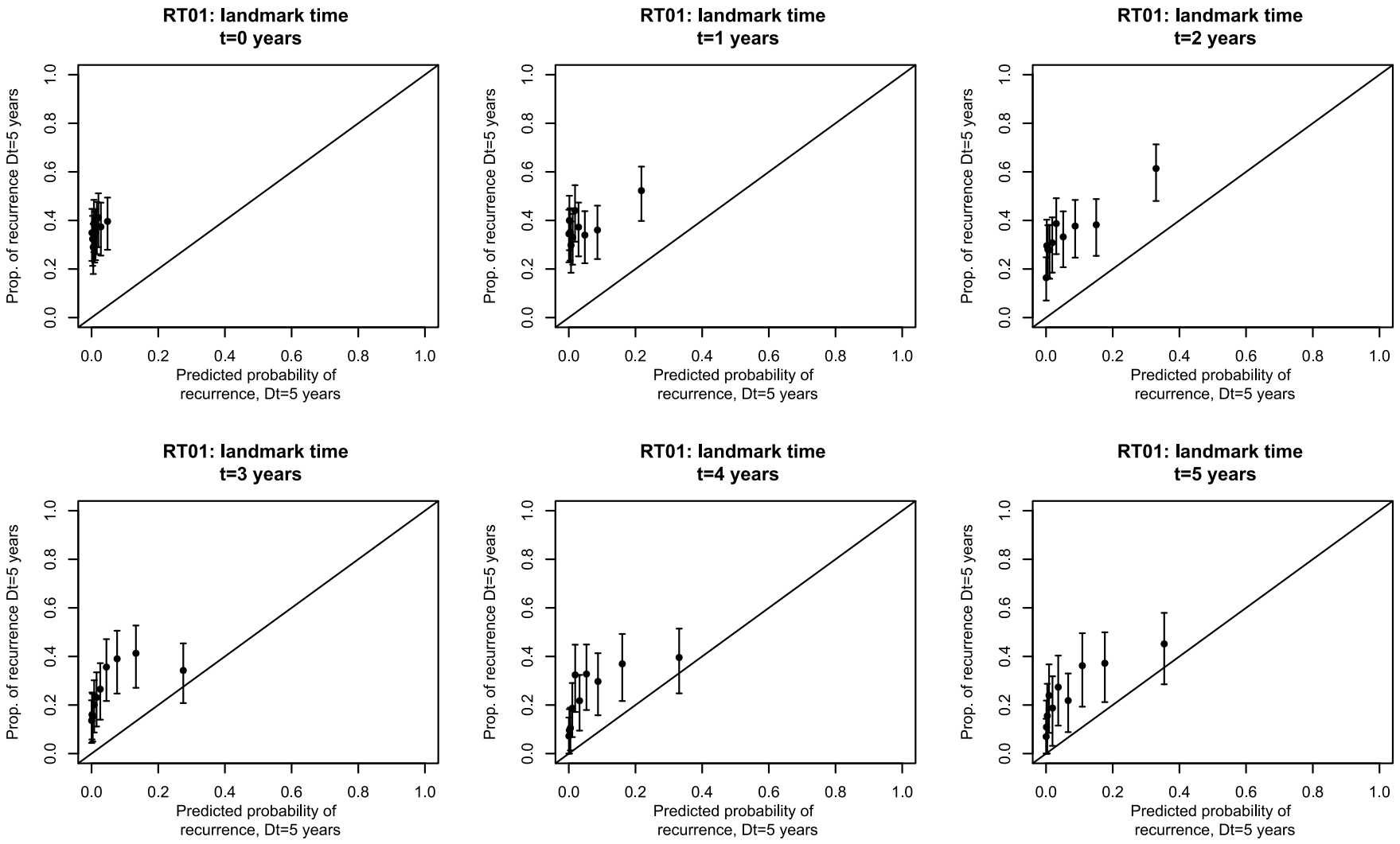
Supplementary Figure B5 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RT01 cohort, after recalibration, for a fixed prediction window of two years.



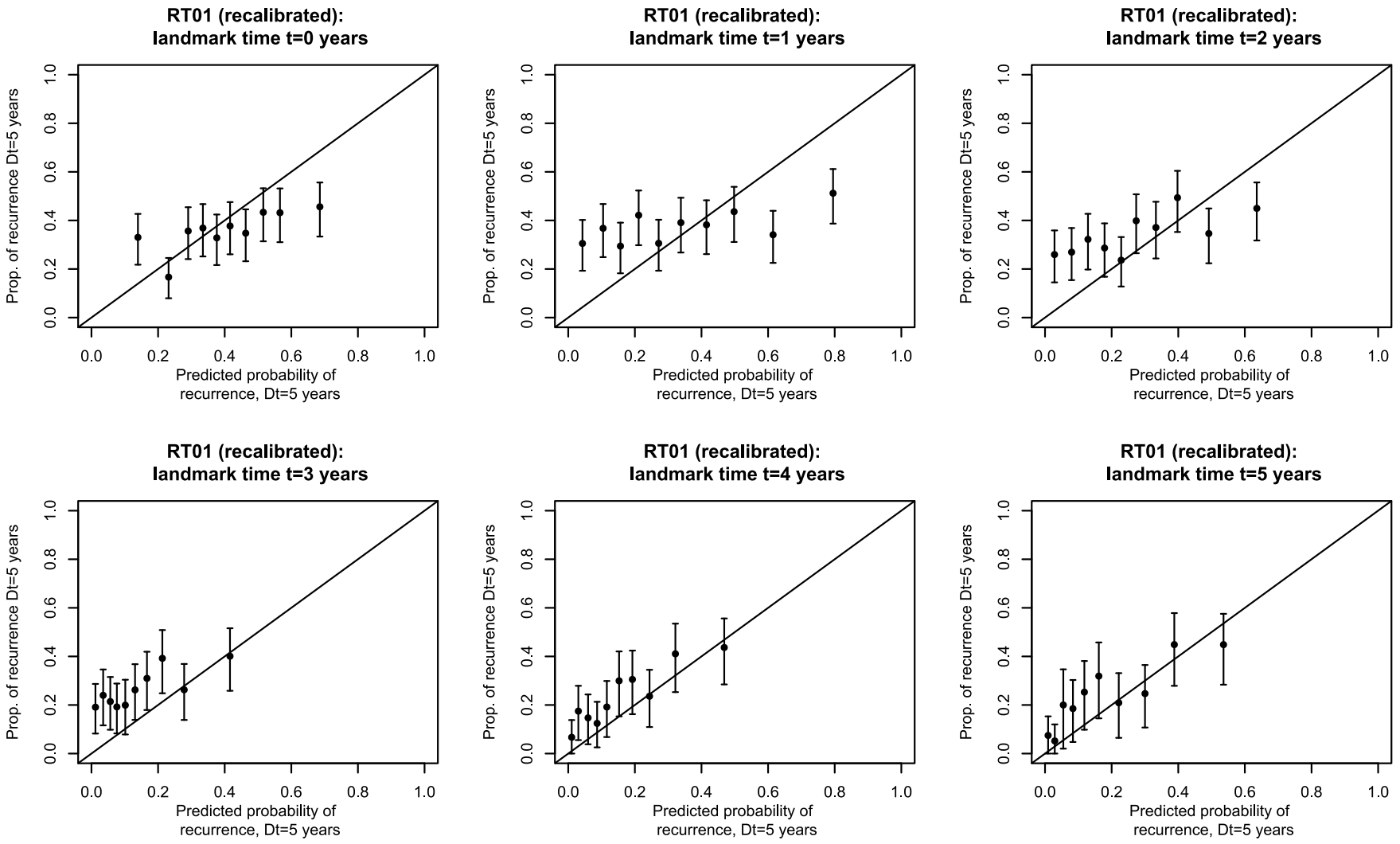
Supplementary Figure B6 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RADAR cohort, before recalibration, for a fixed prediction window of five years.



Supplementary Figure B7 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RADAR cohort, after recalibration, for a fixed prediction window of five years.



Supplementary Figure B8 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RT01 cohort, before recalibration, for a fixed prediction window of five years.



Supplementary Figure B9 visually assessing calibration-in-the-large via graphical smoothed calibration plots of the RT01 cohort, after recalibration, for a fixed prediction window of five years.

A simulation study – minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome for CHHiP using the framework developed by: Riley RD, Collins GS, Ensor J, Archer L, Booth S, Mozumder S, Rutherford M, van Smeden M, Lambert P, Snell K. "Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome" (2021) https://doi.org/10.1002/sim.9275

Overview

There are broadly three steps involved in calculating the minimum sample size required for external calibration, in particular to calculate the calibration slope precisely.

1) Specifying the distribution of the linear predictor.
2) Specifying the distribution of the censoring and recurrence times.
3) Specifying the required standard error for the calibration slope.

When these three components are known, then a simulation can take place to estimate the required sample size necessary for external validation. The simulation-based framework is described by the authors Riley and colleagues and is applied to external validation sample size developed using CHHiP. The workflow followed is given in ***Supplementary Figure B10***, taken from [192].

- Step 1: Set-up process:
  - Specify the time point of interest for checking calibration performance.
  - Specify the model's anticipated linear predictor (LP_i) distribution in the validation population. See main text for guidance on how to choose this, for example based on reported histograms, or the D statistic and *C-index.*
  - Specify the anticipated overall $F(t)$ (or $\mathrm{S}(t)$) in the validation population at the time point of interest.
  - Specify the distribution of survival times in the validation population, conditional on the effect of LP_i. In the absence of other evidence, we suggest assuming an exponential distribution for simplicity, but other distributions should be specified if distributional information exists
  - Specify the effect of LP_i. We suggest assuming the existing model will have good calibration, such that the log hazard ratio for the effect of LP_i is 1 (ie, calibration slope is 1). Ensure the chosen parameter(s) of the distribution (eg, baseline rate parameter for exponentially distributed survival times) correspond to the anticipated *F*(*t*) (or *S*(*t*)) at the time point of interest is met, conditional on assuming the calibration slope is 1.
  - Specify the assumed distribution of censoring times in the validation population, and maximum follow-up time. We suggest assuming an exponential distribution in the absence of other information.
  - Specify the target value for the SE, of the calibration slope (eg, 0.051).
- Step 2: Specify a starting sample size and generate a dataset containing the same number of individuals as this sample size.
- Step 3: For each individual i in the dataset, simulate values of LP_i from the assumed linear predictor distribution.
- Step 4: For the time point of interest (prediction time horizon), generate values of $\hat{F}_i(t)$ for each individual by applying the existing prediction model equation. For example, the existing model will typically have an equation of the form $\hat{F}_i(t) = 1 - \hat{S}_0(t)^{\exp(\mathrm{LP}_i)}$ where $\hat{S}_0(t)$ is the model's reported baseline survival probability, and LP_i is the value of the linear predictor for individual i, generated from step 3.
- Step 5: Randomly generate survival times for each individual according to the assumed distribution from step 1 and conditional on their LP_i value. For example, using the *survsim* package in Stata,[29] or *simsurv* in R.[30] For each individual, set their outcome status to be 1 (ie, event) and their follow-up time to be their survival time.
- Step 6: Randomly generate a censoring time for each individual under the censoring distribution assumed in step 1. Also specify the maximum follow-up time for all individuals in the validation study. For those individuals whose survival time (from step 5) is later than their generated censoring time or the maximum follow-up time, change their event status to 0 (ie, no event) and change their follow-up time to their censoring time or the maximum follow-up time (whichever is earlier).
- Step 7: For the chosen time point of interest for prediction, generate pseudo-observations and fit Equation (3) to estimate the calibration slope and its SE, for example by using the *stcoxcal* package in Stata. Store the results (and those for any other measures of interest, eg, net benefit).
- Step 8: Repeat steps 2 to 7 many times (eg, 1000), and each time store the obtained estimates and SEs of the calibration slope.
- Step 9: Summarize the mean calibration slope (simply to check it equals 1 as assumed in step 1) and the mean SE. If the mean SE equals the targeted value specified in step 1, then the sample size in step 2 is the required sample size. Otherwise, repeat steps 2 to 9 with a different chosen sample size.

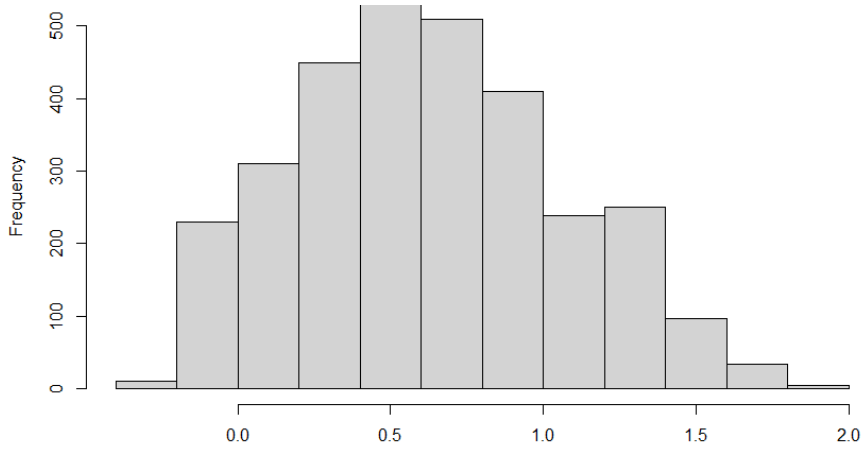
Once the required sample size is identified, plot flexible calibration curves (and confidence intervals) for the simulated datasets, to ascertain whether their spread (and confidence interval width) appears acceptable. Particular attention might be given to regions of risk that are key to clinical decision making.

Supplementary Figure B2 workflow to calculate the required minimum external sample size to precisely estimate the calibration slope at a particular time point. Replicated from [192].

Part 1 - set-up process

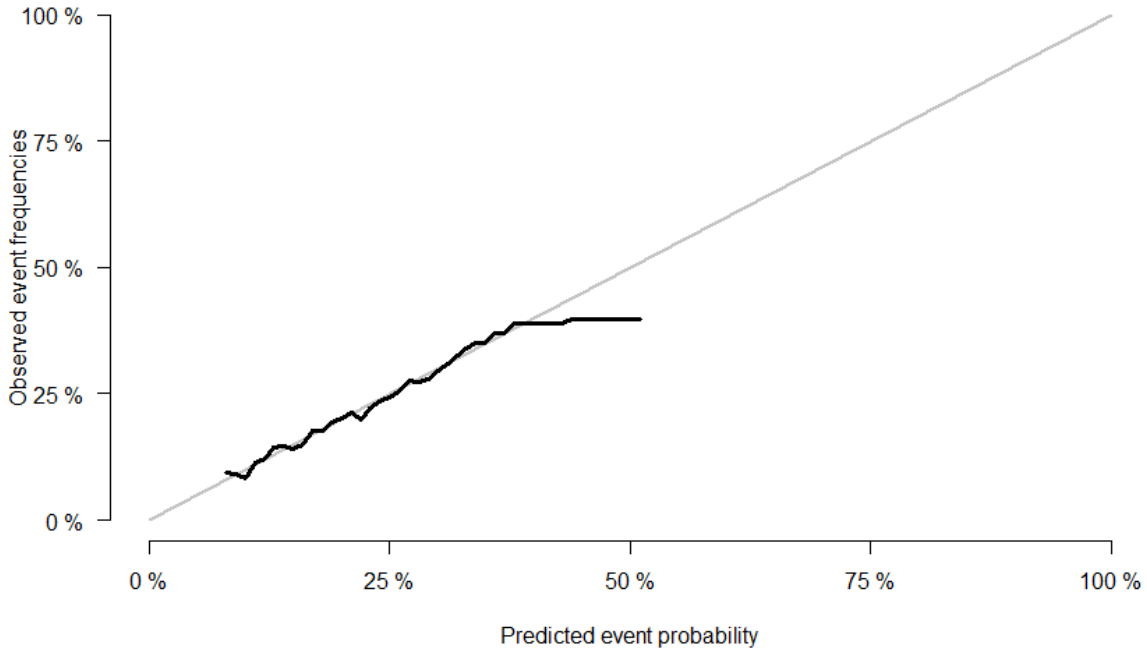
Using [192]'s ***Supplementary Figure B10*** as follows.

- Time point of interest is prediction at the horizon time at eight years.
- The model's linear predictor distribution in the development and validation population is known from CHHiP's Cox proportional hazards submodel. These linear predictors can be extracted to learn its distribution. It gives a skewed near-normal distribution (***Supplementary Figure B11***), with the following moments:
    - Mean = 0.63
    - Variance = 0.19
    - Skewness = 0.25
    - Kurtosis = 2.47



Supplementary Figure B11 histogram of the linear predictor of the Cox submodel for CHHiP.

- The Kaplan-Meier curves shows an event-free rate at 8 years at around 90%, $S(t = 8) = 0.902$ (or $F(t) = 1 - S(t) \rightarrow F(t = 8) = 1 - 0.902 = 0.098$) where 1846 patients are still at risk and there has been 530 cumulative events up to this landmark time. This will be the assumed recurrence-free probability in the validation sample.
- The assumed distribution of survival times in the population, conditional on the effect of the linear predictor. The Cox model predicting at 8 years (from baseline) is reasonably well calibrated, using the *pec::CalPlot* **R** function (see ***Supplementary Figure B12*** below). An exponential distribution with baseline rate parameter $\lambda = 0.0065$ corresponds to $S(t = 8) \approx 0.902$ when the log-hazard ratio for the effect of the linear prediction is 1 (that corresponds to a calibration slope of 1). Therefore, survival times were drawn from this distribution.

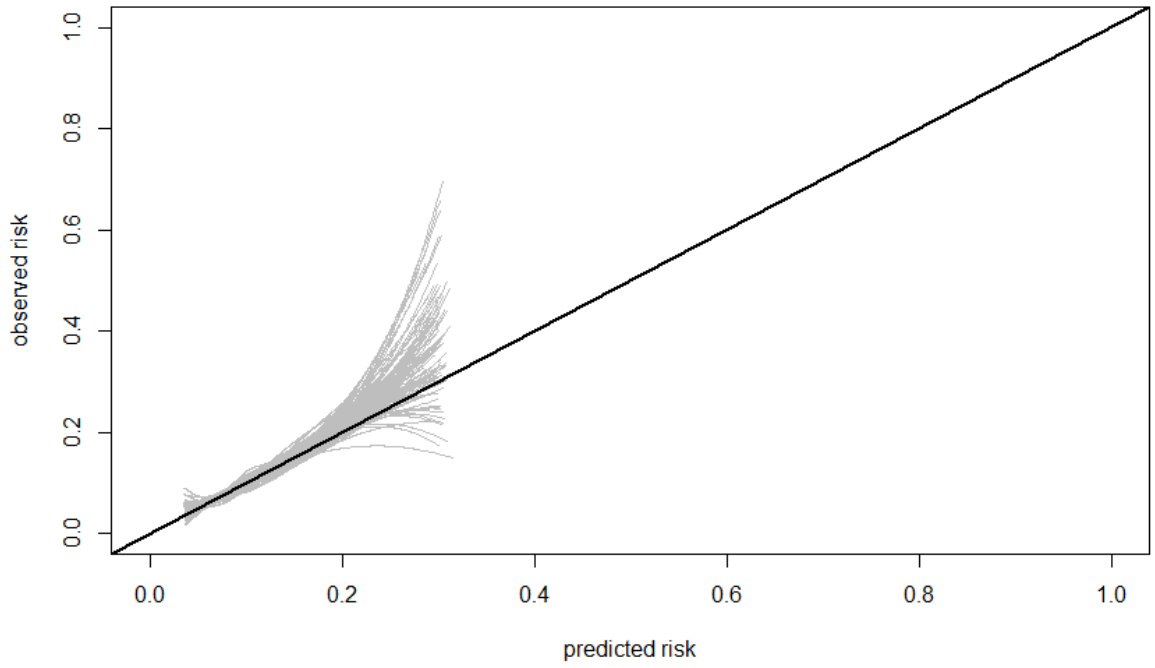


Supplementary Figure B12 calibration plot of the Cox survival submodel for CHHiP at a horizon time of 8 years.

- Specifying the assumed distribution of the censoring times in the population. Censoring was high in CHHiP given the relatively low event rates. For CHHiP, $(3071 - 530 = 2541)$ out of 3071 (83%) participants were censored before, or administratively censored by 8 years in the CHHiP dataset. To replicate this, censoring times were drawn with an assumed constant rate of censoring (which is reasonable given the gradual constant reduction in the red Kaplan-Meier curve, ***Supplementary Figure B1***) from $\exp(\lambda = 0.224)$, giving a censoring probability of $\approx 83\%$ by 8 years. All simulated values would be assigned as censored if the simulated censoring time came before their recurrence time. The maximum horizon time of 8 years was assumed, therefore any values generated after this were administratively censored at 8 years.
- The standard error for the calibration slope was chosen to be ≈ 0.1

Part 2 – simulation

When the assumptions of part 1 are specified then the simulation-based framework provided in Supplementary Figure 10 (steps two to nine) can be performed. This led to approximately 20,000 patients required to estimate the calibration slope with a target standard error of ~0.1. The simulated calibration plots are shown in ***Supplementary Figure B13***. There is some miscalibration at the higher recurrence probabilities, however there is better calibration where the model is to be used for clinical decision-making, i.e., at the lower predicted risks. Slight miscalibration in the higher risk ranges can be acceptable in this circumstance, as decisions to instigate imaging or salvage therapy are more likely to be carried out here in any case, and unlikely to influence decisions at the lower risk thresholds. It is worth noting that this has been carried out with the Cox survival submodel only.



Supplementary Figure B13 simulated calibration curves for a sample of 20,000 patients and a target standard error of 0.1. The grey curves show the estimated and simulated calibration, and the black 45-degree line indicates perfect calibration.

Appendix C (chapter 6)

I calculate the dynamic predictions based on the methodology presented in section **6.2.2**. Using the fitted CRJM, the conditional probabilities of recurrence and death (due to an unrelated cause) were estimated with the Monte Carlo procedure. In particular, I derive the prognosis of two patients (different to the examples used in **Chapter 4**) and their risk of the two competing outcomes. Patient A who presents at 65 years of age, pre-treatment PSA=7.4ng/mL, T-stage=2, Gleason Grade=4+3, hormonal therapy received: LHRHa, randomised fractionation arm: 57Gy/19f. Patient B is 79 years of age at presenting, his baseline covariates and risk factors are: pre-treatment PSA=7.8ng/mL, T-stage=2, Gleason Grade=3+4, hormone therapy received: LHRHa, and randomised fractionation arm: 57Gy/19f.

Predictions up to a horizon time of ten years of follow-up, from presenting PSA ($t = 0$) are presented. In each of the six panels, the above two patients' dynamic predictions are shown, with the left-hand side of each plot indicating PSA values in blue dots, and the predicted PSA trajectory from the longitudinal submodel with the blue curve, provided with shaded 95% credible intervals. The right-hand side indicates the cumulative incidence functions for both patients for each of the two competing outcomes: recurrence (green), and the competing risk of death due to non-disease related causes (red), with the shaded 95% credible intervals for each cumulative incidence function.

Describing each panel of ***Supplementary Figure C1*** (panels **U**—**Z**), both patients have similar initial PSA values at baseline ($t = 0$, first panel **U**). Patient *A* has a slightly higher risk of recurrence at ten years (~20%) than his risk of death (~10%), patient *B* has similar predictions of either outcome of just over 20% by ten years. For both patients, the credible intervals start to increase from 1½—2 years onwards, indicating that with only considering baseline covariates and presenting PSA, the prediction window is precise for the next two years from baseline.

After almost a year of follow-up (second panel **V**, $t \approx 1$) and accrued PSAs for both patients, the predicted risks of each outcome for patient *A* are almost identical. For patient *B*, the risk of death unrelated to cancer remains the same as in the previous landmark $t = 0$.

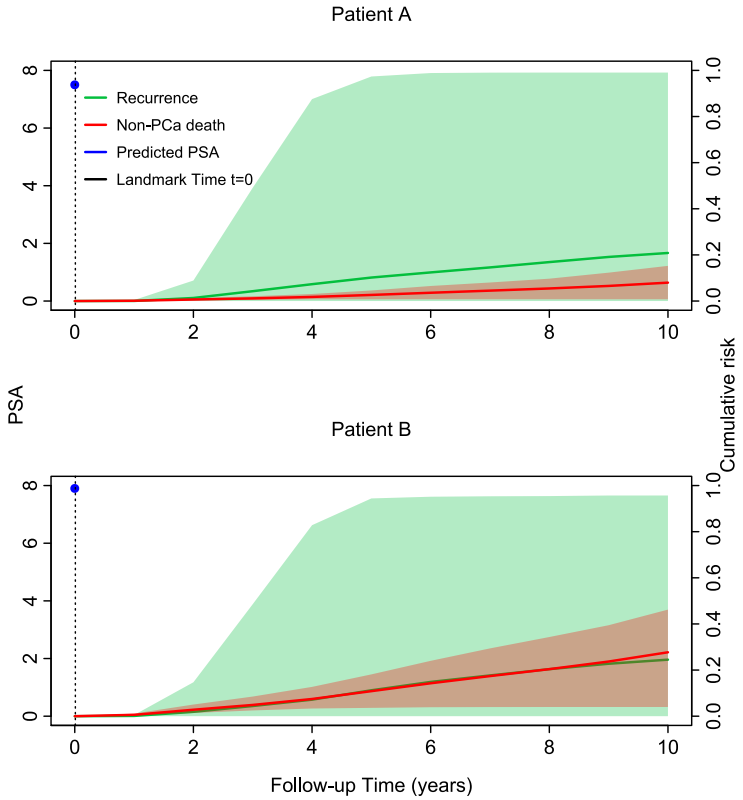
In the third panel **W** ($t \approx 2$), there is a slight increase in PSA post-nadir for *A*, his predicted risk of recurrence by ten years is ~35%, whereas their competing risk of death as decreased slightly with narrower credible intervals. *B*'s PSA remains low and stable with smaller credible intervals for the subsequent two-year prediction window, his risk of recurrence has almost halved with his competing risk remaining similar to the previous landmark time.

In the fourth panel **X** (landmark $t \approx 3$), *A*'s latest PSA has increased beyond 2ng/mL and above the expected upper 97.5% credible interval, their risk of death remains similar, however their 10-year recurrence risk is now 40% (within the next seven years). *B*'s recurrence risk as decreased with much smaller credible intervals for the next 4 years, their risk of death by 10 years remains similar to the previous landmark.

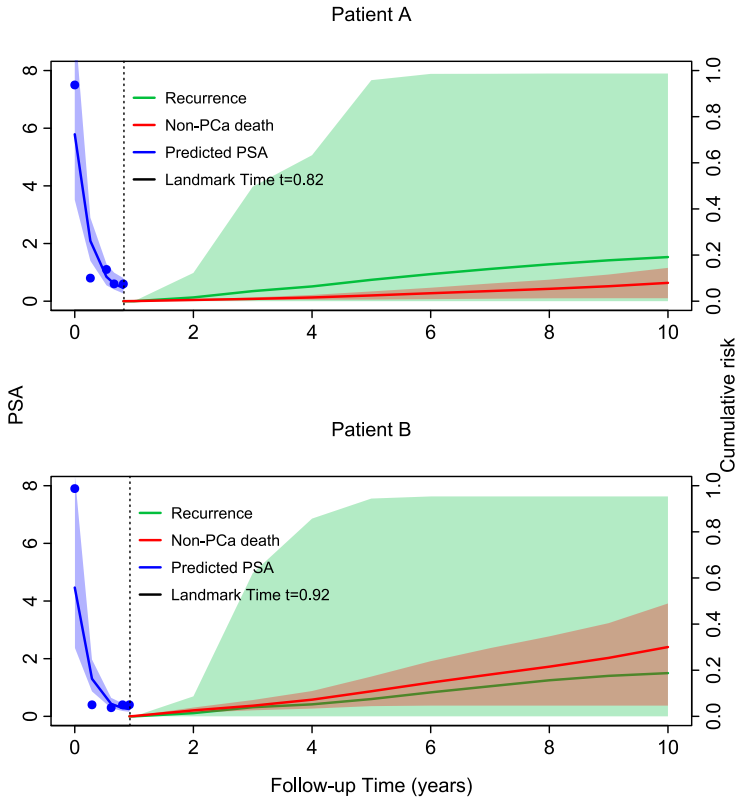
At landmark $t \approx 4$ panel **Y,** *A*'s PSA has decreased slightly however their 10-year recurrence risk remains almost unchanged (~40%) with minimal risk of death occurring first (~5%), *B*'s PSA remains low and stable with their competing risk of death still higher than recurrence (30% *vs* 10% respectively).

In the final panel **Z** landmark time $t = 5.3$ years of follow-up, *A*'s PSA rises yet again with their final two PSAs being above the 97.5th- predicted credible percentile, his risk of recurrence increases to 50% in the next five years. Indeed, this patient does go on to have biochemical failure in the subsequent two years. Patient *B* continues to have low and stable PSAs five years after treatment, therefore has very little risk of recurrence in the next five years and is predicted to have prolonged event-free survival. As this patient ages well into their mid-80s, their risk of death unrelated to prostate cancer expectedly increases to ~25% in the next five years.

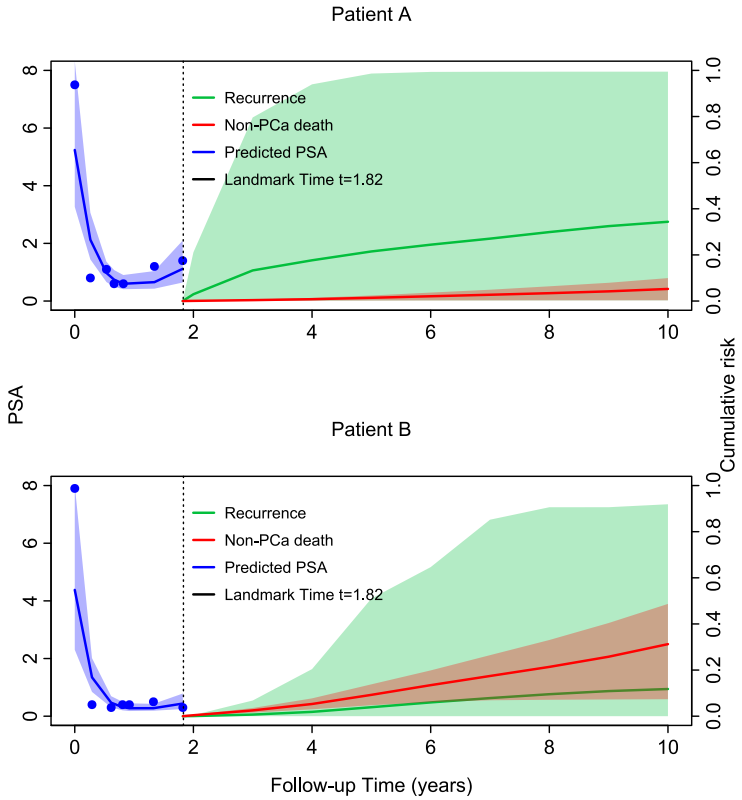
U


Patient A
Recurrence
Non-PCa death
Predicted PSA
Landmark Time t=0
Patient B
PSA
Cumulative risk
Follow-up Time (years)

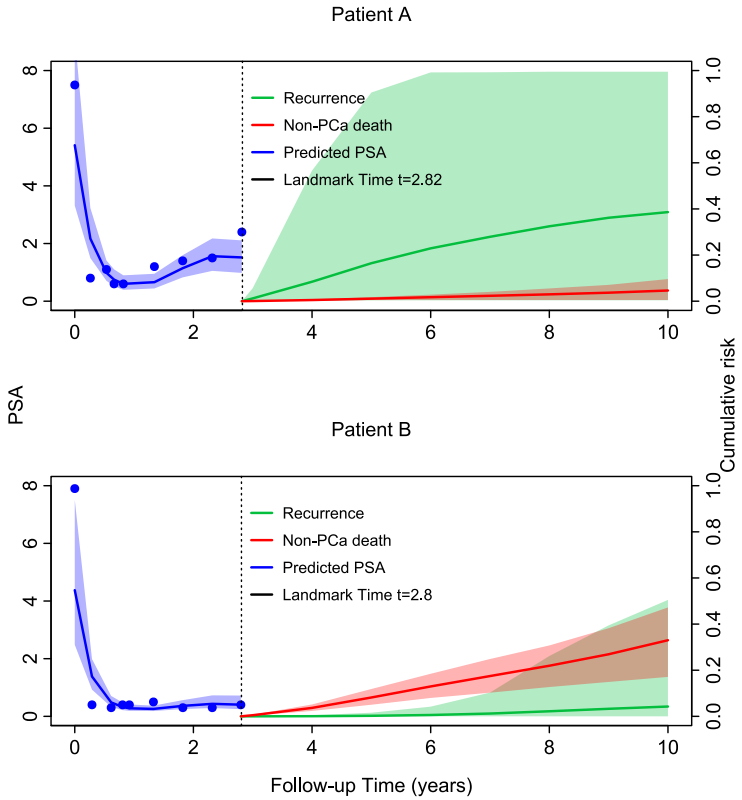
V


Patient A
Recurrence
Non-PCa death
Predicted PSA
Landmark Time t=0.82
Patient B
Recurrence
Non-PCa death
Predicted PSA
Landmark Time t=0.92
PSA
Cumulative risk
Follow-up Time (years)

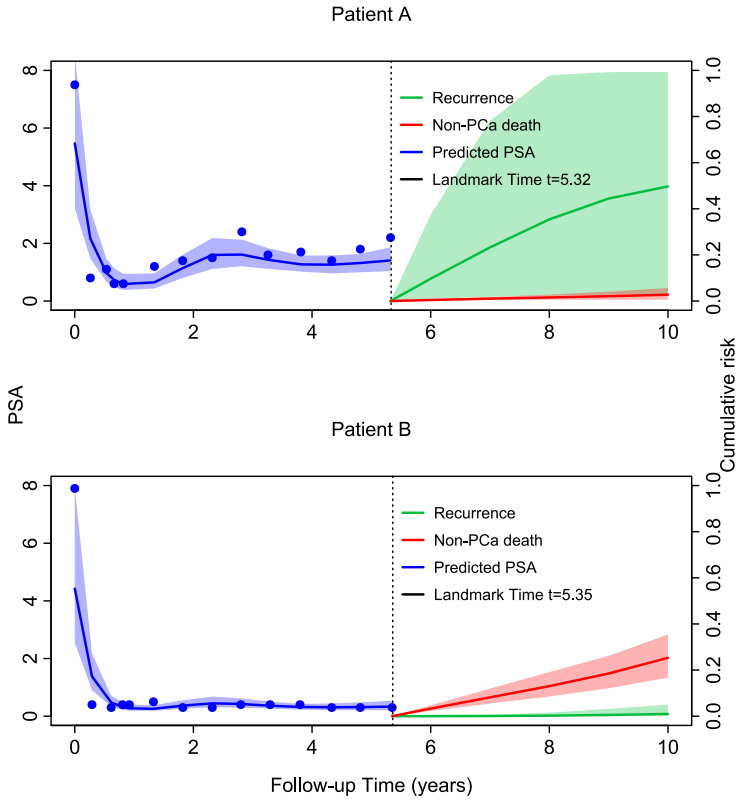
W


Patient A
Recurrence
Non-PCa death
Predicted PSA
Landmark Time t=1.82
PSA
Cumulative risk
Patient B
Recurrence
Non-PCa death
Predicted PSA
Landmark Time t=1.82
Follow-up Time (years)

X


Patient A
Recurrence
Non-PCa death
Predicted PSA
Landmark Time t=2.82
PSA
Cumulative risk
Patient B
Recurrence
Non-PCa death
Predicted PSA
Landmark Time t=2.8
Follow-up Time (years)

Z



Supplementary Figure C3 competing risk dynamic predictions for two patients from the competing risk joint model. Blue on the left-hand side indicates the PSA readings (dots), predicted PSAs (curve), on the right-hand side the cumulative incidence functions of each competing cause: recurrence (green) and death unrelated to prostate cancer (red), with the corresponding shaded 95% credible intervals. The vertical dotted line indicates the landmark time for each patient and their accrued PSAs up to that time point, to then make predictions for the prediction window with horizon time up to ten years.