Personalized Schedules for Invasive Diagnostic Tests

With Applications in Surveillance of Chronic Non-Communicable Diseases

Anirudh Tomer

Rotterdam, July 14, 2020

Personalized Schedules for Invasive Diagnostic Tests

With Applications in Surveillance of Chronic Non-Communicable Diseases

Gepersonaliseerde Schema's voor Invasieve Diagnostische Testen

met Toepassingen voor het Monitoren van Chronische Niet-overdraagbare Ziekten

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op
16 September 2020 om 9:30 uur

door

Anirudh Tomer
geboren te Jorhat, India.

Erasmus University Rotterdam

Promotiecommissie:

Promotoren:	Prof. dr. D. Rizopoulos Prof. dr. E. W. Steyerberg
Overige leden:	Prof. dr. ir. E. H. Boersma Prof. dr. M. J. Roobol Prof. dr. H. Putter

The research described in this thesis was supported by Nederlandse Organisatie voor Wetenschappelijk Onderzoek VIDI grant nr. 016.146.301, and Erasmus University Medical Center funding.

Support by the PRIAS consortium for enabling this research project is gratefully acknowledged.

Support by the Erasmus University Medical Center's Cancer Computational Biology Center for giving access to their IT-infrastructure and software that was used for the computations and data analysis in this research is gratefully acknowledged.

Part of this research was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Dedicated to my parents,
medical science, and patients worldwide

Contents

Contents **i**

1 General Introduction **1**
1.1 Chronic Disease Surveillance 3
1.2 A Joint Model for Time-to-progression and Longitudinal Data 12
1.3 Motivating Studies 14
1.4 Outline of Thesis 17
1.5 References 19

I Methodology **23**

2 Personalized Schedules for Surveillance of Low-risk Prostate Cancer Patients **25**
2.1 Introduction 27
2.2 Joint Model for Time-to-Event and Longitudinal Outcomes 29
2.3 Personalized Schedules for Repeat Biopsies 31
2.4 Evaluation of Schedules 37
2.5 Demonstration of Personalized Schedules 38
2.6 Simulation Study 40

2.7 Discussion 48
2.A Parameter Estimation 51
2.B Ascertainment Bias: PSA Doubling Time-Dependent Biopsies and Competing Events 54
2.C Source Code 56
2.4 References 58

3 Personalized Decision Making for Biopsies in Prostate Cancer Active Surveillance Programs 63
3.1 Introduction 65
3.2 Methods 68
3.3 Results 80
3.4 Discussion 84
3.A Parameter Estimation 87
3.B Source Code 95
3.3 References 97

4 Personalized Schedules for Burdensome Surveillance Tests 101
4.1 Introduction 103
4.2 Joint Model for Time-to-Progression and Longitudinal Outcomes 107
4.3 Personalized Schedule of Invasive Tests for Detecting Progression 109
4.4 Application of Personalized Schedules in Prostate Cancer Surveillance 118
4.5 Simulation Study 121
4.6 Discussion 123
4.A Parameter Estimation 127
4.B Joint Model for the PRIAS Dataset Used in Simulation Study 129
4.C Risk Based Schedules Versus All Possible Schedules 135
4.D Simulation Study Extended Results 136
4.E Partially Observable Markov Decision Processes 138
4.F Source Code 145

4.7 References 146

II Application 151

5 Personalized Biopsy Schedules Based on Risk of Gleason Upgrading for Low-Risk Prostate Cancer Active Surveillance Patients 153

5.1 Introduction 156

5.2 Patients and Methods 160

5.3 Results 165

5.4 Discussion 168

5.5 Conclusions 172

5.A Model Specification 176

5.B Full Results 178

5.C Risk Predictions for Upgrading 180

5.D Source Code 191

5.5 References 196

6 Personalized Screening Intervals for Measurement of N-terminal pro-B-type Natriuretic Peptide Improve Efficiency of Prognostication in Patients with Chronic Heart Failure 203

6.1 Introduction 205

6.2 Methods 206

6.3 Results 212

6.4 Discussion 214

6.A Details: Materials and Methods 218

6.B Supplemental Results 220

6.3 References 224

III Summary 229

7 General Discussion 231

7.1 Background 233

7.2 Subgoals and Research Questions 234

7.3 Recommendations for Practice, and Future Improvements . 240

7.4 General Conclusion 244

7.5 References 246

English Summary, Nederlandse Samenvatting, PhD Portfolio, CV, and Acknowledgements 249

Summary 251

Nederlandse Samenvatting 255

PhD Portfolio 259

CV 262

Acknowledgements 263

Chapter 1

General Introduction

1.1 Chronic Disease Surveillance

Non-communicable diseases (NCDs) such as cancer, diabetes, cardiovascular, and respiratory diseases are a 21st-century global pandemic. They affect men and women equally and cause 60% to 70% of all human deaths worldwide (WHO et al., 2014; Bennett et al., 2018). Often NCDs are chronic. Hence, in many *low-risk* NCD diagnoses (e.g., localized prostate cancer, low-risk dysplasia), immediate serious treatments like surgery, radiotherapy, etc., can induce side-effects and reduce a patient's overall quality of life. A common alternative to immediate treatment is delaying it until the disease has *progressed*, a curable non-terminal disease stage. In this regard, monitoring patients for progression, with curative intent, is called *surveillance*.

The goal of surveillance is to timely detect progression, upon which patients are typically removed and treated. However, the transition of a patient's *disease state* from low-risk to progressed disease is not directly observable. Instead, auxiliary modalities such as biomarkers, physical examinations, medical imaging, biopsies, etc., are used to determine the disease state. Among these, the gold standard *tests* for confirming progression are typically *invasive* (e.g., biopsies). For timely observing the occurrence of progression, invasive tests are conducted repeatedly in surveillance. For example, biopsies are the benchmark test for verifying progression in surveillance of localized prostate cancer (Bokhorst et al., 2015). Similarly, endoscopies are utilized in Barrett's esophagus (Choi and Hur, 2012) and colonoscopies in colorectal cancer (Krist et al., 2007) surveillance. Repeat bronchoscopies, and core biopsies are also employed to detect allograft deterioration in lung (McWilliams et al., 2008) and kidney transplant (Henderson et al., 2011) patients, respectively.

1.1.1 Invasive Test: Burden versus Benefit

Currently, repeated invasive tests are a necessary *burden* for patients. They are indispensable for confirming progression, but they are also difficult to perform, may cause pain, and can lead to severe complications (Loeb et al., 2013; Krist et al., 2007). Consequently, invasive tests are usually planned with a considerable time gap between them. For example, in prostate cancer surveillance, it is recommended to maintain a time difference of one year between consecutive biopsies. However, a time gap between tests also leads to a *time delay* in detecting progression (Figure 1.1). When tests are conducted periodically, this delay can be reduced by scheduling tests frequently. The argument for lowering delay is that detecting progression earlier may provide a larger window of opportunity for curative treatment. Also, timely treatment may also have an impact on the patient's (quality-adjusted) life-years remaining. Hence, a balance between the number and frequency of tests (burden) and time delay in detecting progression (shorter is beneficial) is of crucial importance for patients.

1.1.2 Schedules for Invasive Tests

The frequency of invasive tests varies across diseases and cohorts. However, within a cohort, usually a constant frequency or *fixed schedule* (e.g., every six months) is employed for all patients (Henderson et al., 2011; Bokhorst et al., 2015; Krist et al., 2007). The primary drawback of a fixed schedule is its *one-size-fits-all* assumption. Specifically, high-frequency tests promise shorter delays in detecting progression at the cost of imposing an extra burden on patients who progress slowly and/or patients who never experience progression (e.g., due to comorbidities). The vice versa holds for infrequent tests. Schedules with a skewed burden benefit ratio are also prone to patient non-compliance (Bokhorst et al., 2015; Le Clercq et al., 2015). Reduced compliance for invasive tests may lead to the original problem of delayed detection of disease progression, and reduce the effectiveness of surveillance.

Several improvements have been proposed over one-size-fits-all fixed schedules. The underlying methodology of these advances can be broadly divided into three categories. Namely, sub-group specific fixed schedules, schedules cost-optimized using Markov decision processes, and schedules optimizing a specific utility function of the clinical parameters of interest. Two commonly used terms across these three methodologies are *personal-*

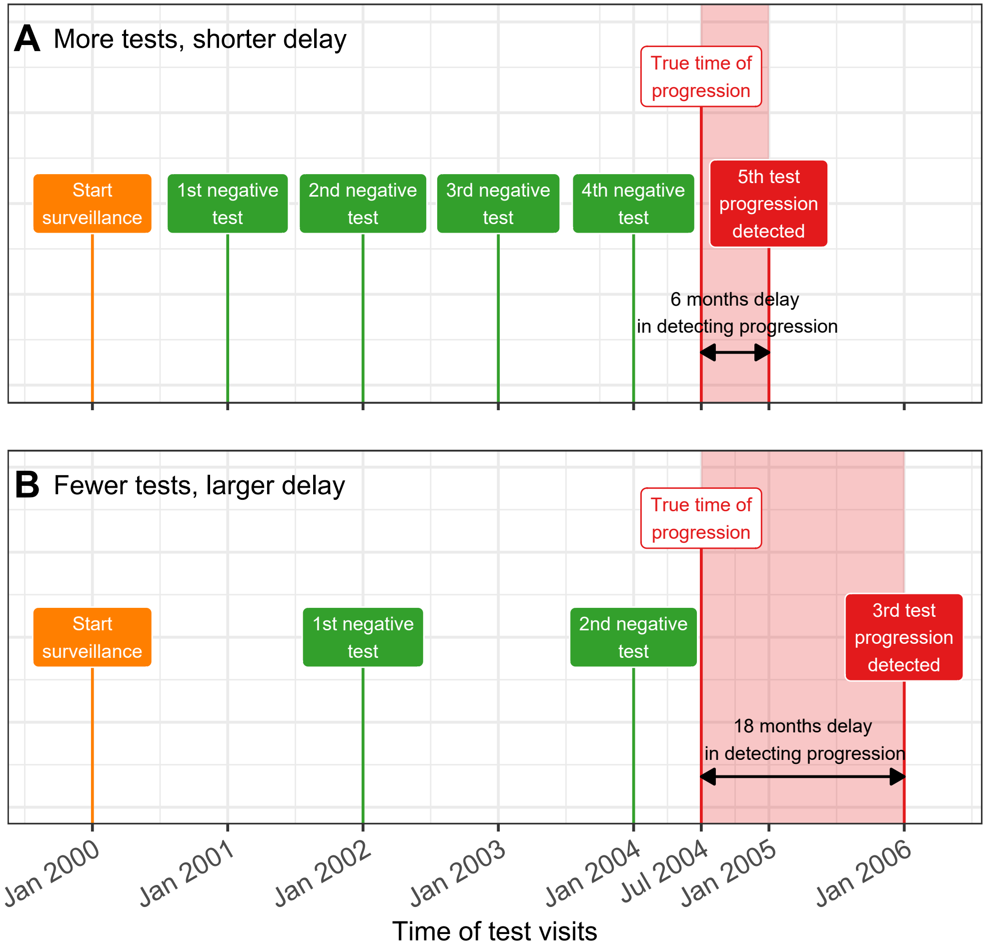

Figure 1.1: **Trade-off between the test frequency and the time delay in detecting disease *progression*:** The true time of disease progression for the patient in this figure is July 2008. More frequent tests in **Panel A**, lead to a shorter time delay in detecting progression, than fewer tests in **Panel B**. Due to the periodical nature of tests, the time of progression is always observed as an interval. For example, between Jan 2004–Jan 2005 in **Panel A** and between Jan 2004–Jan 2006 in **Panel B**.

ized/individualized/tailored schedules, and *optimal* schedules. Loosely, personalization means a unique schedule for each patient in a study population. Optimal refers to mathematical optimization of certain schedule-specific criteria to automatically derive a schedule.

Sub-group specific fixed schedules These schedules are typically prescribed based on observed patient data such as biomarkers, physical examinations, medical imaging, or previous test results. For example, in Barrett's esophagus patients observing low-risk dysplasia on a repeat endoscopy are prescribed future endoscopies every six to twelve months, rather than the standard once every three to five years (Choi and Hur, 2012). Sub-groups are also formed based on multiple results. For example, in the world's largest prostate cancer surveillance PRIAS, the time of biopsies is decided using observed *prostate-specific antigen (PSA)* value, the average rate of change of PSA, the size and shape of the tumor, and previous biopsy results (Figure 1.2). There are two main shortcomings of such heuristic schedules. First, they often create sub-groups based on observed data without accounting for ascertainment biases and measurement error. Second, as illustrated in Figure 1.2, instead of utilizing complete observed data, they typically use only the latest observed value, that too after categorizing continuous ones.

Partially observable Markov decision processes or POMDPs have been utilized in numerous optimal screening and surveillance test schedules for chronic diseases (Steimle and Denton, 2017; Denton, 2018), and especially for nearly all types of cancers (Alagoz et al., 2010). A notable advantage of POMDPs is that they find an optimal schedule from all schedules possible over a set of follow-up visits. The criterion of optimality in POMDPs is the weighted cumulative reward. A reward is a number that is chosen manually for four possible outcomes (true-positive, false-positive, true-negative, and false-negative) of a binary test/no test decision in a schedule. The weighted cumulative reward of a schedule is the weighted sum of all rewards possible with all sequential test decisions in a schedule. The weights are probabilities

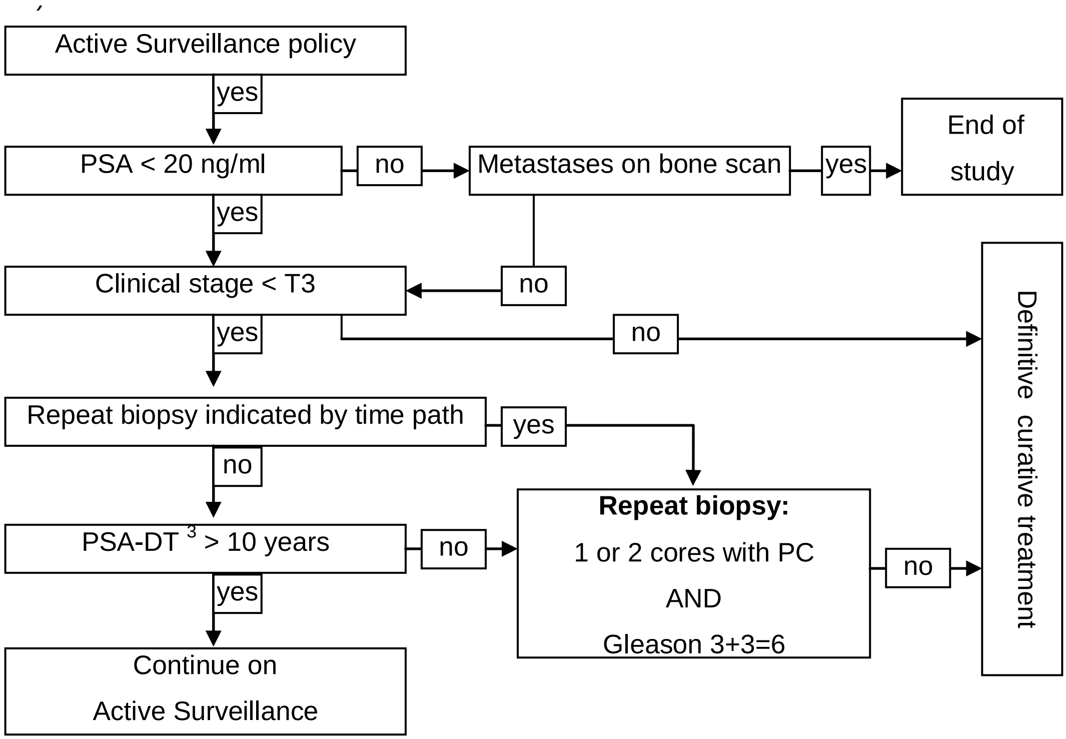

Figure 1.2: **Treatment and biopsy protocol** of the world's largest localized prostate cancer surveillance program PRIAS. Source: https://www.prias-project.org

from a joint probability distribution of the disease state of the patient and the auxiliary outcomes (e.g., biomarkers) that manifest this state. This joint distribution is allowed to change over time.

In general, POMDP algorithms suffer from the curse of dimensionality if continuous longitudinal outcomes or continuous time-space is used (Sunberg and Kochenderfer, 2018). However, a more substantial drawback of POMDPs is their very flexible specification. Specifically, in a simple POMDP with binary test/no test decisions, and binary disease state (low-risk, progressed), it can be shown that there exist infinite possible rewards result in the same optimal schedule (Chapter 4.E). Typically POMDP rewards are chosen based on survey results (Denton, 2018) and translated as quality-adjusted life-years saved. However, with infinite optimal reward sets, any reward set can be cherry-picked, including those that correspond to (improbable) thousands of quality-adjusted life-years saved. Last, to our knowledge, POMDPs are not currently personalized. Since they exploit population-level joint distributions of disease state (e.g., Kaplan-Meier curve) and auxiliary outcomes, the resulting schedules are not personalized.

Schedules optimized for clinical parameters of interest An option to the POMDP framework is optimizing a utility function of the clinical parameters of interest directly (Bebu and Lachin, 2018; Parmigiani, 1996). Examples of clinical parameters are, namely, the financial cost for treating progression, reduction in lifespan due to delayed detection of progression, cost of invasive tests, reduction in quality of life due to invasive tests. Others have proposed optimizing test decision rules for the corresponding sensitivity and specificities in detecting progression (Wang et al., 2019). Alternatively, one may optimize information-theoretic measures such as Wasserstein distance (Hanin et al., 2001) or Kullback-Leibler divergence (Rizopoulos et al., 2016) between the disease state probability distribution at the beginning of surveillance and at a future time point.

In all of these approaches, the expected utility is calculated using the probability distribution of the disease state of the patient. It is standard to use a time-varying disease state distribution. Although, this distribution can be either discrete (e.g., a Markov model with low-risk, medium-risk, progressed disease states) or continuous (e.g., Cox model).

1.1.3 Goal: Developing Personalized Schedules

The overall aim of this work was to develop personalized schedules that better balance the overall burden and benefit of repeated invasive tests in surveillance than one-size-fits-all fixed schedules. The subgoals and specific research questions that we intend to answer in this work are as follows.

- To find a suitable statistical modeling framework to process observed patient data.
- Evaluating the efficacy of different utility functions while planning tests by optimizing clinical parameters of interest (e.g., time delay in detecting progression).
- Evaluating the pros and cons of the widely used POMDP framework for scheduling tests.
- How to schedule invasive tests based on a patient's risk of progression?
- On which criteria should patients chose a personalized schedule over a fixed schedule and vice versa?
- Which factors (e.g., cohort, type of disease) affect the performance of a personalized schedule?
- Can the same test scheduling framework be used across different cohorts and diseases?

To answer our research questions and to develop personalized schedules, the process we followed consisted of four steps. First, processing the observed data of the patient. For example, directly using data via flowcharts (Figure 1.2), using summary statistics, and statistical modeling of observed data, etc. Second, choosing the reward/utility/loss function and the corresponding clinical parameters. Third, defining criteria and methodology for comparing proposed personalized schedules with currently practiced schedules. Fourth, implementing personalized schedules in a computer application for practitioners.

Processing observed data In surveillance, observed data consists of baseline patient characteristics, longitudinally measured outcomes, and previous invasive test results. Since all of these manifest the underlying disease state of the patient, they are usually correlated as well. To accommodate outcomes of various types, we utilized the framework of joint models for time-to-event and longitudinal data (Rizopoulos, 2012; Tsiatis and Davidian, 2004). The motivation of this choice was that joint models combine observed data into a patient-specific cumulative-risk of progression over the entire follow-up period. This risk profile manifests the underlying latent disease state of the patient.

Choosing of reward/utility/loss function and clinical parameters of interest Once a risk profile for progression is available, the next step is to utilize it for optimizing clinical parameters of interest. Examples of these parameters are the time of disease progression, time delay in detecting disease progression given a schedule (Figure 1.1), number and timing of tests in a schedule, cumulative-risk of disease progression, sensitivity/specificity of an invasive test and their derivatives such as Youden index and F1score (López-Ratón et al., 2014). We optimized these parameters via both standard utility functions such as squared loss, absolute loss, multilinear loss (Robert, 2007), and custom utility functions that are a linear sum of multiple clinical parameters of interest.

Comparing personalized versus fixed schedules There are no single perfect criteria to compare schedules. Some important ones, though, are how many patient deaths and/or progression to an advanced disease state (e.g., metastasis) are saved. Reliable data on such metrics are difficult to obtain in low-grade diseases. This is because, in such diseases, the prevalence of death from disease can be quite low (e.g., almost zero in low-grade prostate cancer active surveillance). Hence in this work, we used two other criteria for comparing the performance of proposed personalized schedules with existing fixed schedules; Specifically, the number and timing of invasive tests (burden of tests) and time delay in the detecting progression (shorter is beneficial). Our choice of these criteria is motivated by two reasons. First, we argue that time delay in detection of progression is an easily-quantifiable surrogate for important clinical aspects such as the window of opportunity for curative treatment, risk of adverse downstream outcomes, quality-adjusted remaining lifetime, and additional complications in treating a delayed progression. Similarly, the number and timing of tests manifest financial costs of tests, risk of side-effects, and reduction in quality of life, etc. Second, both the number of tests and time delay in detecting progression are easy to understand for both patients/doctors and can better facilitate *shared decision making* of test schedules.

Computer application implementing personalized schedules While there is no lack of existing methodologies for making invasive test schedules, presenting them in a user-friendly computer/web/phone application may increase their awareness and/or adoption. In this regard, we implemented personalized schedules in a web-application for real patients of the seven largest prostate cancer active surveillance programs. Also, we provide our scheduling methodology as a generic R application programming interface for surveillance of other diseases.

1.2 A Joint Model for Time-to-progression and Longitudinal Data

The first step in developing personalized schedules is processing a patient's surveillance data. This data includes baseline patient features, longitudinally measured outcomes of different types, and previous invasive test results. There are several challenges in modeling such data. First, longitudinal outcomes can be of different types (e.g., binary, continuous), are measured with error, and possibly correlated with each other. Second, usually, longitudinal measurements are not available after the patient is removed from surveillance upon observing progression. Third, patients who observe progression can have more adverse longitudinal data values. Fourth, time of progression is interval-censored (Figure 1.1). Last, combining all this data to obtain a patient's personalized risk of progression. To overcome these challenges, we utilize the framework of joint models for time-to-event and longitudinal data (Rizopoulos, 2012; Tsiatis and Davidian, 2004).

The primary component in joint models is patient-specific random effects (Laird and Ware, 1982). They represent the underlying state of disease, as well as act as the common source of correlation between different outcomes (Figure 1.3) of a patient. Each outcome has a separate sub-model. Usually, mixed-effect sub-models are used for longitudinal outcomes, and a relative risk sub-model is employed for time-to-progression data. The parameters of the different sub-models are estimated jointly. Given a patient's data, the key output from the fitted joint model is a patient's personalized cumulative-risk of progression.

1.2.1 Cumulative-risk of Progression

Consider a joint model is fitted to a particular dataset. Given a new patient's accumulated data, the fitted joint model can predict his cumulative-risk of progression over his entire follow-up period starting from the time of his last negative test. This risk profile manifests the transition of a patient's

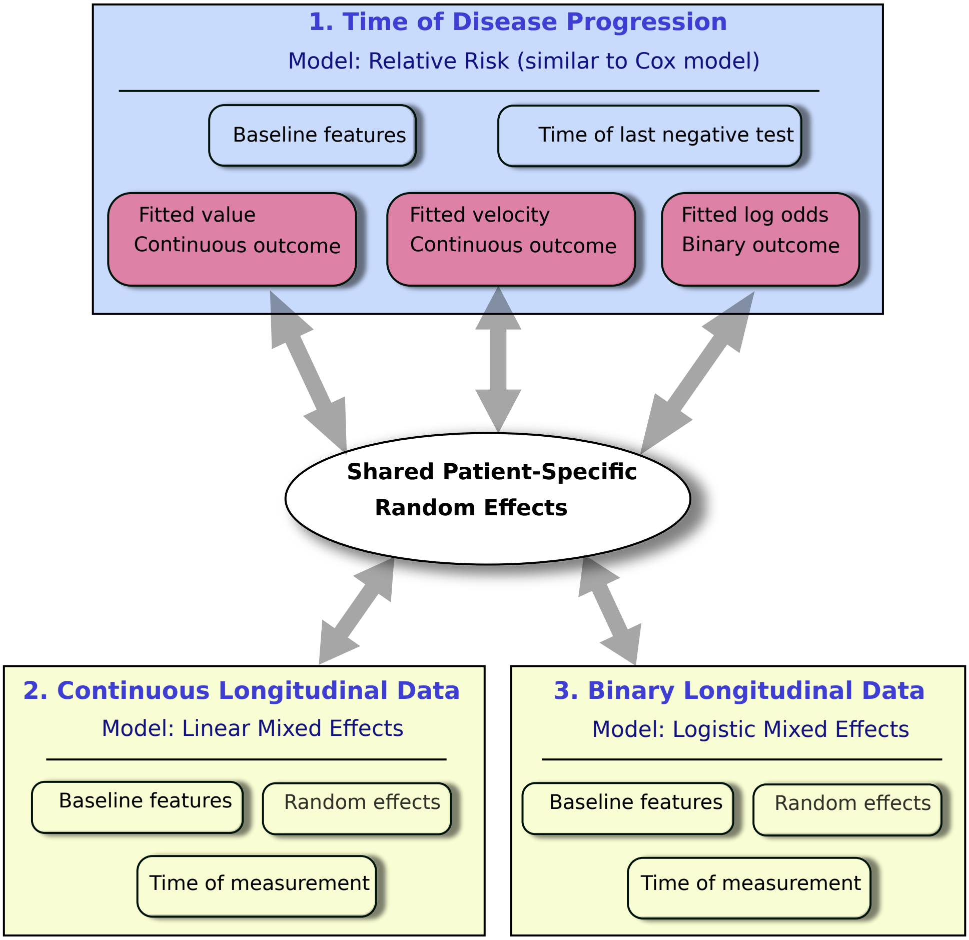
1. Time of Disease Progression
Model: Relative Risk (similar to Cox model)
Baseline features
Time of last negative test
Fitted value Continuous outcome
Fitted velocity Continuous outcome
Fitted log odds Binary outcome
Shared Patient-Specific Random Effects
2. Continuous Longitudinal Data
Model: Linear Mixed Effects
Baseline features
Random effects
Time of measurement
3. Binary Longitudinal Data
Model: Logistic Mixed Effects
Baseline features
Random effects
Time of measurement

Figure 1.3: **Block diagram of a joint model for time-to-progression and longitudinal data**. Typically mixed effect sub-models are utilized for longitudinally measured data, and a relative-risk sub-model is employed for the interval-censored time of progression. The outcomes in these sub-models are conditionally independent of each other, given the common source of correlation patient-specific random-effects (Laird and Ware, 1982). Different features of the longitudinal outcomes such as their fitted value, rate of change, fitted log-odds can be included in the relative-risk sub-model for predicting the risk of progression.

disease state over time from low-risk to progressed. Hence, it can be used to guide the timing of invasive tests. In this regard, we have not only used the cumulative-risk to create personalized schedules but also for calculating a patient's expected time of progression. Although we estimate cumulative-risk using joint models, such estimates can also be obtained via other methods such as landmarking (Van Houwelingen, 2007). In this regard, the scheduling methodology that we propose in this thesis is generic for use with any model that provides the cumulative-risk of progression.

1.3 Motivating Studies

1.3.1 PRIAS: Prostate Cancer Research International Active Surveillance

Our first motivating study is PRIAS (Bul et al., 2013), the world's largest ongoing prostate cancer surveillance study for low- and very-low grade prostate cancer patients. More than 100 medical centers from 17 countries contributed to PRIAS, using a common protocol (`https://www.prias-project.org`). In PRIAS the state of cancer is evaluated via PSA (ng/mL), a blood test; digital rectal examinations (DRE), indicating the shape and size of the tumor; repeat biopsy Gleason grade group (1 to 5), an invasive test; and recently magnetic resonance imaging (MRI). Among these, the biopsy Gleason grade (Epstein et al., 2016) is the strongest indicator of cancer-related outcomes. Consequently, a trigger for treatment in PRIAS is observing an increase in biopsy Gleason grade on repeat biopsy, also informally termed as progression.

Current schedule of biomarkers and biopsies Upon inclusion in PRIAS, PSA (ng/mL) was measured quarterly for the first two years of follow-up and semiannually after that. The DRE was also measured semiannually. The MRI data on tumor volume was very sparsely available in PRIAS. Hence, in this work, we were unable to use it. Biopsies were scheduled at year one,

Table 1.1: **Summary of the PRIAS dataset**. The primary event of interest is cancer progression (increase in biopsy Gleason grade group from grade group 1 to 2 or higher). Abbreviations: PSA is prostate-specific antigen; DRE is digital rectal examination, with level T1c (Schröder et al., 1992) indicating a clinically inapparent tumor which is not palpable or visible by imaging, whereas tumors with DRE $>$ T1c are palpable; IQR is interquartile range; #PSA, #DRE, #biopsies are the number of PSA, DRE, and biopsies conducted, respectively. Chapters 2 and 3 use the December 2016 version of the dataset, but Chapters 4 and 5 utilize the updated April 2019 version.

Characteristic	**Dec 2016 Version**	**Apr 2019 Version**
Total patients	5270	7813
Progression (primary event)	866	1134
Treatment	1488	2250
Watchful waiting	179	334
Lost to follow-up	72	203
Discontinued on request	8	46
Death (other)	61	95
Death (prostate cancer)	2	2
Total DRE measurements	25606	37326
Total PSA measurements	46015	67578
Total biopsies	11042	15686
Median age at diagnosis (years)	70 (IQR: 65–75)	66 (IQR: 61–71)
Median PSA (ng/mL)	5.6 (IQR: 4.0–7.5)	5.7 (IQR: 4.1–7.7)
DRE = T1c (%)	23538/25606 (92%)	34883/37326 (94%)
Median maximum follow-up per patient (years)	1.9 (IQR: 1.0–3.8)	1.8 (IQR: 0.9–4.0)
Median #PSA per patient	7 (IQR: 5–12)	6 (IQR: 4–12)
Median #DRE per patient	4 (IQR: 3–7)	4 (IQR: 2–7)
Median #biopsies per patient	2 (IQR: 1–3)	2 (IQR: 1–2)

four, seven, and ten of follow-up. Additional yearly biopsies were scheduled when PSA doubling time was between zero and ten years (Figure 1.2). The PSA doubling time or PSA-DT is an indicator of the average rate of change of PSA over follow-up. It is measured as the inverse of the slope of the regression line through the base two logarithm of the observed PSA values. Unlike PRIAS's dynamically changing biopsy schedule, in the majority of the prostate cancer surveillance studies worldwide, yearly biopsies are the norm (Loeb et al., 2014; Nieboer et al., 2018).

1.3.2 Bio-SHiFT: The Role of Biomarkers and Echocardiography in Prediction of Prognosis of Chronic Heart Failure Patients

Our second motivating study is called Bio-SHiFT (van Boven et al., 2018), a prospective ongoing study with currently 263 patients followed-up over a period of 30 months. The goal of Bio-SHiFT is to evaluate the performance of blood biomarkers in the prognosis of chronic heart failure. In this thesis, we focused only on one such biomarker, called NT-proBNP (Bhalla et al., 2004). Measuring NT-proBNP requires only a blood sample, and thus it less burdensome than biopsies or endoscopies. However, when measured repeatedly for the prognosis of heart failure, the overall burden accumulates over time. Currently, NT-proBNP is measured once every three months. Since only 70 out of 263 patients had adverse heart failure related events (cardiac death, cardiac transplantation, left ventricular assist device implantation, or heart failure hospitalization), many patients may not require some of the NT-proBNP measurements prescribed in the fixed schedule. Hence, we aimed to reduce patient burden by providing them a personalized schedule for measuring NT-proBNP. To this end, we used an existing scheduling methodology (Rizopoulos et al., 2016). This approach balances information gained from an extra NT-proBNP measurement and the risk of missing an adverse event if NT-proBNP is not measured.

Table 1.2: **Summary of the Bio-SHiFT dataset**. The primary study endpoint (PE) was defined as the composite of cardiac death, cardiac transplantation, left ventricular assist device implantation, or hospitalization for heart failure, whichever occurred first. Abbreviations: NYHA is New York Heart Association Classification (Bredy et al., 2018); IQR is interquartile range.

Characteristic	Value
Total patients	263
PE (primary endpoint)	70
Total NT-proBNP measurements	2022
Median NT-proBNP (pg/mL)	110.3 (IQR: 38.5–240.9)
Median age at inclusion (years)	67.9 (IQR: 58.9–75.8)
Median BMI at inclusion	26.5 (IQR: 24.4–30.1)
Median NYHA (assumed continuous)	2 (IQR: 1–3)
Gender = Female (%)	74/263 (28.1%)
Renal failure history = Yes (%)	136/263 (51.7%)
Type-II diabetes mellitus = Yes (%)	81/263 (30.8%)
Median maximum follow-up per patient (years)	2.1 (IQR: 1.2–2.4)
Median #NT-proBNP per patient	9 (IQR: 5–10)

1.4 Outline of Thesis

The outline of the rest of this thesis is as follows. In Chapter 2, using loss functions from Bayesian decision theory, we develop a methodology for personalized biopsy decisions in prostate cancer active surveillance. In Chapter 3, we extend the joint model proposed in Chapter 2 to account for both PSA and DRE longitudinal outcomes. Also, we focus exclusively on progression-risk based personalized biopsy decisions and conduct a more realistic simulation study than Chapter 2. In Chapter 4, we generalize our model for use surveillance across different chronic diseases and extend single optimal biopsy decisions to full optimal biopsy schedules. To this end, we define and utilize two measures of performance of a schedule. These are, namely, the expected number of invasive tests and the expected time delay in detecting progression. We evaluate the POMDP framework in Chapter 4.E. We also apply our model and methodology in a real-world scenario. Specifically, in Chapter 5, we first externally validate a joint model fitted to the PRIAS prostate cancer dataset in six largest cohorts of the Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) database. Then we implement the validated models and personalized schedules in a web-application. Lastly, in Chapter 6, we demonstrate the use of personalized schedules for planning biomarker measurements.

1.5 References

Alagoz, O., Ayer, T., and Erenay, F. S. (2010). Operations research models for cancer screening. *Wiley Encyclopedia of Operations Research and Management Science*.

Bebu, I. and Lachin, J. M. (2018). Optimal screening schedules for disease progression with application to diabetic retinopathy. *Biostatistics*, 19(1):1–13.

Bennett, J. E., Stevens, G. A., Mathers, C. D., Bonita, R., Rehm, J., Kruk, M. E., Riley, L. M., Dain, K., Kengne, A. P., Chalkidou, K., et al. (2018). NCD countdown 2030: worldwide trends in non-communicable disease mortality and progress towards sustainable development goal target 3.4. *The Lancet*, 392(10152):1072–1088.

Bhalla, V., Willis, S., and Maisel, A. S. (2004). B-Type natriuretic peptide: The level and the drug—partners in the diagnosis and management of congestive heart failure. *Congestive Heart Failure*, 10:3–27.

Bokhorst, L. P., Alberts, A. R., Rannikko, A., Valdagni, R., Pickles, T., Kakehi, Y., Bangma, C. H., Roobol, M. J., and PRIAS study group (2015). Compliance rates with the Prostate Cancer Research International Active Surveillance (PRIAS) protocol and disease reclassification in noncompliers. *European Urology*, 68(5):814–821.

Bredy, C., Ministeri, M., Kempny, A., Alonso-Gonzalez, R., Swan, L., Uebing, A., Diller, G.-P., Gatzoulis, M. A., and Dimopoulos, K. (2018). New York Heart Association (NYHA) classification in adults with congenital heart disease: relation to objective measures of exercise and outcome. *European Heart Journal-Quality of Care and Clinical Outcomes*, 4(1):51–58.

Bul, M., Zhu, X., Valdagni, R., Pickles, T., Kakehi, Y., Rannikko, A., Bjartell, A., Van Der Schoot, D. K., Cornel, E. B., Conti, G. N., et al.

(2013). Active surveillance for low-risk prostate cancer worldwide: the PRIAS study. *European Urology*, 63(4):597–603.

Choi, S. E. and Hur, C. (2012). Screening and surveillance for Barrett's esophagus: current issues and future directions. *Current Opinion in Gastroenterology*, 28(4):377.

Denton, B. T. (2018). Optimization of sequential decision making for chronic diseases: From data to decisions. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, pages 316–348. INFORMS.

Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., and Humphrey, P. A. (2016). The 2014 International Society of Urological Pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 40(2):244–252.

Hanin, L., Tsodikov, A., and Yakovlev, A. Y. (2001). Optimal schedules of cancer surveillance and tumor size at detection. *Mathematical and Computer Modelling*, 33(12-13):1419–1430.

Henderson, L., Nankivell, B., and Chapman*, J. (2011). Surveillance protocol kidney transplant biopsies: their evolving role in clinical practice. *American Journal of Transplantation*, 11(8):1570–1575.

Krist, A. H., Jones, R. M., Woolf, S. H., Woessner, S. E., Merenstein, D., Kerns, J. W., Foliaco, W., and Jackson, P. (2007). Timing of repeat colonoscopy: disparity between guidelines and endoscopists' recommendation. *American Journal of Preventive Medicine*, 33(6):471–478.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Le Clercq, C., Winkens, B., Bakker, C., Keulen, E., Beets, G., Masclee, A., and Sanduleanu, S. (2015). Metachronous colorectal cancers result from missed lesions and non-compliance with surveillance. *Gastrointestinal Endoscopy*, 82(2):325–333.e2.

Loeb, S., Carter, H. B., Schwartz, M., Fagerlin, A., Braithwaite, R. S., and Lepor, H. (2014). Heterogeneity in active surveillance protocols worldwide. *Reviews in Urology*, 16(4):202–203.

Loeb, S., Vellekoop, A., Ahmed, H. U., Catto, J., Emberton, M., Nam, R., Rosario, D. J., Scattoni, V., and Lotan, Y. (2013). Systematic review of complications of prostate biopsy. *European Urology*, 64(6):876–892.

López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C., Gude-Sampedro, F., et al. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, 61(8):1–36.

McWilliams, T. J., Williams, T. J., Whitford, H. M., and Snell, G. I. (2008). Surveillance bronchoscopy in lung transplant recipients: risk versus benefit. *The Journal of Heart and Lung Transplantation*, 27(11):1203–1209.

Nieboer, D., Tomer, A., Rizopoulos, D., Roobol, M. J., and Steyerberg, E. W. (2018). Active surveillance: a review of risk-based, dynamic monitoring. *Translational andrology and Urology*, 7(1):106–115.

Parmigiani, G. (1996). Optimal scheduling of fallible inspections. *Operations Research*, 44(2):360–367.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press.

Rizopoulos, D., Taylor, J. M., Van Rosmalen, J., Steyerberg, E. W., and Takkenberg, J. J. (2016). Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics*, 17(1):149–164.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Schröder, F., Hermanek, P., Denis, L., Fair, W., Gospodarowicz, M., and Pavone-Macaluso, M. (1992). The TNM classification of prostate cancer. *The Prostate*, 21(S4):129–138.

Steimle, L. N. and Denton, B. T. (2017). Markov decision processes for screening and treatment of chronic diseases. In *Markov Decision Processes in Practice*, pages 189–222. Springer.

Sunberg, Z. N. and Kochenderfer, M. J. (2018). Online algorithms for pomdps with continuous state, action, and observation spaces. In *Twenty-Eighth International Conference on Automated Planning and Scheduling*.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

van Boven, N., Battes, L. C., Akkerhuis, K. M., Rizopoulos, D., Caliskan, K., Anroedh, S. S., Yassi, W., Manintveld, O. C., Cornel, J.-H., Constantinescu, A. A., et al. (2018). Toward personalized risk assessment in patients with chronic heart failure: detailed temporal patterns of NT-proBNP, troponin T, and CRP in the Bio-SHiFT study. *American Heart Journal*, 196:36–48.

Van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85.

Wang, Y., Zhao, Y.-Q., and Zheng, Y. (2019). Learning-based biomarker-assisted rules for optimized clinical benefit under a risk-constraint. *Biometrics*, pages 1–10.

WHO, W. H. O. et al. (2014). *Global status report on noncommunicable diseases 2014*. Number WHO/NMH/NVI/15.1. World Health Organization.

Part I

Methodology

Chapter 2

Personalized Schedules for Surveillance of Low-risk Prostate Cancer Patients

This chapter is based on the paper

Tomer, A., Nieboer, D., Roobol, M.J., Steyerberg, E.W., and Rizopoulos, D. (2019), Personalized schedules for surveillance of low-risk prostate cancer patients. *Biometrics*, 75: 153–162. doi: `https://doi.org/10.1111/biom.12940`

Abstract

Low-risk prostate cancer patients enrolled in active surveillance (AS) programs commonly undergo biopsies on a frequent basis for examination of cancer progression. AS programs employ a fixed schedule of biopsies for all patients. Such fixed and frequent schedules may schedule unnecessary biopsies. Since biopsies are burdensome, patients do not always comply with the schedule, which increases the risk of delayed detection of cancer progression. Motivated by the world's largest AS program, Prostate Cancer Research International Active Surveillance (PRIAS), we present personalized schedules for biopsies to counter these problems. Using joint models for time-to-event and longitudinal data, our methods combine information from historical prostate-specific antigen levels and repeat biopsy results of a patient, to schedule the next biopsy. We also present methods to compare personalized schedules with existing biopsy schedules.

2.1 Introduction

Prostate cancer (PCa) is the second most frequently diagnosed cancer (14% of all cancers) in males worldwide (Torre et al., 2015). The increase in the diagnosis of low-grade PCa has been attributed to an increase in life expectancy and an increase in the number of screening programs (Potosky et al., 1995). An issue of screening programs that has also been established in other types of cancers (e.g., breast cancer) is over-diagnosis. To avoid overtreatment, patients diagnosed with low-grade PCa are commonly advised to join active surveillance (AS) programs. In order to delay serious treatments such as surgery, chemotherapy, or radiotherapy, in AS PCa progression is routinely examined via serum prostate-specific antigen (PSA) levels, digital rectal examination, medical imaging, and biopsy, etc.

Biopsies are the most painful, prone to medical complications (Loeb et al., 2013) and yet also the most reliable PCa progression examination technique used in AS. When a patient's biopsy Gleason grading becomes larger than 6 (Gleason reclassification or GR), he is advised to switch from AS to active treatment (Bokhorst et al., 2015). Hence the timing of biopsies has significant medical implications. The world's largest AS program, Prostate Cancer Research International Active Surveillance (PRIAS) conducts biopsies at year one, year four, year seven and year ten of follow-up, and every five years thereafter. However, it switches to a more frequent, annual biopsy schedule for faster-progressing patients. These are patients with PSA doubling time (PSA-DT) between 0 and 10 years, which is measured as the inverse of the slope of the regression line through the base two logarithm of PSA values. In contrast, many AS programs use annual schedule for all patients (Tosoian et al., 2011; Welty et al., 2015). Consequently, for slowly-progressing PCa patients, many unnecessary biopsies are scheduled. Furthermore, patients may not always comply with such schedules (Bokhorst et al., 2015), which can lead to delayed detection of PCa and reduce the effectiveness of AS.

This paper is motivated by the need to reduce the medical burden of repeat biopsies while simultaneously avoiding the late detection of PCa progression. To this end, we intend to develop personalized schedules for biopsies using historical PSA measurements and biopsy results of patients. Personalized schedules for screening have received much interest in the literature, especially in the medical decision making context. For example, Markov decision process (MDP) models have been used to create personalized screening schedules for diabetic retinopathy (Bebu and Lachin, 2018), breast cancer (Ayer et al., 2012), cervical cancer (Akhavan-Tabatabaei et al., 2017), and colorectal cancer (Erenay et al., 2014). Another type of model called a joint model for time-to-event and longitudinal data (Tsiatis and Davidian, 2004; Rizopoulos, 2012) has also been used to create personalized schedules for the measurement of longitudinal biomarkers (Rizopoulos et al., 2016). In the context of PCa, Zhang et al. (2012) have used partially observable MDP models to personalize the decision of (not) deferring a biopsy to the next check-up time during the screening process. This decision is based on the baseline characteristics as well as a discretized PSA level of the patient at the current check-up time.

In comparison to the work referenced above, the schedules we propose in this paper account for the latent between-patient heterogeneity. We achieve this by using joint models, which are inherently patient-specific because they utilize random effects. Secondly, joint models allow a continuous time scale and utilize the entire history of PSA levels. Lastly, instead of making a binary decision of (not) deferring a biopsy to the next pre-scheduled check-up time, we schedule biopsies at a per-patient optimal future time. To this end, using joint models, we first obtain a full specification of the joint distribution of PSA levels and time of GR. We then use it to define a patient-specific posterior predictive distribution of the time of GR, given the observed PSA measurements and repeat biopsies up to the current check-up time. Using the general framework of Bayesian decision theory, we propose a set of loss functions that are minimized to find the optimal time of conducting a biopsy. These loss functions yield us two categories of personalized schedules, those based on the expected time of GR and those based on the risk of GR. In addition, we analyze an approach where the two types of schedules are combined. We also present methods to evaluate and compare the various schedules for biopsies.

The rest of the paper is organized as follows. Section 2.2 briefly covers the joint modeling framework. Section 2.3 details the personalized scheduling approaches we have proposed in this paper. In Section 2.4 we discuss methods for evaluation and selection of a schedule. In Section 2.5 we demonstrate the personalized schedules by employing them for the patients from the PRIAS program. Lastly, in Section 2.6, we present the results of a simulation study we conducted to compare personalized schedules with PRIAS and annual schedule.

2.2 Joint Model for Time-to-Event and Longitudinal Outcomes

We start with a short introduction of the joint modeling framework we will use in our following developments. Let T_i^* denote the true GR time for the i-th patient and let S be the schedule of his biopsies. Let the vector of the time of biopsies be denoted by $T_i^S = \{T_{i0}^S, T_{i1}^S, \ldots, T_{iN_i^S}^S; T_{ij}^S < T_{ik}^S, \forall j < k\}$, where N_i^S are the total number of biopsies conducted. Because biopsy schedules are periodical, T_i^* cannot be observed directly and it is only known to fall in an interval $l_i < T_i^* \leq r_i$, where $l_i = T_{iN_i^S - 1}^S, r_i = T_{iN_i^S}^S$ if GR is observed, and $l_i = T_{iN_i^S}^S, r_i = \infty$ if GR is not observed yet. Further let $\boldsymbol{y}_i$ denote the $n_i \times 1$ vector of PSA levels for the i-th patient. For a sample of n patients the observed data is denoted by $\mathcal{D}_n = \{l_i, r_i, \boldsymbol{y}_i; i = 1, \ldots, n\}$.

The longitudinal outcome of interest, namely PSA level, is continuous in nature and thus to model it the joint model utilizes a linear mixed effects model (LMM) of the form:

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) \\ &= \boldsymbol{x}_i^T(t)\boldsymbol{\beta} + \boldsymbol{z}_i^T(t)\boldsymbol{b}_i + \varepsilon_i(t), \end{aligned}$$

where $\boldsymbol{x}_i(t)$ and $\boldsymbol{z}_i(t)$ denote the row vectors of the design matrix for fixed and random effects, respectively. The fixed and random effects are denoted by $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, respectively. The random effects are assumed to be normally distributed with mean zero and $q \times q$ covariance matrix $\boldsymbol{D}$. The true and unobserved, error free PSA level at time t is denoted by $m_i(t)$. The error $\varepsilon_i(t)$ is assumed to be t-distributed with three degrees of freedom and scale σ, and is independent of the random effects $\boldsymbol{b}_i$.

To model the effect of PSA on hazard of GR, joint models utilize a relative risk sub-model. The hazard of GR for patient i at any time point t, denoted by $h_i(t)$, depends on a function of subject specific linear predictor $m_i(t)$ and/or the random effects:

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) &= \lim_{\Delta t \to 0} \frac{\text{Pr}\Big\{t \leq T_i^* < t + \Delta t \mid T_i^* \geq t, \mathcal{M}_i(t), \boldsymbol{w}_i\Big\}}{\Delta t} \\ &= h_0(t) \exp\Big[\boldsymbol{\gamma}^T \boldsymbol{w}_i + f\{\mathcal{M}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\}\Big], \quad t > 0, \end{aligned}$$

where $\mathcal{M}_i(t) = \{m_i(v), 0 \leq v \leq t\}$ denotes the history of the underlying PSA levels up to time t. The vector of baseline covariates is denoted by $\boldsymbol{w}_i$, and $\boldsymbol{\gamma}$ are the corresponding parameters. The function $f(\cdot)$ parametrized by vector $\boldsymbol{\alpha}$ specifies the functional form of PSA levels (Brown, 2009; Rizopoulos, 2012; Taylor et al., 2013; Rizopoulos et al., 2014) that is used in the linear predictor of the relative risk model. Some functional forms relevant to the problem at hand are the following:

$$\begin{cases} f\{\mathcal{M}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\} = \alpha m_i(t), \\ f\{\mathcal{M}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\} = \alpha_1 m_i(t) + \alpha_2 m_i'(t), \quad \text{with } m_i'(t) = \frac{\mathrm{d}m_i(t)}{\mathrm{d}t}. \end{cases}$$

These formulations of $f(\cdot)$ postulate that the hazard of GR at time t may be associated with the underlying level $m_i(t)$ of the PSA at t, or with both the level and velocity $m_i'(t)$ of the PSA at t. Lastly, $h_0(t)$ is the baseline hazard at time t, and is modeled flexibly using P-splines. More specifically:

$$\log h_0(t) = \gamma_{h_0,0} + \sum_{q=1}^{Q} \gamma_{h_0,q} B_q(t, \boldsymbol{v}),$$

where $B_q(t, \boldsymbol{v})$ denotes the q-th basis function of a B-spline with knots $\boldsymbol{v} = v_1, \ldots, v_Q$ and vector of spline coefficients γ_{h_0}. To avoid choosing the number and position of knots in the spline, a relatively high number of knots (e.g., 15 to 20) are chosen and the corresponding B-spline regression coefficients γ_{h_0} are penalized using a differences penalty (Eilers and Marx, 1996). Parameter estimation using the Bayesian approach is presented in Appendix 2.A.

2.3 Personalized Schedules for Repeat Biopsies

We intend to use the joint model fitted to $\mathcal{D}_n$, to create personalized schedules of biopsies. To this end, let us assume that a schedule is to be created for a new patient j, who is not present in $\mathcal{D}_n$. Let t be the time of his latest biopsy, and $\mathcal{Y}_j(s)$ denote his historical PSA measurements up to time s. The goal is to find the optimal time $u > \max(t, s)$ of the next biopsy.

2.3.1 Posterior Predictive Distribution for Time to GR

The information from $\mathcal{Y}_j(s)$ and repeat biopsies is manifested by the posterior predictive distribution $g(T_j^*)$, given by (baseline covariates $\boldsymbol{w}_i$ are not shown for brevity hereafter):

$$
\begin{aligned}
g(T_j^*) &= p\Big\{T_j^* \mid T_j^* > t, \mathcal{Y}_j(s), \mathcal{D}_n\Big\}\\
&= \int p\Big\{T_j^* \mid T_j^* > t, \mathcal{Y}_j(s), \boldsymbol{\theta}\Big\} p\Big(\boldsymbol{\theta} \mid \mathcal{D}_n\Big) \mathrm{d}\boldsymbol{\theta}\\
&= \int \int p\Big(T_j^* \mid T_j^* > t, \boldsymbol{b}_j, \boldsymbol{\theta}\Big) p\Big\{\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(s), \boldsymbol{\theta}\Big\} p\Big(\boldsymbol{\theta} \mid \mathcal{D}_n\Big) \mathrm{d}\boldsymbol{b}_j \mathrm{d}\boldsymbol{\theta}.
\end{aligned}
$$

The distribution $g(T_j^*)$ depends on $\mathcal{Y}_j(s)$ and $\mathcal{D}_n$ via the posterior distribution of random effects $\boldsymbol{b}_j$ and posterior distribution of the vector of all parameters $\boldsymbol{\theta}$, respectively.

2.3.2 Loss Functions

To find the time u of the next biopsy, we use principles from statistical decision theory in a Bayesian setting (Berger, 1985; Robert, 2007). More specifically, we propose to choose u by minimizing the posterior expected loss $E_g\Big\{L(T_j^*, u)\Big\}$, where the expectation is taken with respect to $g(T_j^*)$. The former is given by:

$$E_g\Big\{L(T_j^*, u)\Big\} = \int_t^\infty L(T_j^*, u) p\Big\{T_j^* \mid T_j^* > t, \mathcal{Y}_j(s), \mathcal{D}_n\Big\} \mathrm{d}T_j^*.$$

Various loss functions $L(T_j^*, u)$ have been proposed in literature (Robert, 2007). The ones we utilize, and the corresponding motivations are presented next.

Given the burden of biopsies, ideally only one biopsy performed at the exact time of GR is sufficient. Hence, neither a time which overshoots the true GR time T_j^*, nor a time which undershoots it, is preferred. In this regard, the squared loss function $L(T_j^*, u) = (T_j^* - u)^2$ and the absolute loss function $L(T_j^*, u) = \Big|T_j^* - u\Big|$ have the properties that the posterior expected loss is symmetric on both sides of T_j^*. Secondly, both loss functions have well known solutions available. The posterior expected loss for the squared loss function is given by:

$$\begin{aligned} E_g\Big\{L(T_j^*, u)\Big\} &= E_g\Big\{(T_j^* - u)^2\Big\} \\ &= E_g\Big\{(T_j^*)^2\Big\} + u^2 - 2uE_g(T_j^*). \end{aligned} \tag{2.1}$$

The posterior expected loss in (2.1) attains its minimum at $u = E_g(T_j^*)$, that is, the expected time of GR. The posterior expected loss for the absolute loss function is given by:

$$\begin{aligned} E_g\Big\{L(T_j^*, u)\Big\} &= E_g\Big(\Big|T_j^* - u\Big|\Big) \\ &= \int_u^\infty (T_j^* - u) g(T_j^*) \mathrm{d}T_j^* + \int_t^u (u - T_j^*) g(T_j^*) \mathrm{d}T_j^*. \end{aligned} \tag{2.2}$$

The posterior expected loss in (2.2) attains its minimum at $u = \text{median}_g(T_j^*)$, that is, the median time of GR. It can also be expressed as $\pi_j^{-1}(0.5 \mid t, s)$, where $\pi_j^{-1}(\cdot)$ is the inverse of dynamic survival probability $\pi_j(u \mid t, s)$ of patient j (Rizopoulos, 2011). It is given by:

$$\pi_j(u \mid t, s) = \text{Pr}\Big\{T_j^* \geq u \mid T_j^* > t, \mathcal{Y}_j(s), D_n\Big\}, \quad u \geq t.$$

Even though $E_g(T_j^*)$ or $\text{median}_g(T_j^*)$ may be obvious choices from a statistical perspective, from the viewpoint of doctors or patients, it could be more intuitive to make the decision for the next biopsy by placing a cutoff $1-\kappa$, where $0 \leq \kappa \leq 1$, on the dynamic incidence/risk of GR. This approach would be successful if κ can sufficiently well differentiate between patients who will obtain GR in a given period of time versus others. This approach is also useful when patients are apprehensive about delaying biopsies beyond a certain risk cutoff. Thus, a biopsy can be scheduled at a time point u such that the dynamic risk of GR is higher than a certain threshold $1-\kappa$ beyond u. To this end, the posterior expected loss for the following multilinear loss function can be minimized to find the optimal u:

$$L_{k_1,k_2}(T_j^*, u) = \begin{cases} k_2(T_j^* - u), k_2 > 0 & \text{if } T_j^* > u, \\ k_1(u - T_j^*), k_1 > 0 & \text{otherwise,} \end{cases}$$

where k_1, k_2 are constants parameterizing the loss function. The posterior expected loss $E_g\Big\{L_{k_1,k_2}(T_j^*, u)\Big\}$ obtains its minimum at $u = \pi_j^{-1}\Big\{k_1/(k_1+k_2) \mid t, s\Big\}$ (Robert, 2007). The choice of the two constants k_1 and k_2 is equivalent to the choice of $\kappa = k_1/(k_1+k_2)$.

In practice, for some patients, we may not have sufficient information to estimate their PSA profile accurately. The resulting high variance of $g(T_j^*)$ could lead to a mean (or median) time of GR, which overshoots the true T_j^* by a big margin. In such cases, the approach based on the dynamic risk of GR with smaller risk thresholds is more risk-averse. It thus could be more robust to large overshooting margins. This consideration leads us to a hybrid approach, namely, to select u using the dynamic risk of GR based approach when the spread of $g(T_j^*)$ is large, while using $E_g(T_j^*)$ or $\text{median}_g(T_j^*)$ when the spread of $g(T_j^*)$ is small. What constitutes a large spread will be application-specific. In PRIAS, within the first ten years, the maximum possible delay in detection of GR is three years. Thus we propose that if the difference between the 0.025 quantile of $g(T_j^*)$, and $E_g(T_j^*)$ or median$_g(T_j^*)$ is more than three years, then proposals based on the dynamic risk of GR be used instead.

2.3.3 Estimation

Since there is no closed form solution available for $E_g(T_j^*)$, for its estimation we utilize the following relationship between $E_g(T_j^*)$ and $\pi_j(u \mid t, s)$:

$$E_g(T_j^*) = t + \int_t^\infty \pi_j(u \mid t, s) \mathrm{d}u. \tag{2.3}$$

However, as mentioned earlier, selection of the optimal biopsy time based on $E_g(T_j^*)$ alone will not be practically useful when the var$_g(T_j^*)$ is large, which is given by:

$$\text{var}_g(T_j^*) = 2\int_t^\infty (u-t)\pi_j(u \mid t, s)\mathrm{d}u - \left\{\int_t^\infty \pi_j(u \mid t, s)\mathrm{d}u\right\}^2. \tag{2.4}$$

Since there is no closed form solution available for the integrals in (2.3) and (2.4), we approximate them using Gauss-Kronrod quadrature. The variance depends both on the last biopsy time t and the PSA history $\mathcal{Y}_j(s)$, as demonstrated in Section 2.5.2.

For schedules based on the dynamic risk of GR, the choice of threshold κ has important consequences because it dictates the timing of biopsies. Often it may depend on the amount of risk that is acceptable to the patient (if the maximum acceptable risk is 5%, $\kappa = 0.95$). When κ cannot be chosen based on the input of the patients, we propose to automate its choice. More specifically, given the time t of the latest biopsy, we propose to choose a κ for which a binary classification accuracy measure (López-Ratón et al., 2014), discriminating between cases (patients who experience GR) and controls, is maximized. In joint models, a patient j is predicted to be a case in the time window Δt if $\pi_j(t + \Delta t \mid t, s) \leq \kappa$, or a control if $\pi_j(t + \Delta t \mid t, s) >$

κ (Rizopoulos, 2016; Rizopoulos et al., 2017). We choose Δt to be one year. This is because, in AS programs at any point in time, it is of interest to identify and provide extra attention to patients who may obtain GR in the next one year. As for the choice of the binary classification accuracy measure, we chose F_1 score since it is in line with our goal to focus on potential cases in time window Δt. The F_1 score combines both sensitivity and positive predictive value (PPV) and is defined as:

$$\begin{aligned}
\mathsf{F}_1(t, \Delta t, s, \kappa) &= 2\frac{\mathsf{TPR}(t, \Delta t, s, \kappa)\ \mathsf{PPV}(t, \Delta t, s, \kappa)}{\mathsf{TPR}(t, \Delta t, s, \kappa) + \mathsf{PPV}(t, \Delta t, s, \kappa)},\\
\mathsf{TPR}(t, \Delta t, s, \kappa) &= \mathsf{Pr}\Big\{\pi_j(t + \Delta t \mid t, s) \leq \kappa \mid t < T_j^* \leq t + \Delta t\Big\},\\
\mathsf{PPV}(t, \Delta t, s, \kappa) &= \mathsf{Pr}\Big\{t < T_j^* \leq t + \Delta t \mid \pi_j(t + \Delta t \mid t, s) \leq \kappa\Big\},
\end{aligned}$$

where $\mathsf{TPR}(\cdot)$ and $\mathsf{PPV}(\cdot)$ denote time-dependent true positive rate (sensitivity) and positive predictive value (precision), respectively. The estimation for both is similar to the estimation of $\mathsf{AUC}(t, \Delta t, s)$ given by Rizopoulos et al. (2017). Since a high F_1 score is desired, the corresponding value of κ is $\arg\max_\kappa \mathsf{F}_1(t, \Delta t, s, \kappa)$. We compute the latter using a grid search approach. That is, first, the F_1 score is computed using the available dataset over a fine grid of κ values between 0 and 1, and then κ corresponding to the highest F_1 score is chosen. Furthermore, in this paper, we use κ chosen only based on the F_1 score.

2.3.4 Algorithm

When a biopsy gets scheduled at a time $u < T_j^*$, then GR is not detected at u, and at least one more biopsy is required at an optimal time $u^{new} >$ max(u, s). This process is repeated until GR is detected. To aid in medical decision making, we elucidate this process via an algorithm in Figure 2.1. AS programs strongly advise that two biopsies have a gap of at least one year. Thus, when $u - t < 1$, the algorithm postpones u to $t + 1$ because it is the time nearest to u, at which the one-year gap condition is satisfied.

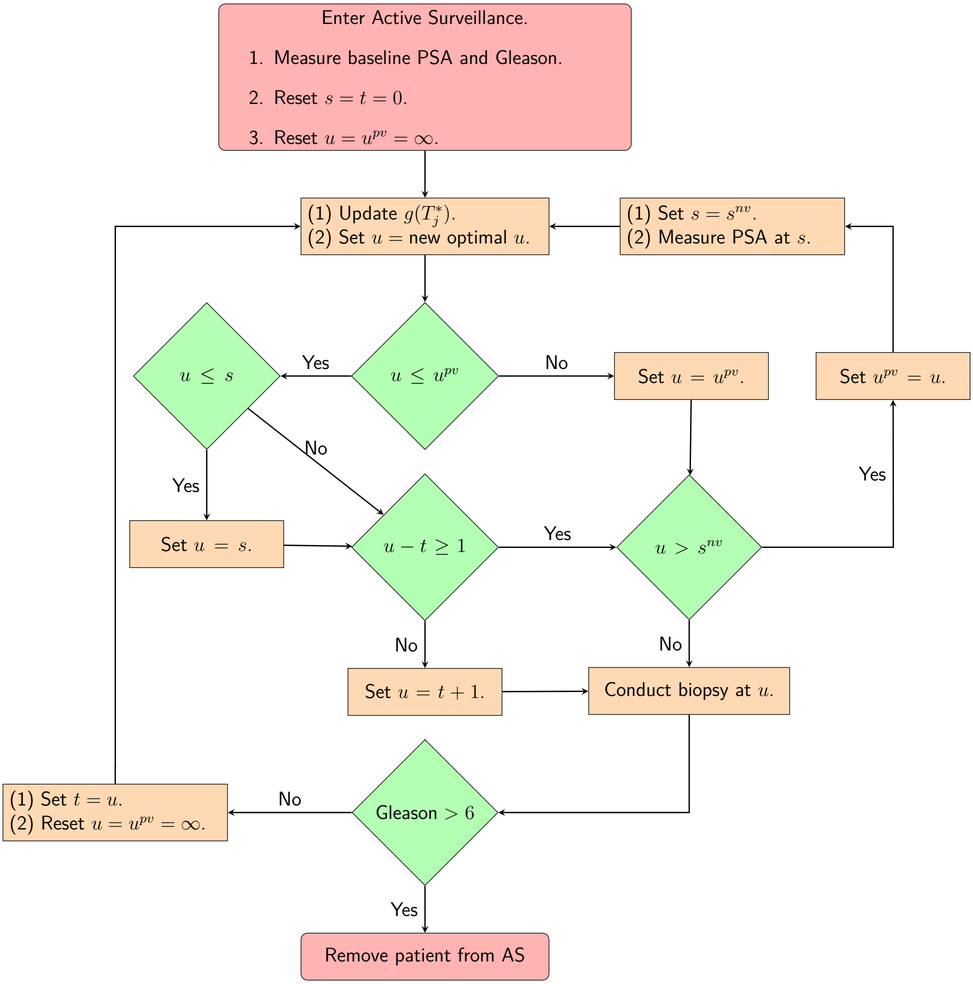

Figure 2.1: **Algorithm for creating a personalized schedule** for patient j. The time of the latest biopsy is denoted by t. The time of the latest available PSA measurement is denoted by s. The proposed personalized time of biopsy is denoted by u. The time at which a repeat biopsy was proposed on the last visit to the hospital is denoted by u^{pv}. The time of the next visit for the measurement of PSA is denoted by s^{nv}.

2.4 Evaluation of Schedules

In order to compare various schedules of biopsies, we require measures of their efficacy. We propose to use two measures, namely the number of biopsies (burden) $N_j^S \geq 1$ a schedule S conducts for the j-th patient to detect GR, and the offset $O_j^S \geq 0$ by which it overshoots T_j^*. The offset O_j^S is defined as $O_j^S = T^S_{jN_j^S} - T_j^*$, where $T^S_{jN_j^S} \geq T_j^*$ is the time at which GR is detected. Our interest lies in the joint distribution $p(N_j^S, O_j^S)$ of the number of biopsies and the offset. The least burdensome scenario is when $N_j^S = 1$ and $O^S = 0$. Hence, realistically we should select a schedule with a low mean number of biopsies $E(N_j^S)$ as well a low mean offset $E(O_j^S)$. It is also desired that a schedule has a low variance for both the number of biopsies var(N_j^S) and offset var(O_j^S) so that the schedule works similarly for most patients.

2.4.1 Choosing a Schedule

Given the multiple schedules of biopsies, it is of clinical interest to choose a suitable schedule. Using principles from compound optimal designs (Läuter, 1976) we propose to choose a schedule S which minimizes a loss function of the following form:

$$L(S) = \sum_{r=1}^{R} \eta_r \mathcal{R}_r(N_j^S), \tag{2.5}$$

where $\mathcal{R}_r(\cdot)$ is a function of either N_j^S or O_j^S (for brevity, only N_j^S is used in the equation above). Some examples of $\mathcal{R}_r(\cdot)$ are mean, median, variance and quantile function. Constants $\eta_1, \dots, \eta_R$, where $0 \leq \eta_r \leq 1$ and $\sum_{r=1}^R \eta_r = 1$, are weights to differentially weigh-in the contribution of each of the R criteria. An example loss function is:

$$L(S) = \eta_1 E(N_j^S) + \eta_2 E(O_j^S). \tag{2.6}$$

The choice of η_1 and η_2 is not easy, because the burden of a biopsy cannot be compared to a unit increase in offset easily. To obviate this problem we utilize the equivalence between compound and constrained optimal designs (Cook and Wong, 1994). More specifically, it can be shown that for any η_1 and η_2 there exists a constant $C > 0$ for which minimization of the loss function in (2.6) is equivalent to minimization of the loss function subject to the constraint that $E(N_j^S) < C$. That is, a schedule which conducts at most C biopsies on average and detects GR earliest should be chosen. The choice of C could be based on the number of biopsies a patient is willing to undergo. In the more generic case in (2.5), a schedule can be chosen by minimizing $\mathcal{R}_R(\cdot)$ under the constraint $\mathcal{R}_r(\cdot) < C_r; r = 1, \ldots, R-1$.

2.5 Demonstration of Personalized Schedules

To demonstrate the personalized schedules, we apply them to the patients enrolled in the PRIAS study. To this end, we divide the PRIAS dataset into a training part (5264 patients) and a demonstration part (three patients). We fit a joint model to the training dataset and then use it to create schedules for the demonstration patients. We fit the joint model using the R package **JMbayes** (Rizopoulos, 2016), which uses the Bayesian approach for parameter estimation.

2.5.1 Fitting the Joint Model to the PRIAS Dataset

For each of the PRIAS patients, we know their age at the time of inclusion in AS, PSA history and the time interval in which GR is detected. For the longitudinal analysis of PSA we use $\log_2(\text{PSA} + 1)$ measurements instead of the raw data (Lin et al., 2000; Pearson et al., 1994). The longitudinal sub-model of the joint model we fit is given by:

$$\begin{aligned}\log_2(\text{PSA}_i + 1)(t) &= \beta_0 + \beta_1(\text{Age}_i - 70) + \beta_2(\text{Age}_i - 70)^2 \\ &\quad + \sum_{k=1}^{4} \beta_{k+2} B_k(t, \mathcal{K}) \\ &\quad + b_{i0} + b_{i1} B_7(t, 0.1) + b_{i2} B_8(t, 0.1) + \varepsilon_i(t), \qquad (2.7)\end{aligned}$$

where $B_k(t, \mathcal{K})$ denotes the k-th basis function of a B-spline with three internal knots at $\mathcal{K} = \{0.1, 0.5, 4\}$ years, and boundary knots at zero and seven (0.99 quantile of the observed follow-up times) years. The spline for the random effects consists of one internal knot at 0.1 years and boundary knots at zero and seven years. For the relative risk sub-model the hazard function we fit is given by:

$$\begin{aligned} h_i(t) = h_0(t) \exp\Big\{&\gamma_1(\text{Age}_i - 70) + \gamma_2(\text{Age}_i - 70)^2 \\ &+ \alpha_1 m_i(t) + \alpha_2 m'_i(t)\Big\}, \end{aligned} \tag{2.8}$$

where α_1 and α_2 are measures of strength of the association between hazard of GR and $\log_2(\text{PSA}_i + 1)$ value $m_i(t)$ and $\log_2(\text{PSA}_i + 1)$ velocity $m'_i(t)$, respectively.

From the fitted joint model, we found that $\log_2(\text{PSA} + 1)$ velocity and the age at the time of inclusion in AS were significantly associated with the hazard of GR. For any patient, an increase in $\log_2(\text{PSA} + 1)$ velocity from -0.06 to 0.14 (first and third quartiles of the fitted velocities, respectively) corresponds to a 2.05 fold increase in the hazard of GR. In terms of the predictive performance, we found that the area under the receiver operating characteristic curves (Rizopoulos et al., 2017) was 0.61, 0.65, and 0.59 at year one, year two, and year three of follow-up, respectively. Parameter estimates are presented in detail in Appendix 2.A.

In PRIAS, the interval $l_i < T^*_i \leq r_i$ in which GR is detected depends on the PSA-DT of the patient. However, because the parameters are estimated using a full likelihood approach (Tsiatis and Davidian, 2004), the joint model gives valid estimates for all of the parameters, under the condition that the model is correctly specified (Appendix 2.B). To this end, we performed several sensitivity analyses in our model (e.g., changing the position of the knots, etc.) to investigate the fit of the model and also the robustness of the results. In all of our attempts, the same conclusions were reached, namely that the velocity of the longitudinal outcome is more strongly associated with the hazard of GR than the value.

2.5.2 Personalized Schedules for a Demonstration Patient

We now demonstrate the functioning of the personalized schedules for the first demonstration patient. The fitted and observed $\log_2(\text{PSA} + 1)$ profile, time of latest biopsy and proposed biopsy times u for him are shown in Figure 2.2. We can see that with a consistently decreasing PSA and negative repeat biopsy between year 3 (Panel A of Figure 2.2) and year 4.5 (Panel B of Figure 2.2), the proposed time of biopsy based on the dynamic risk of GR has increased from 3.05 years ($\kappa = 0.94$) to 14.73 years ($\kappa = 0.96$) in this period. The proposed time of biopsy based on the expected time of GR has also increased from 14.53 years to 16.05 years. We can also see in Figure 2.3 that after each negative repeat biopsy, $\text{SD}(T^*_j) = \sqrt{\text{var}_g(T^*_j)}$ decreases sharply. Thus, if the expected time of GR based approach is used, then the offset O^S_j will be smaller on average for biopsies scheduled after the second repeat biopsy than those scheduled after the first repeat biopsy.

2.6 Simulation Study

In Section 2.5.2 we demonstrated that the personalized schedules, schedule future biopsies according to the historical data of each patient. However, we could not perform a full-scale comparison between personalized and PRIAS schedules, because the true time of GR was not known for the PRIAS patients. To this end, we conducted a simulation study comparing personalized schedules with PRIAS and annual schedule, whose details are presented next.

2.6.1 Simulation Setup

The population of AS patients in this simulation study is assumed to have the same entrance criteria as that of PRIAS. The PSA and hazard of GR for these patients follow a joint model of the form postulated in Section 2.5.1, with the only change that $\log_2$ PSA levels are used as the outcome. The

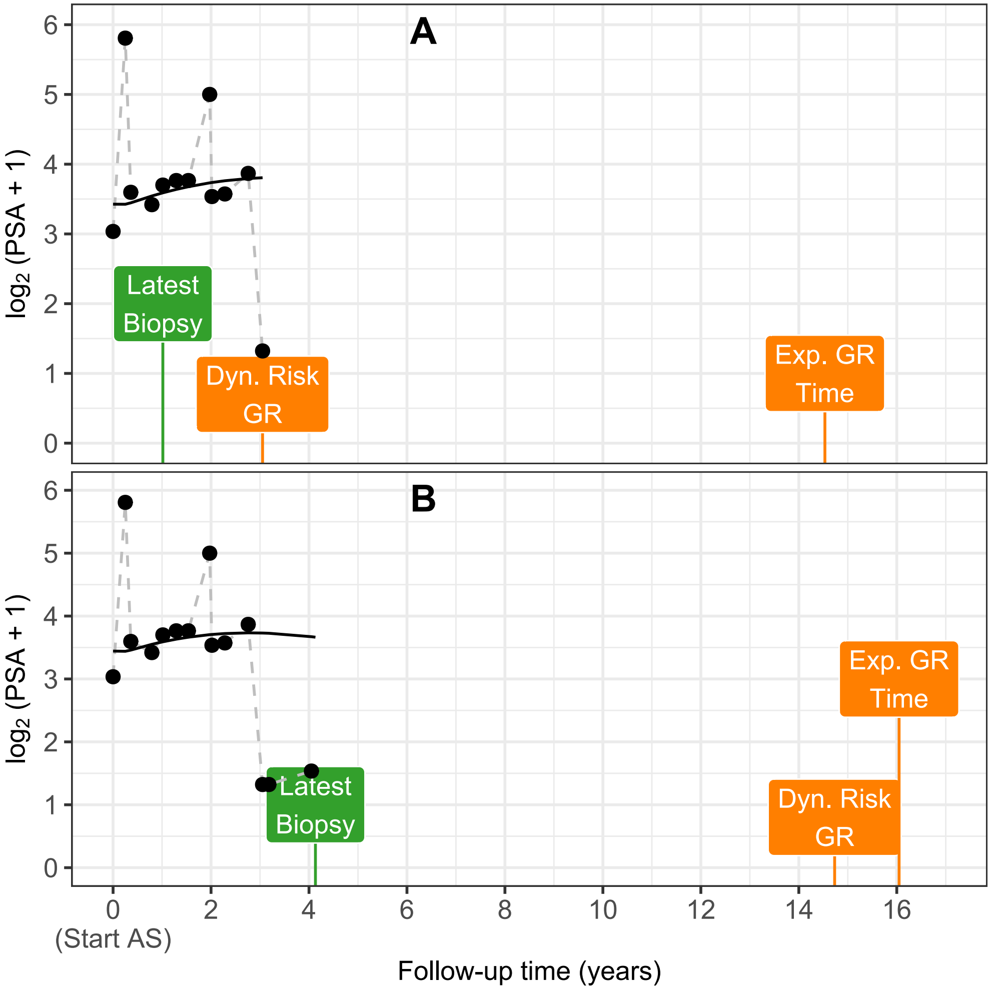

Figure 2.2: **Demonstration of personalized schedules at two different visits**. Panels A and B show fitted (solid black line) versus observed $\log_2(\text{PSA}+1)$ profile, time of latest biopsy, and personalized time of biopsies for the first demonstration patient. **Types of personalized schedules**: Exp. GR Time schedules a biopsy at the expected time of GR (Gleason reclassification) and Dyn. Risk GR schedules a biopsy when the dynamic risk of GR is higher than a certain threshold.

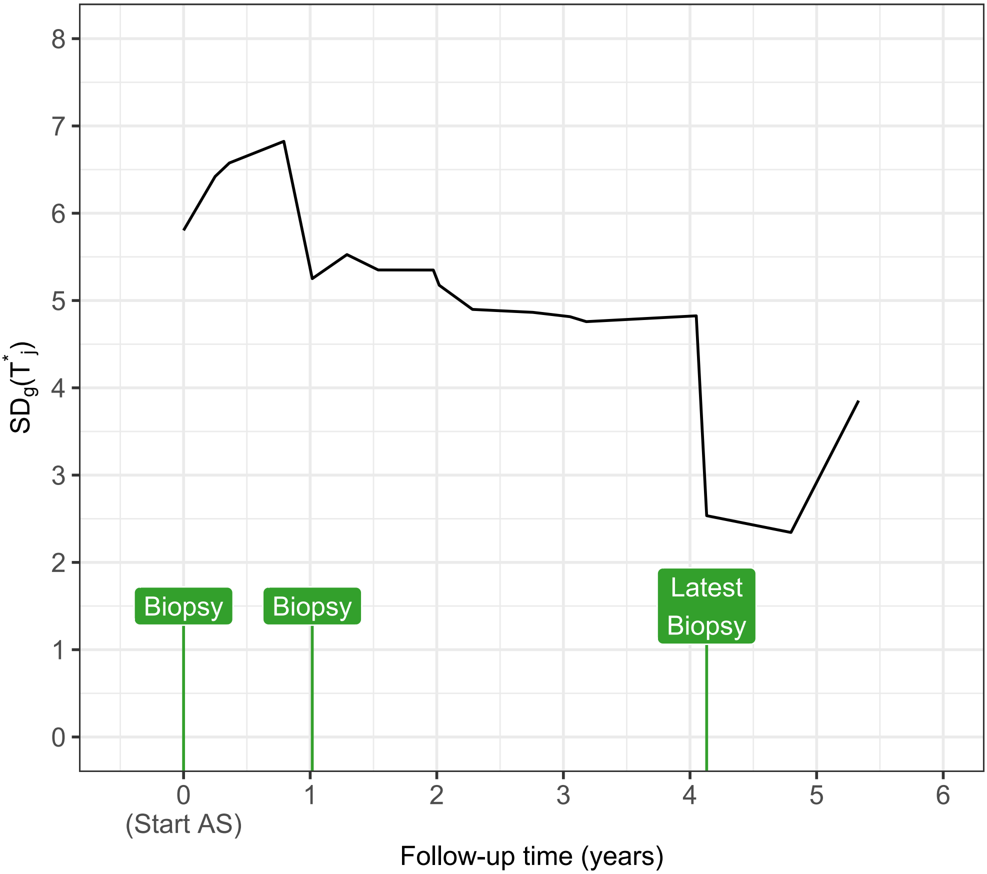

Figure 2.3: History of repeat biopsies and standard deviation $\text{SD}_g(T_j^*) = \sqrt{\text{var}_g(T_j^*)}$ of the posterior predictive distribution of time of Gleason reclassification (see Section 2.3.1), over time, for the first demonstration patient.

population joint model parameters are equal to the posterior mean of parameters estimated from the corresponding joint model fitted to the PRIAS dataset. We intend to test the efficacy of different schedules for a population which has patients with both faster as well as slowly-progressing PCa. This rate of progression is not only manifested via PSA profiles but also via the baseline hazard. We assume that there are three equal sized subgroups G_1, G_2 and G_3 of patients in the population, each with a baseline hazard from a Weibull distribution, with the following shape and scale parameters (k, λ): $(1.5, 4)$, $(3, 5)$ and $(4.5, 6)$ for G_1, G_2 and G_3, respectively. The effect of these parameters is that the mean GR time is lowest in G_1 (fast PCa progression) and highest in G_3 (slow PCa progression).

From this population, we have sampled 500 datasets with 1000 patients each. We generate a true GR time for each of the patients, and then sample a set of PSA measurements at the same time points as given in PRIAS protocol (quarterly for the first two years of AS, semiannually thereafter). We then split the dataset into a training (750 patients) and a test (250 patients) part, and generate a random and non-informative censoring time for the training patients. We next fit a joint model of the specification given in (2.7) and (2.8) to each of the 500 training datasets and obtain MCMC samples from the 500 sets of the posterior distribution of the parameters. Using these fitted joint models, we obtain the posterior predictive distribution of time of GR for each of the 500×250 test patients. This distribution is further used to create personalized biopsy schedules for the test patients. For every test patient we conduct hypothetical biopsies using the following six types of schedules (abbreviated names in parenthesis): personalized schedules based on expected time of GR (Exp. GR Time) and median time of GR (Med. GR Time), personalized schedules based on dynamic risk of GR (Dyn. Risk GR), a hybrid approach between median time of GR and dynamic risk of GR (Hybrid), PRIAS schedule and the annual schedule. The biopsies are conducted as per the algorithm in Figure 2.1.

To compare the aforementioned schedules we require estimates of the various measures of efficacy described in Section 2.4. To this end, for schedule S, we compute pooled estimates of mean offset $E(O_j^S)$ and variance of offset var(O^S_j), as below (estimates for N^S_j are similar):

$$\widehat{E(O^S_j)} = \frac{\sum_{k=1}^{500} n_k \widehat{E(O^S_k)}}{\sum_{k=1}^{500} n_k},$$

$$\widehat{\text{var}(O^S_j)} = \frac{\sum_{k=1}^{500} (n_k - 1)\widehat{\text{var}(O^S_k)}}{\sum_{k=1}^{500} (n_k - 1)},$$

where n_k denotes the number of test patients, $\widehat{E(O^S_k)} = \sum_{l=1}^{n_k} O^S_{kl}/n_k$ is the estimated mean and $\widehat{\text{var}(O^S_k)} = \sum_{l=1}^{n_k} \left\{O^S_{kl} - \widehat{E(O^S_k)}\right\}^2/(n_k - 1)$ is the estimated variance of the offset for the k-th simulation. The offset for the l-th test patient of the k-th dataset is denoted by O^S_{kl}.

2.6.2 Results

The pooled estimates of the aforementioned measures are summarized in Table 2.1 and Table 2.2. In addition, estimated values of $E(O^S_j)$ are plotted against $E(N^S_j)$ in Figure 2.4. The figure shows that across the schedules, there is an inverse relationship between number $E(O^S_j)$ and $E(N^S_j)$. For example, the annual schedule conducts, on average, 5.2 biopsies to detect GR, which is the highest among all schedules. However, it has the least average offset of 6 months as well. On the other hand, the schedule based on the expected time of GR conducts only 1.9 biopsies on average to detect GR, the least among all schedules, but it also has the highest average offset of 15 months (similar for the median time of GR). Since the annual schedule attempts to contain the offset within a year it has the least SD$(O^S_j) = \sqrt{\text{var}(O^S_j)}$. However to achieve this, it conducts a wide range of number of biopsies from patient to patient, i.e., highest SD$(N^S_j) = \sqrt{\text{var}(N^S_j)}$. In this regard, schedules based on expected and median time of GR perform the opposite of the annual schedule.

The PRIAS schedule conducts only 0.3 biopsies less than the annual schedule, but with a higher SD(O^S_j), early detection is not always guaranteed. In comparison, the dynamic risk of GR based schedule performs slightly better

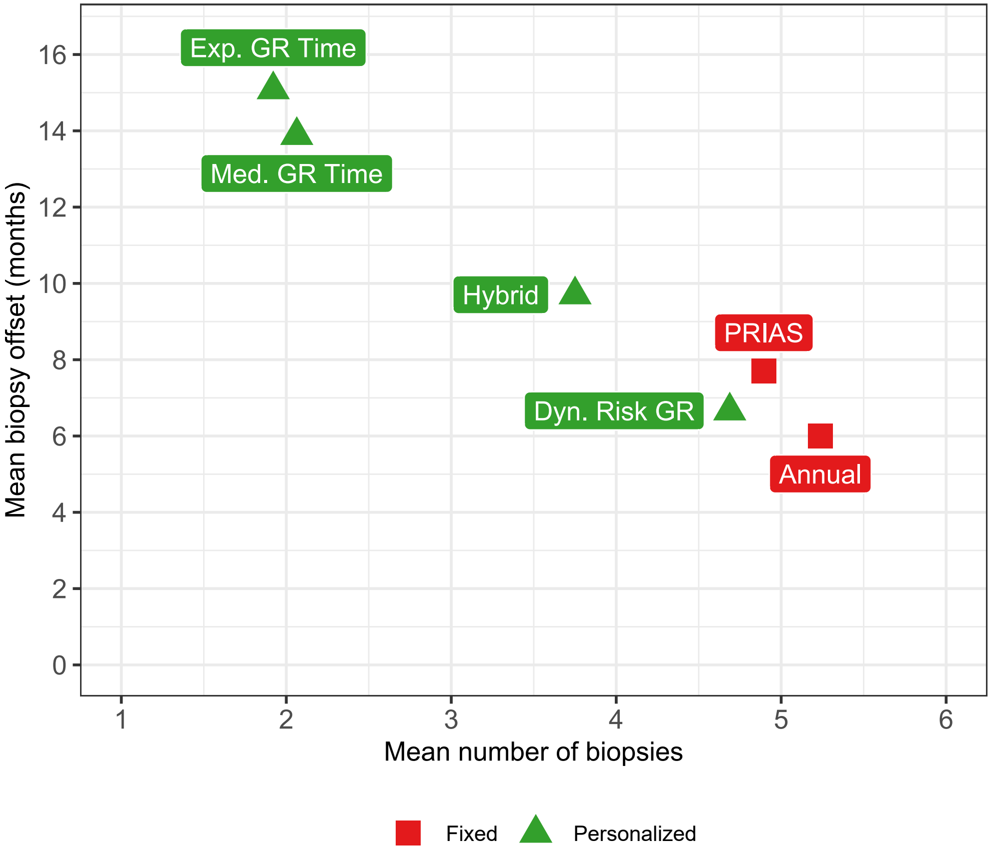

Figure 2.4: **Estimated mean number of biopsies and offset (in months)**. Biopsies are conducted until Gleason reclassification (GR) is detected. Offset is the difference in time at which GR is detected and the true time of GR. Results are based on the simulated (500 datasets) test patients. **Types of personalized schedules**: Exp. GR Time (expected time of GR), Med. GR Time (median time of GR), Dyn. Risk GR (schedules based on the dynamic risk of GR), Hybrid (a hybrid approach between Med. GR Time and Dyn. Risk GR). **Annual**: yearly biopsies. **PRIAS**: biopsies as per PRIAS protocol.

Table 2.1: **Estimated mean and standard deviation (SD), of the number of biopsies N_j^S and offset O_j^S.** Offset (in months) is defined as difference in time at which GR (Gleason reclassification) is detected and the true time of GR. Results are based on all simulated (500 datasets) test patients. **Types of personalized schedules**: Exp. GR Time (expected time of GR), Med. GR Time (median time of GR), Dyn. Risk GR (schedules based on dynamic risk of GR), Hybrid (a hybrid approach between Med. GR Time and Dyn. Risk GR). **Annual**: yearly biopsies. **PRIAS**: biopsies as per PRIAS protocol.

Schedule	$E(N_j^S)$	$E(O_j^S)$	SD(N_j^S)	SD(O_j^S)
Annual	5.24	6.01	2.53	3.46
PRIAS	4.90	7.71	2.36	6.31
Dyn. Risk GR	4.69	6.66	2.19	4.38
Hybrid	3.75	9.70	1.71	7.25
Med. GR Time	2.06	13.88	1.41	11.80
Exp. GR Time	1.92	15.08	1.19	12.11

than the PRIAS schedule in all four criteria. The hybrid approach combines the benefits of methods with low $E(N_j^S)$ and SD(N_j^S), and methods with low $E(O_j^S)$ and SD(O_j^S). It conducts 1.5 biopsies less than the annual schedule on average, and with a $E(O_j^S)$ of 9.7 months, it detects GR within a year since its occurrence. Moreover, it has both SD(N_j^S) and SD(O_j^S) comparable to PRIAS.

The performance of each schedule differs for the three subgroups G_1, G_2, and G_3. The annual schedule remains the most consistent across subgroups in terms of the offset, but it conducts two extra biopsies for the subgroup G_3 (slowly-progressing PCa) than G_1 (faster-progressing PCa). The performance of schedule based on expected time of GR is the most consistent in terms of the number of biopsies, but it detects GR a year later on average in subgroup G_1 than G_3. For the dynamic risk of GR based schedule and the hybrid schedule, the dynamics are similar to that of the annual schedule. Unlike the latter two schedules, the PRIAS schedule not only conducts more

Table 2.2: **Subgroup Estimated mean and standard deviation (SD), of the number of biopsies N_j^S and offset O_j^S**. Offset (in months) is defined as difference in time at which GR (Gleason reclassification) is detected and the true time of GR. Results based on simulated (500 datasets) test patients, with **Subgroup** G_1 and **Subgroup** G_3 having the fastest, and slowest progressing cancer patients, respectively. **Types of personalized schedules**: Exp. GR Time (expected time of GR), Med. GR Time (median time of GR), Dyn. Risk GR (schedules based on dynamic risk of GR), Hybrid (a hybrid approach between Med. GR Time and Dyn. Risk GR). **Annual**: yearly biopsies. **PRIAS**: biopsies as per PRIAS protocol.

b) Hypothetical subgroup G_1

Schedule	$E(N_j^S)$	$E(O_j^S)$	SD(N_j^S)	SD(O_j^S)
Annual	4.32	6.02	3.13	3.44
PRIAS	4.07	7.44	2.88	6.11
Dyn. Risk GR	3.85	6.75	2.69	4.44
Hybrid	3.25	10.25	2.16	8.07
Med. GR Time	1.84	20.66	1.76	14.62
Exp. GR Time	1.72	21.65	1.47	14.75

c) Hypothetical subgroup G_2

Schedule	$E(N_j^S)$	$E(O_j^S)$	SD(N_j^S)	SD(O_j^S)
Annual	5.18	5.98	2.13	3.47
PRIAS	4.85	7.70	2.00	6.29
Dyn. Risk GR	4.63	6.66	1.82	4.37
Hybrid	3.68	10.32	1.37	7.45
Med. GR Time	1.89	12.33	1.16	9.44
Exp. GR Time	1.77	13.54	0.98	9.83

d) Hypothetical subgroup G_3

Schedule	$E(N_j^S)$	$E(O_j^S)$	SD(N_j^S)	SD(O_j^S)
Annual	6.20	6.02	1.76	3.46
PRIAS	5.76	7.98	1.71	6.51
Dyn. Risk GR	5.58	6.58	1.56	4.33
Hybrid	4.32	8.55	1.26	5.91
Med. GR Time	2.45	8.70	1.15	6.32
Exp. GR Time	2.27	10.09	0.99	7.47

biopsies in G_3 than G_1 but also detects GR later in G_3 than G_1.

The choice of a suitable schedule using (2.5) depends on the chosen measure for evaluation of schedules. In this regard, the schedules we compared either have high SD(O_j^S) and low SD(N_j^S), or vice versa (Table 2.1 and Table 2.2). Thus, applying a cutoff on $E(O_j^S)$ when SD(O_j^S) is high may not be as fruitful (same for N_j^S) as applying a cutoff on SD(O_j^S) or quantile(s) of O_j^S. For example, the schedule based on the dynamic risk of GR is suitable if, on average, the least number of biopsies are to be conducted to detect GR, while simultaneously making sure that at least 90% of the patients have an average offset less than one year.

2.7 Discussion

In this paper, we presented personalized schedules based on joint models for time-to-event and longitudinal data for the surveillance of PCa patients. These schedules are dynamic, and at any given follow-up time, utilize a patient's historical PSA measurements and repeat biopsies conducted up to that time. We proposed two types of personalized schedules, namely those based on expected and median time of GR of a patient, and those based on the dynamic risk of GR. We also proposed a combination (hybrid approach) of these two approaches, which is useful in scenarios where the variance of time of GR for a patient is high. We then proposed criteria for the evaluation of various schedules and a method to select a suitable schedule.

We demonstrated the dynamic and personalized nature of our schedules using the PRIAS dataset. We observed that a recent biopsy impacts the schedules more than recent PSA measurements, which correlates with biopsies being more reliable. Since true GR time is not known for PRIAS patients, we conducted a simulation study to compare personalized schedules with PRIAS and annual schedules. The latter two schedules are already in practice. Hence it can be argued that the maximum possible offsets due to these schedules (one and three years, respectively) are acceptable to doctors. Thus, less frequent schedules with offset under one year may reduce

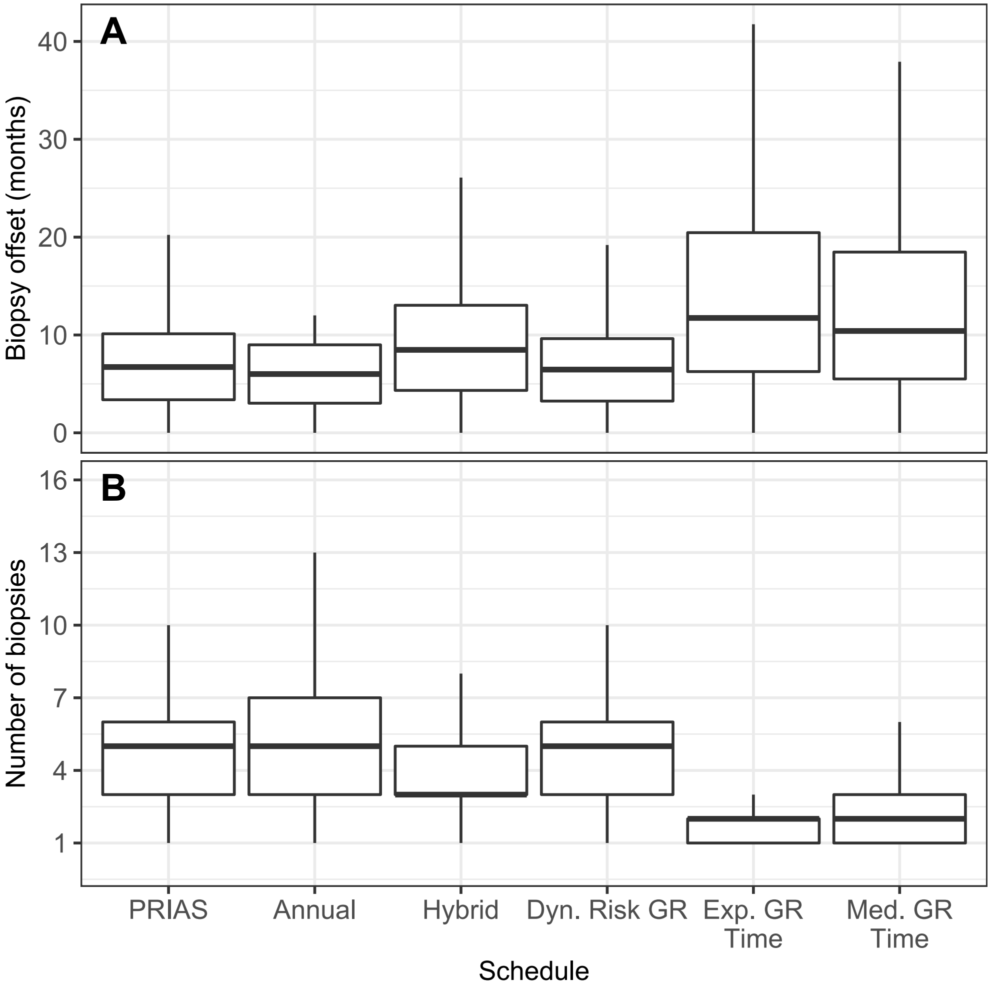

Figure 2.5: Variation in the number of biopsies and biopsy offset (difference in time at which Gleason reclassification / GR is detected and the true time of GR, in months). Results are based on the simulated (500 datasets) test patients. Biopsies are conducted until Gleason reclassification (GR) is detected. **Types of personalized schedules**: Exp. GR Time (expected time of GR), Med. GR Time (median time of GR), Dyn. Risk GR (schedules based on the dynamic risk of GR), Hybrid (a hybrid approach between Med. GR Time and Dyn. Risk GR). **Annual**: yearly biopsies. **PRIAS**: biopsies as per PRIAS protocol.

the burden of biopsies while simultaneously being practical. For example, for slowly-progressing patients in our simulation study, we observed that the schedule based on the expected time of GR conducts on average two biopsies and has an average offset of 10 months. In comparison, the annual schedule conducts six biopsies on average and gives an offset smaller by only four months, making the personalized schedule a suitable alternative. For high-risk patients, however, early detection (annual or PRIAS schedule) may be necessary, given the rapidness of progression. When it is not known in advance, if a patient will have a fast or slow-progression of PCa, the hybrid approach may be used. It conducts one biopsy less than the annual schedule in faster-progressing PCa patients and has an average offset of 10.25 months. For slowly-progressing PCa patients, it conducts two biopsies less than the annual schedule and has an average offset of 8.55 months.

More personalized schedules can be added to the current set, using loss functions that asymmetrically penalize overshooting/undershooting the target GR time. For dynamic risk of GR based schedules, more simulations are required to compare data-driven κ values (e.g., F_1 score), with κ chosen using decision analytic approaches such as the net benefit measure (Vickers and Elkin, 2006), and with various fixed κ values used by doctors in practice. In general, the Gleason scores are susceptible to inter-observer variation (Carlson et al., 1998). Schedules that account for error in the measurement of time of GR will be interesting to investigate further (Coley et al., 2017). Lastly, there is potential for including diagnostic information from magnetic resonance imaging (MRI) or DRE. When such information is not continuous, our proposed methodology can be easily extended by utilizing the framework of generalized linear mixed models.

Acknowledgements The first and last authors would like to acknowledge support by the Netherlands Organization for Scientific Research's VIDI grant nr. 016.146.301, and Erasmus MC funding. The authors also thank the Erasmus MC Cancer Computational Biology Center for giving access to their IT-infrastructure and software that was used for the computations and

data analysis in this study. Lastly, we thank Frank-Jan H. Drost from the Department of Urology, Erasmus University Medical Center, for helping us in accessing the PRIAS data set.

Appendix

2.A Parameter Estimation

We estimate parameters of the joint model using Markov chain Monte Carlo (MCMC) methods under the Bayesian framework. Let $\boldsymbol{\theta}$ denote the vector of the parameters of the joint model. The joint model postulates that given the random effects, time to GR and longitudinal responses taken over time are all mutually independent. Under this assumption the posterior distribution of the parameters is given by:

$$\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D}_n) &\propto \prod_{i=1}^{n} p(l_i, r_i, \boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\
&\propto \prod_{i=1}^{n} p(l_i, r_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}), \\
p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) &= \frac{1}{\sqrt{(2\pi)^q \text{det}(\boldsymbol{D})}} \exp(\boldsymbol{b}_i^T \boldsymbol{D}^{-1} \boldsymbol{b}_i),
\end{aligned}$$

where the likelihood contribution of longitudinal outcome conditional on random effects is:

$$\begin{aligned}
p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^{n_i}} \exp\left(- \frac{\|\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{Z}_i \boldsymbol{b}_i\|^2}{\sigma^2} \right), \\
\boldsymbol{X}_i &= \{\boldsymbol{x}_i(t_{i1})^T, \dots, \boldsymbol{x}_i(t_{in_i})^T\}^T, \\
\boldsymbol{Z}_i &= \{\boldsymbol{z}_i(t_{i1})^T, \dots, \boldsymbol{z}_i(t_{in_i})^T\}^T.
\end{aligned}$$

The likelihood contribution of the time to GR outcome is given by:

$$\begin{aligned} p(l_i, r_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \exp\Big\{ -\int_0^{l_i} h_i(s \mid \mathcal{M}_i(s), \boldsymbol{w}_i) \mathrm{d}s \Big\} \\ - \exp\Big\{ -\int_0^{r_i} h_i(s \mid \mathcal{M}_i(s), \boldsymbol{w}_i) \mathrm{d}s \Big\}. \end{aligned} \tag{2.9}$$

The integral in (2.9) does not have a closed-form solution, and therefore we use a 15-point Gauss-Kronrod quadrature rule to approximate it.

We use independent normal priors with zero mean and variance 100 for the fixed effects $\boldsymbol{\beta}$, and inverse Gamma prior with shape and rate both equal to 0.01 for the parameter σ^2. For the variance-covariance matrix $\boldsymbol{D}$ of the random effects, we take inverse Wishart prior with an identity scale matrix and degrees of freedom equal to q (number of random effects). For the relative risk model's parameters $\boldsymbol{\gamma}$ and the association parameters $\boldsymbol{\alpha}$, we use independent normal priors with zero mean and variance 100.

2.A.1 Parameter Estimates

The longitudinal evolution of $\log_2(\text{PSA}+1)$ is modeled with non-linear terms. Hence, the interpretation of the coefficients in this model is not straightforward. Instead of the parameter estimates, in Figure 2.6, we present the fitted marginal evolution of $\log_2(\text{PSA} + 1)$ over a period of 10 years for a hypothetical patient who is included in AS at the age of 70 years.

For the relative risk sub-model, the parameter estimates in Table 2.3 show that $\log_2(\text{PSA} + 1)$ velocity and the age at the time of inclusion in AS are strongly associated with the hazard of GR. For any patient, an increase in $\log_2(\text{PSA} + 1)$ velocity from -0.061 to 0.136 (first and third quartiles of the fitted velocities, respectively) corresponds to a 2.046 fold increase in the hazard of GR. An increase in age at the time of inclusion in AS from 65 years to 75 years (first and third quartiles of age in PRIAS dataset) corresponds to a 1.428 fold increase in the hazard of GR.

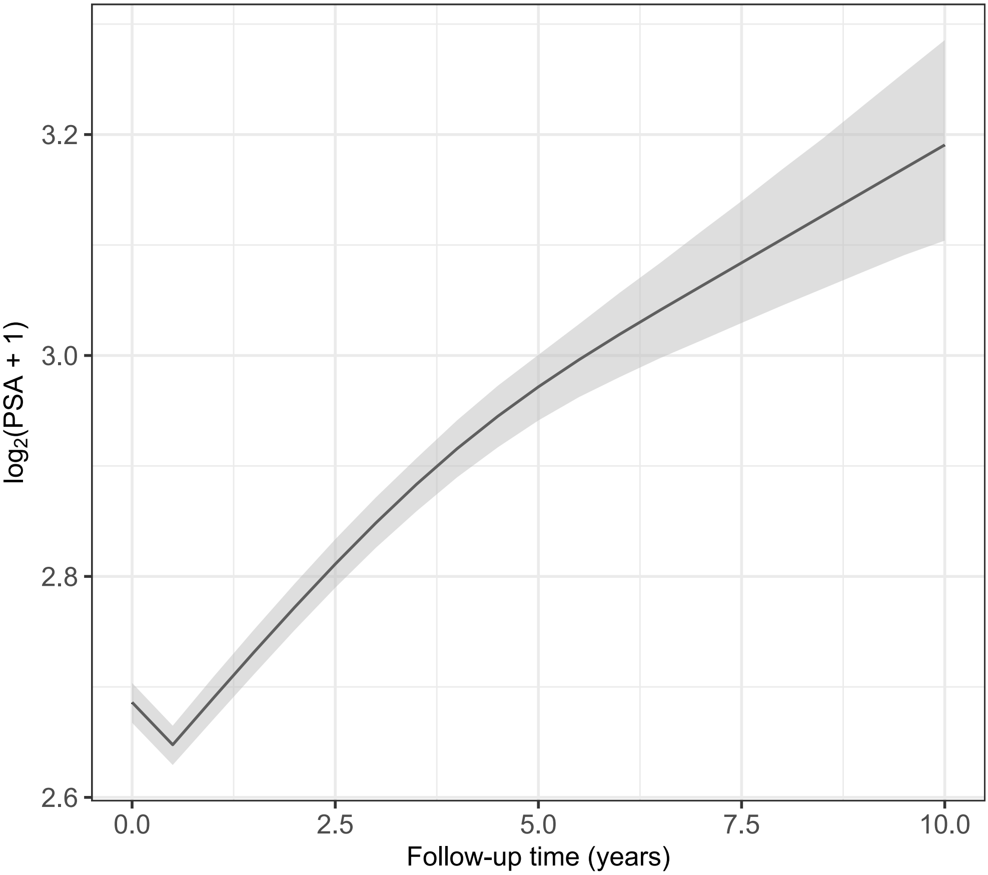

Figure 2.6: **Fitted marginal evolution** of $\log_2(\text{PSA}+1)$ measurements over a period of 10 years with 95% credible interval, for a hypothetical patient who is included in AS at the age of 70 years.

Table 2.3: **Parameters of the relative-risk sub-model**: Estimated mean and 95% credible interval. Age is median centered.

Variable	Mean	Std. Dev	2.5%	97.5%	P
$(\text{Age} - 70)$	0.036	0.006	0.024	0.047	<0.000
$(\text{Age} - 70)^2$	-0.001	0.001	-0.003	7.861×10^{-5}	0.084
$\log_2(\text{PSA} + 1)$	-0.084	0.080	-0.241	0.072	0.296
$\text{Slope}(\log_2(\text{PSA} + 1))$	3.580	0.403	2.815	4.373	<0.000

2.B Ascertainment Bias: PSA Doubling Time-Dependent Biopsies and Competing Events

PSA dependent interval-censored time of upgrading: The true time of upgrading T_i^* is not known for any of the patients in PRIAS. To detect upgrading, PRIAS uses a fixed schedule of biopsies wherein biopsies are conducted at year one, year four, year seven and year ten of follow-up, and every five years after that. However, PRIAS switches to a more frequent annual biopsy schedule for faster-progressing patients. These are patients with PSA doubling time (PSA-DT) between 0 and 10 years, which is measured as the inverse of the slope of the regression line through the base two logarithm of PSA values. Thus, the interval $l_i < T_i^* \leq r_i$ in which upgrading is detected depends on the observed PSA values.

Competing events: The primary event of interest in this paper is upgrading observed via a positive biopsy. There are three types of competing events, namely death, removal of patients from AS on the basis of their observed DRE and PSA measurements, watchful-waiting, and loss to follow-up of patients because of patient anxiety or unknown reasons.

The number of patients obtaining the event death is small compared to the number of patients who obtain the primary event upgrading. Hence in this paper, considering death as non-informative censoring may be viable.

We also consider the loss to follow-up as non-informative censoring, which may not always be true. This is especially the case when the reason for loss to follow-up is unknown. However, when the reason for loss to follow-up is patient anxiety, it is often on the basis of their observed results. Given the large number of loss to follow-up patients, considering these patients as censored is a limitation of our work. However, the problem of the unknown reason for dropout is not specific to only our model. For the remaining patients who are removed from AS on the basis of their observed longitudinal data (e.g., treatment, watchful-waiting), in the next paragraph, we show that the removal of these patients is non-informative about the parameters of the model for the true time of upgrading.

Given the aforementioned issues of PSA dependent interval censoring and removal of patients on the basis of their observed longitudinal data is natural to question in this scenario if the parameters of the joint model are affected by these two. However, because the parameters of the joint model are estimated using a full likelihood approach (Tsiatis and Davidian, 2004), the joint model allows the schedule of biopsies, as well as censoring to depend upon the observed PSA measurements (e.g., via PSA-DT), under the condition that the model is correctly specified. To show this, consider the following full general specification of the joint model that we use. Let $\boldsymbol{y}_i$ denote the observed PSA measurements for the i-th patient, and l_i, r_i denote the two time points of the interval in which upgrading occurs for the i-th patient. In addition, let T_i^S and $\mathcal{V}_i$ denote the schedule of biopsies, and the schedule PSA measurements, respectively. Let G_i^* denote the time of removal from AS without observing upgrading. Under the assumption that $T_i^S, G_i^*, \mathcal{V}_i$ may depend upon only the observed data $\boldsymbol{y}_i$, the joint likelihood of the various processes is given by:

$$p(\boldsymbol{y}_i, l_i, r_i, T_i^S, G_i^*, \mathcal{V}_i \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = p(\boldsymbol{y}_i, l_i, r_i \mid \boldsymbol{\theta}) \times p(T_i^S, G_i^*, \mathcal{V}_i \mid \boldsymbol{y}_i, \boldsymbol{\psi}).$$

where, $\boldsymbol{\psi}$ is the vector of parameters for the processes $T_i^S, G_i^*, \mathcal{V}_i$. From this decomposition, we can see that even if the processes $T_i^S, G_i^*, \mathcal{V}_i$ may be determined from $\boldsymbol{y}_i$, if we are interested in the parameters $\boldsymbol{\theta}$ of the joint distribution of longitudinal and event outcomes, we can maximize the likelihood based on the first term and ignore the second term. In other words, the second term will not carry information for $\boldsymbol{\theta}$. Lastly, since we use a full likelihood approach with an interval censoring specification, the estimates that we obtain are consistent and asymptotically unbiased (Gentleman and Geyer, 1994), despite the interval censoring observed.

We also demonstrate the validity of our argument via a simulated dataset of 750 patients. The true event times T_i^* for these patients were generated using parameters from a joint model fitted to the PRIAS dataset (with the only change that $\log_2$ PSA levels are used as the outcome). However, this joint model did not include the association between the velocity of log PSA values and the hazard of GR. That is, the hazard of GR $h_i(t)$ at any time t was dependent only on the underlying $\log_2$ PSA value $m_i(t)$ at that time. Furthermore, for these patients, we used the schedule of PRIAS to generate the interval $l_i \leq T_i^* \leq r_i$ in which GR is detected. Thus the observed data for i-th patient is $\{\boldsymbol{y}_i, l_i, r_i\}$. Our aim is to show that if there is no association between $h_i(t)$ and velocity of log PSA value $m_i'(t)$, then even though the biopsy schedule depends on PSA-DT (which is a crude measure of PSA velocity), a joint model fitted with both value and velocity associations will have an insignificant velocity association. In the fitted joint model, we found the value association (95% credible interval in brackets) to be 0.182 [0.090, 0.274], and the velocity association to be -0.001 [-0.295, 0.254]. That is, even though the schedule of biopsies depended upon observed PSA values, it did not lead to a spurious velocity association.

2.C Source Code

The source code for fitting the joint model is available at `https://raw.githubusercontent.com/anirudhtomer/prias/master/src/chapter2_biometricspaper/Gleason%20as%20event/log2psaplus1_and_pluspt1.R`.

The code generating the simulation population is available at `https://github.com/anirudhtomer/prias/blob/master/src/chapter2_biometricspape` simulation_study/SimulateJM.R.

The code for scheduling biopsies using fixed schedules and utility functions is available at https://github.com/anirudhtomer/prias/blob/master/src/chapter2_biometricspaper/simulation_study/nbAndOffset.R.

2.4 References

Akhavan-Tabatabaei, R., Sánchez, D. M., and Yeung, T. G. (2017). A Markov decision process model for cervical cancer screening policies in Colombia. *Medical Decision Making*, 37(2):196–211.

Ayer, T., Alagoz, O., and Stout, N. K. (2012). A POMDP approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034.

Bebu, I. and Lachin, J. M. (2018). Optimal screening schedules for disease progression with application to diabetic retinopathy. *Biostatistics*, 19(1):1–13.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Bokhorst, L. P., Alberts, A. R., Rannikko, A., Valdagni, R., Pickles, T., Kakehi, Y., Bangma, C. H., Roobol, M. J., and PRIAS study group (2015). Compliance rates with the Prostate Cancer Research International Active Surveillance (PRIAS) protocol and disease reclassification in noncompliers. *European Urology*, 68(5):814–821.

Brown, E. R. (2009). Assessing the association between trends in a biomarker and risk of event with an application in pediatric HIV/AIDS. *The Annals of Applied Statistics*, 3(3):1163–1182.

Carlson, G. D., Calvanese, C. B., Kahane, H., and Epstein, J. I. (1998). Accuracy of biopsy Gleason scores from a large uropathology laboratory: use of a diagnostic protocol to minimize observer variability. *Urology*, 51(4):525–529.

Coley, R. Y., Zeger, S. L., Mamawala, M., Pienta, K. J., and Carter, H. B. (2017). Prediction of the pathologic Gleason score to inform a personalized management program for prostate cancer. *European Urology*, 72(1):135–141.

Cook, R. D. and Wong, W. K. (1994). On the equivalence of constrained and compound optimal designs. *Journal of the American Statistical Association*, 89(426):687–692.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Erenay, F. S., Alagoz, O., and Said, A. (2014). Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufacturing & Service Operations Management*, 16(3):381–400.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, 81(3):618–623.

Läuter, E. (1976). Optimal multipurpose designs for regression models. *Mathematische Operationsforschung und Statistik*, 7(1):51–68.

Lin, H., McCulloch, C. E., Turnbull, B. W., Slate, E. H., and Clark, L. C. (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine*, 19(10):1303–1318.

Loeb, S., Vellekoop, A., Ahmed, H. U., Catto, J., Emberton, M., Nam, R., Rosario, D. J., Scattoni, V., and Lotan, Y. (2013). Systematic review of complications of prostate biopsy. *European Urology*, 64(6):876–892.

López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C., and Gude-Sampedro, F. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, 61(8):1–36.

Pearson, J. D., Morrell, C. H., Landis, P. K., Carter, H. B., and Brant, L. J. (1994). Mixed-effects regression models for studying the natural history of prostate disease. *Statistics in Medicine*, 13(5-7):587–601.

Potosky, A. L., Miller, B. A., Albertsen, P. C., and Kramer, B. S. (1995). The role of increasing detection in the rising incidence of prostate cancer. *JAMA*, 273(7):548–552.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software*, 72(7):1–46.

Rizopoulos, D., Hatfield, L. A., Carlin, B. P., and Takkenberg, J. J. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association*, 109(508):1385–1397.

Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276.

Rizopoulos, D., Taylor, J. M. G., Van Rosmalen, J., Steyerberg, E. W., and Takkenberg, J. J. M. (2016). Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics*, 17(1):149–164.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Taylor, J. M., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T., and Sandler, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1):206–213.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, 65(2):87–108.

Tosoian, J. J., Trock, B. J., Landis, P., Feng, Z., Epstein, J. I., Partin, A. W., Walsh, P. C., and Carter, H. B. (2011). Active surveillance program for prostate cancer: an update of the Johns Hopkins experience. *Journal of Clinical Oncology*, 29(16):2185–2190.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.

Welty, C. J., Cowan, J. E., Nguyen, H., Shinohara, K., Perez, N., Greene, K. L., Chan, J. M., Meng, M. V., Simko, J. P., Cooperberg, M. R., and Carroll, P. R. (2015). Extended followup and risk factors for disease reclassification in a large active surveillance cohort for localized prostate cancer. *The Journal of Urology*, 193(3):807–811.

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., and Inman, B. A. (2012). Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management*, 14(4):529–547.

Chapter 3

Personalized Decision Making for Biopsies in Prostate Cancer Active Surveillance Programs

This chapter is based on the paper
Tomer, A., Rizopoulos, D., Nieboer, D., Drost, F.J., Roobol, M.J., and Steyerberg, E.W. (2019). Personalized decision making for biopsies in prostate cancer active surveillance programs. *Medical Decision Making*, 39(5): 499–508. doi: `https://doi.org/10.1177/0272989X19861963`

Abstract

Background. Low-risk prostate cancer patients enrolled in active surveillance programs commonly undergo biopsies for examination of cancer progression. Biopsies are conducted as per a fixed and frequent schedule (e.g., annual biopsies). Since biopsies are burdensome, patients do not always comply with the schedule, which increases the risk of delayed detection of cancer progression.

Objective. Our aim is to better balance the number of biopsies (burden) and the delay in detection of cancer progression (less is beneficial), by personalizing the decision of conducting biopsies.

Data Sources. We use patient data of the world's largest active surveillance program (PRIAS). It enrolled 5270 patients, had 866 cancer progressions, and an average of nine prostate-specific antigen (PSA) and five digital rectal examination (DRE) measurements per patient.

Methods. Using joint models for time-to-event and longitudinal data, we model the historical DRE and PSA measurements, and biopsy results of a patient at each follow-up visit. This results in a visit and patient-specific cumulative-risk of cancer progression. If this risk is above a certain threshold, we schedule a biopsy. We compare this personalized approach with the currently practiced biopsy schedules via an extensive and realistic simulation study, based on a replica of the patients from the PRIAS program.

Results. The personalized approach saved a median of six biopsies (median: 4, IQR: 2–5), compared to the annual schedule (median: 10, IQR: 3–10). However, the delay in detection of progression (years) is similar for the personalized (median: 0.7, IQR: 0.3–1.0) and the annual schedule (median: 0.5, IQR: 0.3–0.8).

Conclusions. We conclude that personalized schedules provide substantially better balance in the number of biopsies per detected progression for men with low-risk prostate cancer.

3.1 Introduction

Prostate cancer is the second most frequently diagnosed cancer in men worldwide (Torre et al., 2015). In prostate cancer screening programs, many of the diagnosed tumors are clinically insignificant/over-diagnosed (Etzioni et al., 2002). To avoid further over-treatment, patients diagnosed with low-grade prostate cancer are commonly advised to join active surveillance (AS) programs. In AS, invasive treatments such as surgery are delayed until cancer progresses. Cancer progression is routinely monitored via serum prostate-specific antigen (PSA) measurements, a protein biomarker; digital rectal examination (DRE) measurements, a measure of the size and location of the tumor; and biopsies.

While larger values for PSA and/or DRE, may indicate cancer progression, biopsies are the most reliable cancer progression examination technique used in AS. When a patient's biopsy Gleason score becomes larger than 6 (positive biopsy, cancer progression detected), AS is stopped, and the patient is advised treatment (Bokhorst et al., 2015). However, biopsies are invasive, painful, and prone to medical complications (Ehdaie et al., 2014; Fujita et al., 2009). Hence, they are conducted intermittently until a positive biopsy. Consequently, at the time of a positive biopsy, cancer progression may be observed with a delay of unknown duration. This delay is defined as the difference between the time of the positive biopsy and the unobserved true time of cancer progression. Thus, the decision to conduct biopsies requires a compromise between the burden of biopsy and the potential delay in the detection of cancer progression.

In AS, a delay in the detection of cancer progression around 12 to 14 months is assumed to be unlikely to substantially increase the risk of adverse downstream outcomes (Inoue et al., 2018; de Carvalho et al., 2017). However, for biopsies, there is little consensus on the time gap between them (Loeb et al., 2014; Bruinsma et al., 2016; Nieboer et al., 2018). Many AS programs focus on minimizing the delay in the detection of cancer progression by scheduling biopsies annually for all patients. A drawback of annual biopsies, and other currently practiced fixed/heuristic schedules (Loeb et al., 2014; Bruinsma et al., 2016; Nieboer et al., 2018), is that they ignore the large variation in the time of cancer progression of AS patients. While they may work well for patients who progress early (*fast progressing*) in AS, but for a large proportion of patients who do not progress, or progress late (*slow progressing*) in AS, many unnecessary, burdensome biopsies are scheduled. To mediate the burden between the *fast* and *slow progressing* patients, the world's largest AS program, Prostate Cancer Research International Active Surveillance, PRIAS, (Bokhorst et al., 2016), schedules annual biopsies only for patients with a low PSA doubling time (Bokhorst et al., 2015). For everyone else, PRIAS schedules biopsies at following fixed follow-up times: year one, four, seven, and ten, and every five years thereafter. Despite this effort in PRIAS, patients may get scheduled for four to ten biopsies over a period of ten years. Therefore, compliance for biopsies is low in PRIAS (Bokhorst et al., 2015). This can lead to a delay in the detection of cancer progression and reduce the effectiveness of AS.

We aim to better balance the number of biopsies (more are burdensome), and the delay in the detection of cancer progression (less is beneficial), than currently practiced schedules. We intend to achieve this by personalizing the decision to conduct biopsies (see Figure 3.1). These decisions are made at a patient's pre-scheduled follow-up visits for DRE and PSA measurements. To develop the personalized decision-making methodology, we utilize the data of the patients enrolled in the PRIAS study. We model this data and develop the personalized approach using joint models for time-to-event and longitudinal data (Tsiatis and Davidian, 2004; Rizopoulos, 2012). In order to compare the personalized approach with current schedules, we conduct an extensive simulation study based on a replica of the patients from the PRIAS program.

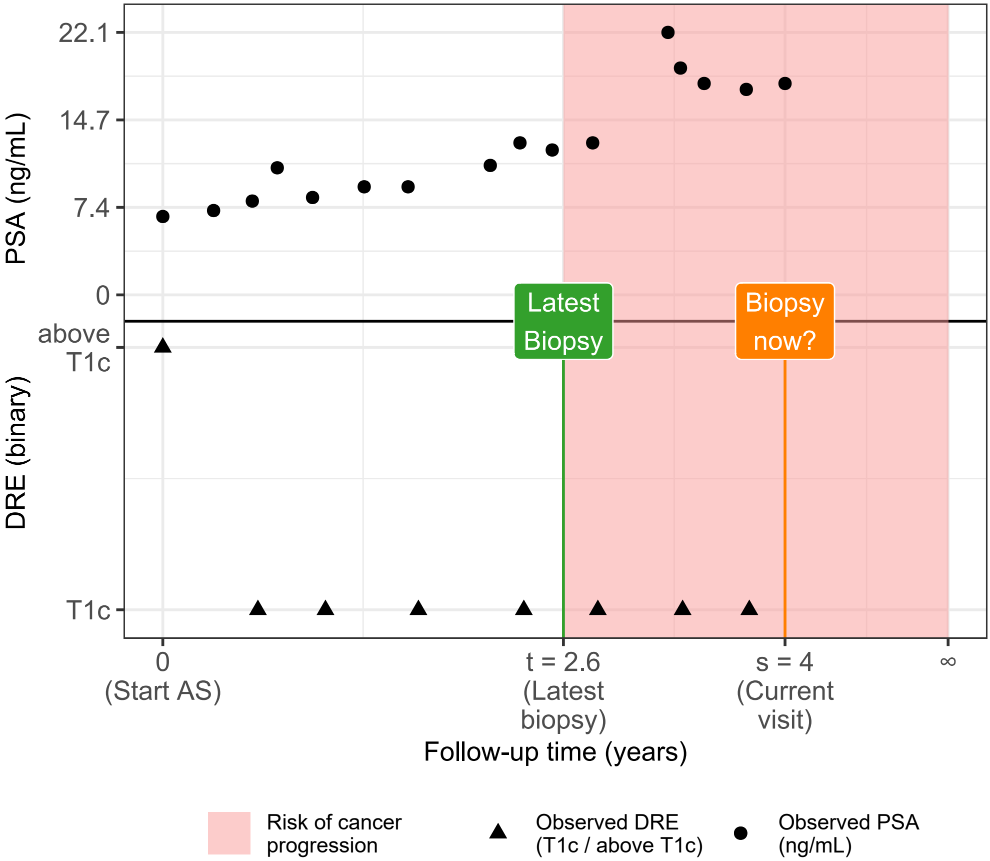

Figure 3.1: **The personalized decision-making problem:** Available data of a patient j, who had his latest negative biopsy at $t = 2.6$ years. The shaded region shows the time period in which the patient is at risk of cancer progression. His current pre-scheduled follow-up visit for measurement of DRE and PSA is at $s = 4$ years. Using his entire history of DRE $\mathcal{Y}_{dj}(s)$ and PSA $\mathcal{Y}_{pj}(s)$ measurements up to the current visit s, and the time of the latest biopsy t, we intend to make a decision on scheduling a biopsy at the current visit.

3.2 Methods

3.2.1 Study Population

To develop our methodology, we use the data (see Table 3.1) of prostate cancer patients from the world's largest AS study called PRIAS (Bokhorst et al., 2016). More than 100 medical centers from 17 countries worldwide contribute to the collection of data, utilizing a common study protocol and a web-based tool, both available at www.prias-project.org. We use data collected over ten years, between December 2006 (beginning of PRIAS study) and December 2016. The primary event of interest is cancer progression detected upon a positive biopsy. The time of cancer progression is interval-censored because biopsies are scheduled periodically. Biopsies are scheduled as per the PRIAS protocol (see Section 3.1). There are three types of competing events, namely death, removal of patients from AS based on their observed DRE and PSA measurements, and loss to follow-up. We assume these three types of events to be censored observations. However, our model allows the removal of patients to depend on observed longitudinal data and baseline covariates of the patient. Under the aforementioned assumption of censoring, Figure 3.2 shows the cumulative-risk of cancer progression over the study follow-up period.

For all patients, PSA measurements (ng/mL) are scheduled every three months for the first two years and every six months thereafter. The DRE measurements are scheduled every six months. We use the DRE measurements as DRE $=$ T1c versus DRE $>$ T1c. A DRE measurement equal to T1c (Schröder et al., 1992) indicates a clinically inapparent tumor that is not palpable or visible by imaging. In contrast, tumors with DRE $>$ T1c are palpable.

Data Accessibility: The PRIAS database is not openly accessible. However, access to the database can be requested based on a study proposal approved by the PRIAS steering committee. The website of the PRIAS program is www.prias-project.org.

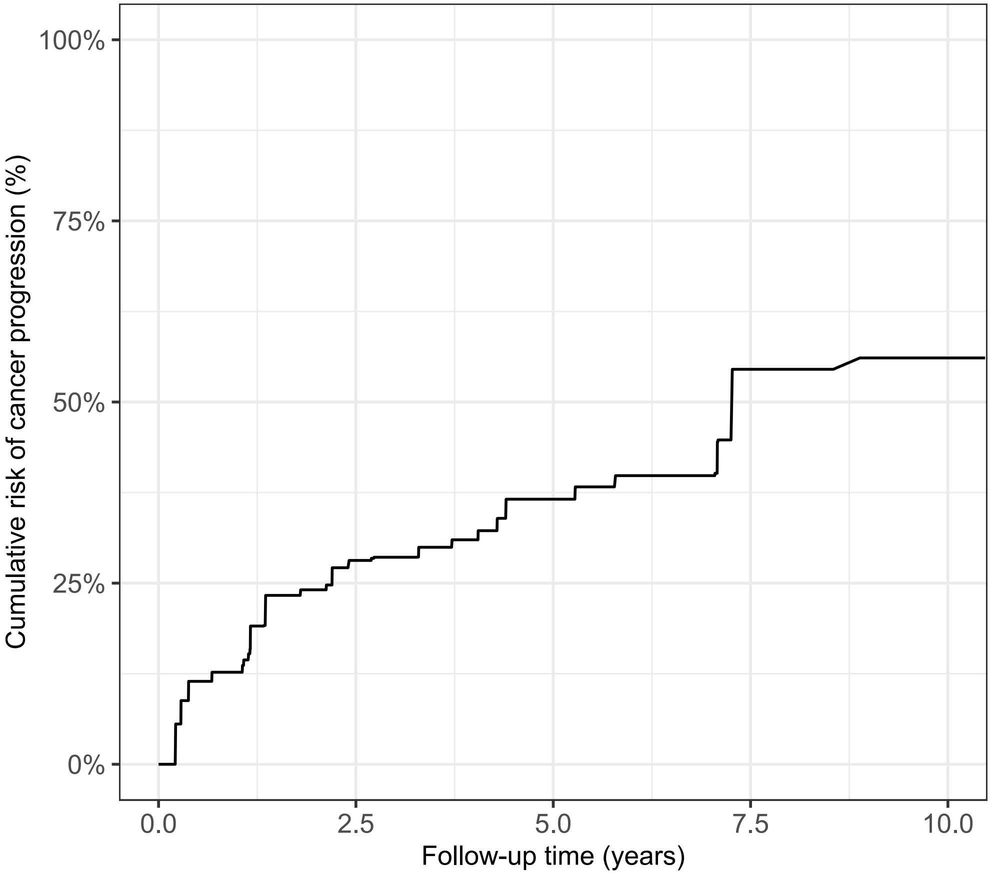

Figure 3.2: **Estimated cumulative-risk of cancer progression in AS** for patients in the Prostate Cancer Research International Active Surveillance (PRIAS) dataset. Nearly 50% patients (*slow progressing*) do not progress in the ten year follow-up period. Cumulative-risk is estimated using nonparametric maximum likelihood estimation (Turnbull, 1976), to account for interval censored cancer progression times observed in the PRIAS dataset. Censoring includes death, removal from AS on the basis of observed longitudinal data, and patient dropout.

Table 3.1: **Summary of the PRIAS dataset**. The primary event of interest is cancer progression. A DRE measurement equal to T1c (Schröder et al., 1992) indicates a clinically inapparent tumor which is not palpable or visible by imaging, while tumors with DRE $>$ T1c are palpable. IQR: interquartile range.

Data	Value
Total patients	5270
Cancer progression (primary event)	866
Loss to follow-up (anxiety or unknown)	685
Removal on the basis of PSA and DRE	464
Death (unrelated to prostate cancer)	61
Death (related to prostate cancer)	2
Median Age (years)	70 (IQR: 65–75)
Total PSA measurements	46015
Median number of PSA measurements per patient	7 (IQR: 5–12)
Median PSA value (ng/mL)	5.6 (IQR: 4.0–7.5)
Total DRE measurements	25606
Median number of DRE measurements per patient	4 (IQR: 3–7)
DRE = T1c (%)	23538/25606 (92%)

3.2.2 A Bivariate Joint Model for the Longitudinal PSA, and DRE Measurements, and Time of Cancer Progression

Let T_i^* denote the true cancer progression time of the i-th patient included in PRIAS. Since biopsies are conducted periodically, T_i^* is observed with interval censoring $l_i < T_i^* \leq r_i$. When progression is observed for the patient at his latest biopsy time r_i, then l_i denotes the time of the second latest biopsy. Otherwise, l_i denotes the time of the latest biopsy and $r_i = \infty$. Let $\boldsymbol{y}_{di}$ and $\boldsymbol{y}_{pi}$ denote his observed DRE and PSA longitudinal measurements, respectively. The observed data of all n patients is denoted by $\mathcal{D}_n = \{l_i, r_i, \boldsymbol{y}_{di}, \boldsymbol{y}_{pi}; i = 1, \ldots, n\}$.

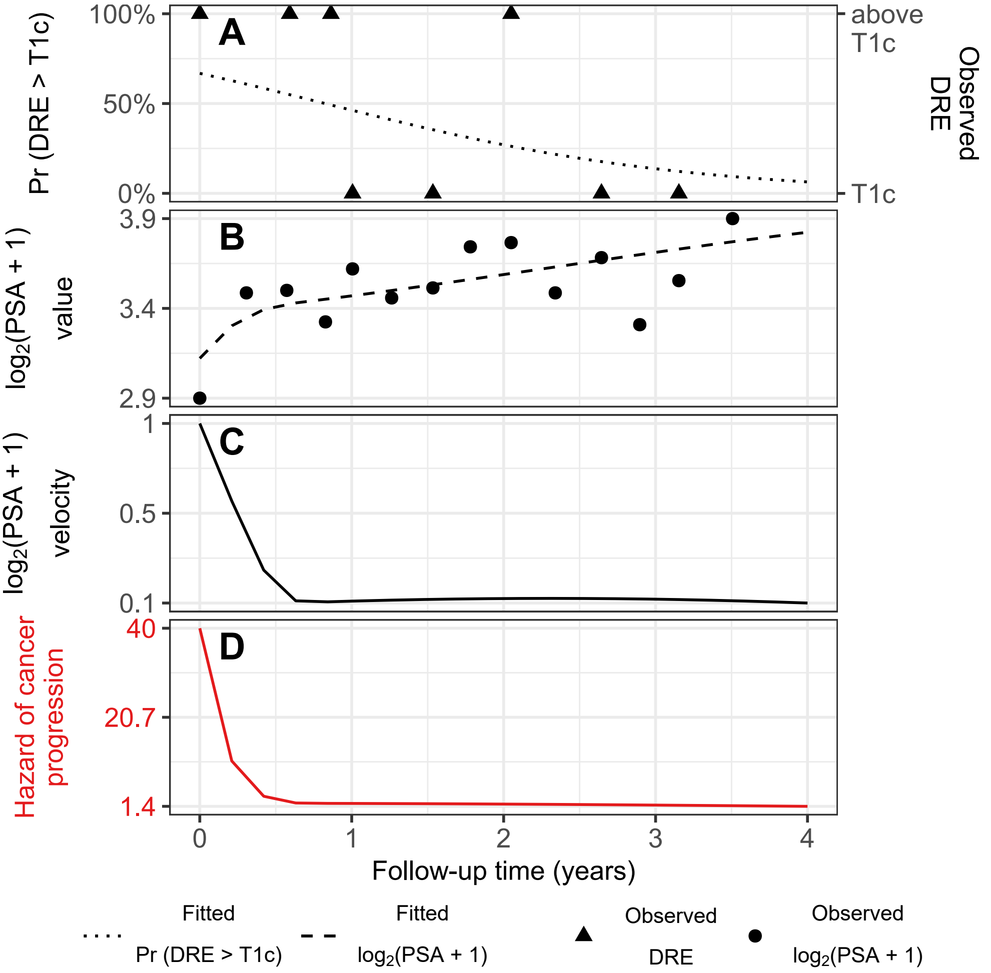

Figure 3.3: **Illustration of the joint model fitted to the PRIAS dataset.** **Panel A:** Observed DRE measurements and the fitted probability of obtaining DRE $>$ T1c (3.1). **Panel B:** Observed and fitted $\log_2(\text{PSA} + 1)$ values (3.2). **Panel C:** Estimated $\log_2(\text{PSA} + 1)$ velocity over time. **Panel D**: Estimated hazard of cancer progression (3.3). It depends on the fitted log odds of having a DRE $>$ T1c, and the fitted $\log_2(\text{PSA} + 1)$ value and velocity.

In our joint model, the patient-specific DRE and PSA measurements over time are modeled using a bivariate generalized linear mixed effects sub-model. The sub-model for DRE is given by (see Panel A, Figure 3.3):

$$\begin{aligned}\mathsf{logit}\Big[\mathsf{Pr}\{y_{di}(t) > \mathsf{T1c}\}\Big] &= \beta_{0d} + b_{0di} + (\beta_{1d} + b_{1di})t \\ &\quad + \beta_{2d}(\mathsf{Age}_i - 70) + \beta_{3d}(\mathsf{Age}_i - 70)^2 \qquad (3.1)\end{aligned}$$

where, t denotes the follow-up visit time, and Age_i is the age of the i-th patient at the time of inclusion in AS. We have centered the Age variable around the median age of 70 years for better convergence during parameter estimation. However, this does not change the interpretation of the parameters corresponding to the Age variable. The fixed effect parameters are denoted by $\{\beta_{0d}, \ldots, \beta_{3d}\}$, and $\{b_{0di}, b_{1di}\}$ are the patient specific random effects. With this definition, we assume that the patient-specific log odds of obtaining a DRE measurement larger than T1c remain linear over time.

The mixed effects sub-model for PSA is given by (see Panel B, Figure 3.3):

$$\begin{aligned}\log_2\Big\{y_{pi}(t) + 1\Big\} &= m_{pi}(t) + \varepsilon_{pi}(t), \\ m_{pi}(t) &= \beta_{0p} + b_{0pi} + \sum_{k=1}^{4}(\beta_{kp} + b_{kpi})B_k(t, \mathcal{K}) \\ &\quad + \beta_{5p}(\mathsf{Age}_i - 70) + \beta_{6p}(\mathsf{Age}_i - 70)^2, \qquad (3.2)\end{aligned}$$

where, $m_{pi}(t)$ denotes the measurement error free value of $\log_2(\mathsf{PSA} + 1)$ transformed (Pearson et al., 1994; Lin et al., 2000) measurements at time t. We model it non-linearly over time using B-splines (De Boor, 1978). To this end, our B-spline basis function $B_k(t, \mathcal{K})$ has 3 internal knots at $\mathcal{K} = \{0.1, 0.7, 4\}$ years, and boundary knots at 0 and 5.42 years (95-th percentile of the observed follow-up times). This specification allows fitting the $\log_2(\mathsf{PSA} + 1)$ levels in a piecewise manner for each patient separately. The internal and boundary knots specify the different time periods (analogously pieces) of this piecewise nonlinear curve. The fixed effect parameters are denoted by $\{\beta_{0p}, \ldots, \beta_{6p}\}$, and $\{b_{0pi}, \ldots, b_{4pi}\}$ are the patient specific random effects. The error $\varepsilon_{pi}(t)$ is assumed to be t-distributed with three degrees of freedom (Figure 3.12) and scale σ, and is independent of the random effects.

To account for the correlation between the DRE and PSA measurements of a patient, we link their corresponding random effects. More specifically, the complete vector of random effects $\boldsymbol{b}_i = (b_{0di}, b_{1di}, b_{0pi}, \ldots, b_{4pi})^T$ is assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{D}$.

To model the impact of DRE and PSA measurements on the risk of cancer progression, our joint model uses a relative risk sub-model. More specifically, the hazard of cancer progression $h_i(t)$ at a time t is given by (see Panel D, Figure 3.3):

$$\begin{aligned} h_i(t) = h_0(t) \exp \biggl(& \gamma_1(\text{Age}_i - 70) + \gamma_2(\text{Age}_i - 70)^2 \\ & + \alpha_{1d}\text{logit}\Bigl[\text{Pr}\{y_{di}(t) > \text{T1c}\}\Bigr] + \alpha_{1p} m_{pi}(t) + \alpha_{2p}\frac{\partial m_{pi}(t)}{\partial t}\biggr), \end{aligned} \tag{3.3}$$

where, γ_1, γ_2 are the parameters for the effect of age. The parameter α_{1d} models the impact of log odds of obtaining a DRE $>$ T1c on the hazard of cancer progression. The impact of PSA on the hazard of cancer progression is modeled in two ways: a) the impact of the error free underlying PSA value $m_{pi}(t)$ (see Panel B, Figure 3.3), and b) the impact of the underlying PSA velocity $\partial m_{pi}(t)/\partial t$ (see Panel C, Figure 3.3). The corresponding parameters are α_{1p} and α_{2p}, respectively. Lastly, $h_0(t)$ is the baseline hazard at time t, and is modeled flexibly using P-splines (Eilers and Marx, 1996). More specifically:

$$\log h_0(t) = \gamma_{h_0,0} + \sum_{q=1}^{Q} \gamma_{h_0,q} B_q(t, \boldsymbol{v}),$$

where $B_q(t, \boldsymbol{v})$ denotes the q-th basis function of a B-spline with knots $\boldsymbol{v} = v_1, \ldots, v_Q$ and vector of spline coefficients γ_{h_0}. To avoid choosing the number and position of knots in the spline, a relatively high number of knots (e.g., 15 to 20) are chosen and the corresponding B-spline regression coefficients γ_{h_0} are penalized using a differences penalty (Eilers and Marx, 1996).The detailed specification of the baseline hazard $h_0(t)$, and the joint parameter estimation of the two sub-models using the Bayesian approach (R package **JMbayes**) are presented in Appendix 3.A.

3.2.3 Personalized Decisions for Biopsy

Let us assume that a decision of conducting a biopsy is to be made for a new patient j shown in Figure 3.1, at his current follow-up visit time s. Let $t \leq s$ be the time of his latest negative biopsy. Let $\mathcal{Y}_{dj}(s)$ and $\mathcal{Y}_{pj}(s)$ denote his observed DRE and PSA measurements up to the current visit, respectively. From the observed measurements we want to extract the underlying measurement error free trend of $\log_2(\text{PSA}+1)$ values and velocity, and the log odds of obtaining DRE $>$ T1c. We intend to combine them to inform us when the cancer progression is to be expected, and to further guide the decision making on whether to conduct a biopsy at the current follow-up visit. The combined information is given by the following posterior predictive distribution $g(T^*_j)$ of his time of cancer progression $T^*_j > t$:

$$g(T^*_j) = p\Big\{T^*_j \mid T^*_j > t, \mathcal{Y}_{dj}(s), \mathcal{Y}_{pj}(s), \mathcal{D}_n\Big\}. \tag{3.4}$$

The distribution $g(T^*_j)$ is not only patient-specific, but also updates as extra information is recorded at future follow-up visits.

A key ingredient in the decision of conducting a biopsy for patient j at the current follow-up visit time s is the personalized cumulative-risk of observing a cancer progression at time s (illustrated in Figure 3.4, and Figure 3.5). This risk can be derived from the posterior predictive distribution $g(T^*_j)$ (Rizopoulos, 2011), and for $s \geq t$ it is given by:

$$R_j(s \mid t) = \text{Pr}\Big\{T^*_j \leq s \mid T^*_j > t, \mathcal{Y}_{dj}(s), \mathcal{Y}_{pj}(s), \mathcal{D}_n\Big\}. \tag{3.5}$$

A simple and straightforward approach to decide upon conducting a biopsy for patient j at the current follow-up visit would be to do so if his personalized cumulative-risk of cancer progression at the visit is higher than a certain

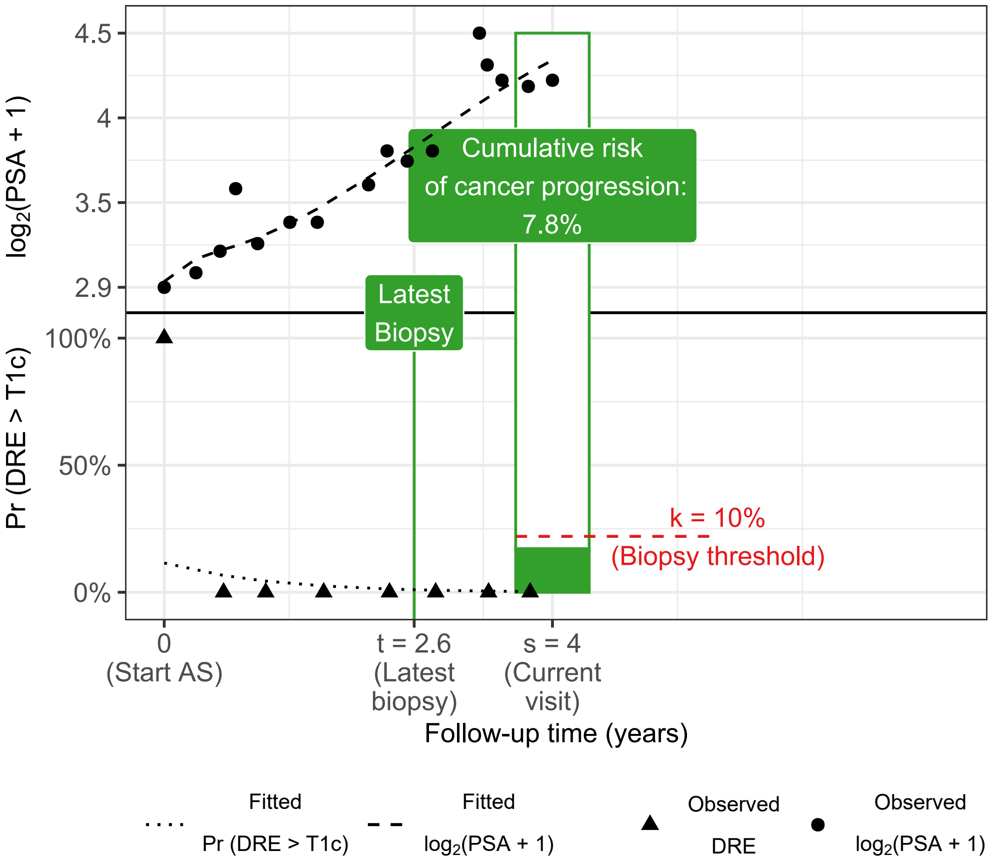

Figure 3.4: **Personalized decision biopsy not recommended**: Biopsy is recommended only if the personalized cumulative-risk of cancer progression estimated from the joint model fitted to the observed data of the j-th patient, is higher than the example risk threshold for biopsy ($\kappa = 10\%$). The cumulative-risk of cancer progression at the current visit time ($s = 4$ years) is 7.8%.

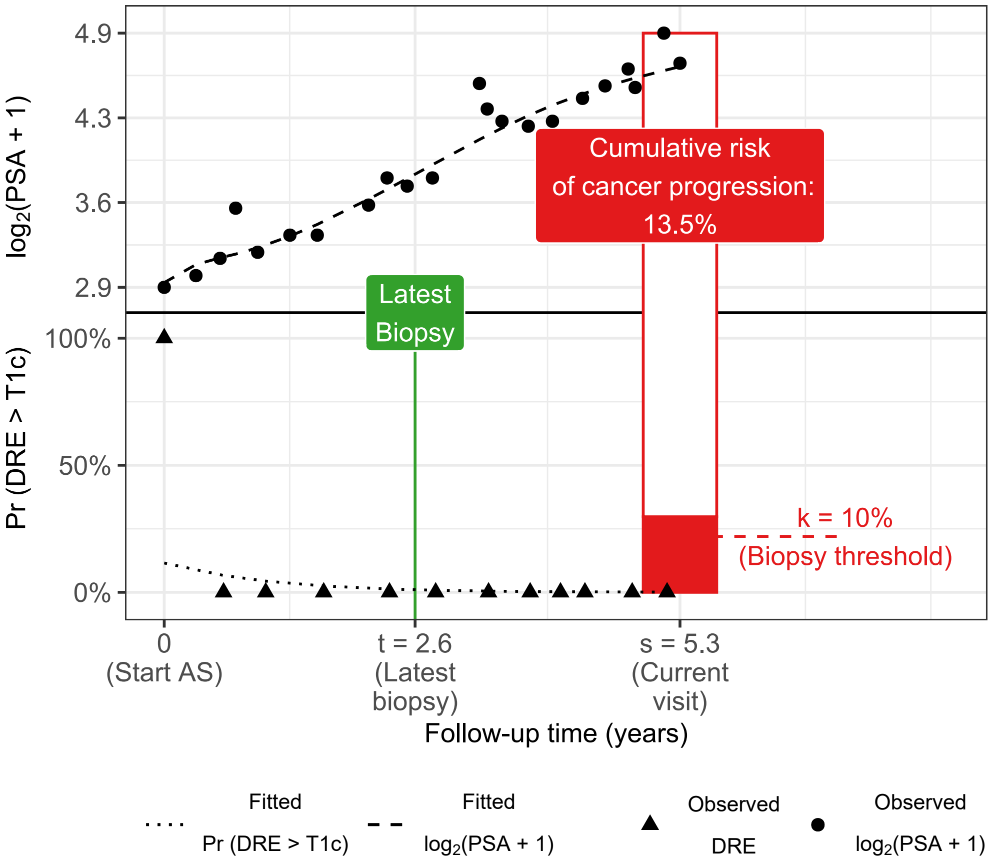

Figure 3.5: **Personalized decision of biopsy recommended**: Biopsy is recommended only if the personalized cumulative-risk of cancer progression estimated from the joint model fitted to the observed data of the j-th patient, is higher than the example risk threshold for biopsy ($\kappa = 10\%$). The cumulative-risk of cancer progression at the current visit time ($s = 5.3$ years) is 13.5%.

threshold $0 \leq \kappa \leq 1$. For example, as shown in Figure 3.4, and Figure 3.5, biopsy at a visit may be scheduled if the personalized cumulative-risk is higher than 10% (example risk threshold). This decision making process is iterated over the follow-up period, incorporating on each subsequent visit the newly observed data, until a positive biopsy is observed. Subsequently, an entire personalized schedule of biopsies for each patient can be obtained.

The choice of the risk threshold dictates the schedule of biopsies and has to be made on each subsequent follow-up visit of a patient. In this regard, a straightforward approach is choosing a fixed risk threshold, such as 5% or 10% risk, at all follow-up visits. Fixed risk thresholds may be chosen by patients and/or doctors according to how they weigh the relative harms of doing an unnecessary biopsy versus a missed cancer progression (e.g., 10% threshold means a 1:9 ratio) if the biopsy is not conducted (Vickers and Elkin, 2006). An alternative approach is that at each follow-up visit, a unique threshold is chosen on the basis of its classification accuracy. More specifically, given the time of latest biopsy t of patient j, and his current visit time s we find a visit-specific biopsy threshold κ, which gives the highest cancer progression detection rate (true positive rate, or TPR) for the period $(t, s]$. However, we also intend to balance for unnecessary biopsies (high false-positive rate), or a low number of correct detections (high false-negative rate) when the false positive rate is minimized. An approach to mitigating these issues is to maximize the TPR and positive predictive value (PPV) simultaneously. To this end, we utilize the F$_1$ score, which is a composite of both TPR and PPV [estimated as in Rizopoulos et al. (2017)] and is defined as:

$$\begin{aligned} \mathsf{F}_1(t, s, \kappa) &= 2\frac{\mathsf{TPR}(t, s, \kappa)\ \mathsf{PPV}(t, s, \kappa)}{\mathsf{TPR}(t, s, \kappa) + \mathsf{PPV}(t, s, \kappa)}, \\ \mathsf{TPR}(t, s, \kappa) &= \mathsf{Pr}\Big\{R_j(s \mid t) > \kappa \mid t < T_j^* \leq s\Big\}, \\ \mathsf{PPV}(t, s, \kappa) &= \mathsf{Pr}\Big\{t < T_j^* \leq s \mid R_j(s \mid t) > \kappa\Big\}, \end{aligned} \tag{3.6}$$

where, TPR(t, s, κ) and PPV(t, s, κ) are the time-dependent true positive rate and positive predictive value, respectively. These values are unique for each combination of the time period $(t, s]$ and the risk threshold κ that is used to discriminate between the patients whose cancer progresses in this time period versus the patients whose cancer does not progress. The same holds true for the resulting F_1 score denoted by $\mathsf{F}_1(t, s, \kappa)$. The F_1 score ranges between 0 and 1, where a value equal to 1 indicates perfect TPR and PPV. Thus the highest F_1 score is desired in each time period $(t, s]$. This can be achieved by choosing a risk threshold κ, which maximizes $\mathsf{F}_1(t, s, \kappa)$. That is, during a patient's visit at time s, given that his latest biopsy was at the time t, the visit-specific risk threshold to decide a biopsy is given by $\kappa = \arg\max_{\kappa} \mathsf{F}_1(t, s, \kappa)$. The criteria on which we evaluate the personalized schedules based on fixed and visit-specific risk thresholds is the total number of biopsies scheduled, and the delay in detection of cancer progression (details in Results).

3.2.4 Simulation Study

Although the personalized decision-making approach is motivated by the PRIAS study, it is not possible to evaluate it directly on the PRIAS dataset. This is because the patients in PRIAS have already had their biopsies as per the PRIAS protocol. In addition, the true time of cancer progression is interval or right-censored for all patients, making it impossible to correctly estimate the delay in the detection of cancer progression due to a particular schedule. To this end, we conduct an extensive simulation study to find the utility of personalized, PRIAS, and fixed/heuristic schedules. For a realistic comparison, we simulate patient data from the joint model fitted to the PRIAS dataset. The simulated population has the same ten year follow-up period as the PRIAS study. In addition, the estimated relations between DRE and PSA measurements, and the risk of cancer progression are retained in the simulated population.

From this population, we first sample 500 datasets, each representing a hypothetical AS program with 1000 patients in it. We generate a true cancer progression time for each of the 500 × 1000 patients and then sample a set of DRE and PSA measurements at the same follow-up visit times as given in PRIAS protocol. We then split each dataset into training (750 patients) and test (250 patients) parts, and generate a random and non-informative censoring time for the training patients. We next fit a joint model of the specification given in (3.1), (3.2), and (3.3) to each of the 500 training datasets and obtain MCMC samples from the 500 sets of the posterior distribution of the parameters.

In each of the 500 hypothetical AS programs, we utilize the corresponding fitted joint models to develop cancer progression risk profiles for each of the 500 × 250 test patients. We make the decision of biopsies for patients at their pre-scheduled follow-up visits for DRE and PSA measurements (see Section 3.2.1), on the basis of their estimated personalized cumulative-risk of cancer progression. These decisions are made iteratively until a positive biopsy is observed. A recommended gap of one year between consecutive biopsies (Bokhorst et al., 2015) is also maintained. Subsequently, for each patient, an entire personalized schedule of biopsies is obtained.

We evaluate and compare both personalized and currently practiced schedules of biopsies in this simulation study. A comparison of the schedules is based on the number of biopsies scheduled and the corresponding delay in the detection of cancer progression. We evaluate the following currently practiced fixed/heuristic schedules: biopsy annually, biopsy every one and a half years, biopsy every two years, and biopsy every three years. We also evaluate the biopsy schedule of the PRIAS program (see Section 3.1). For the personalized biopsy schedules, we evaluate schedules based on three fixed risk thresholds: 5%, 10%, and 15%, corresponding to a missed cancer progression being 19, 9, and 5.5 times more harmful than an unnecessary biopsy (Vickers and Elkin, 2006), respectively. We also implement a personalized schedule wherein for each patient, visit-specific risk thresholds are chosen using F_1 score.

3.3 Results

From the joint model fitted to the PRIAS dataset, we found that both $\log_2\{\text{PSA} + 1\}$ velocity, and log odds of having DRE $>$ T1c were significantly associated with the hazard of cancer progression. For any patient, an increase in $\log_2\{\text{PSA} + 1\}$ velocity from -0.03 to 0.16 (first and third quartiles of the fitted velocities, respectively) corresponds to a 1.94 fold increase in the hazard of cancer progression. Whereas, an increase in odds of DRE $>$ T1c from -6.650 to -4.356 (first and third quartiles of the fitted log-odds, respectively) corresponds to a 1.40 fold increase in the hazard of cancer progression. Detailed results pertaining to the fitted joint model are presented in Appendix 3.A.1.

3.3.1 Comparison of Various Approaches for Biopsies

From the simulation study, we obtain the number of biopsies and the delay in detection of cancer progression for each of the 500 × 250 test patients using different schedules. Figure 3.6 shows that the personalized and PRIAS approaches fall in the region of a better balance between the median number of biopsies and the median delay than fixed/heuristic schedules. Next evaluate these schedules on the basis of both median and interquartile range (IQR) of the number of biopsies and delay (see Figure 3.7). For brevity, only the most widely used annual and PRIAS schedules, the proposed personalized approach with fixed risk thresholds of 5% and 10%, and visit-specific threshold chosen using F_1 score are discussed next (see Table 3.4 for remaining).

Since patients have varying cancer progression speeds, the impact of each schedule also varies with it. To highlight these differences, we divide results for three types of patients, as per their time of cancer progression. They are *fast, intermediate,* and *slow progressing* patients. Although such a division may be imperfect and can only be done retrospectively in a simulation setting, we show results for these three groups for illustration. Roughly 50% of the patients did not obtain cancer progression in the ten year follow-up period of the simulation study. We assume these patients to be *slow progressing*

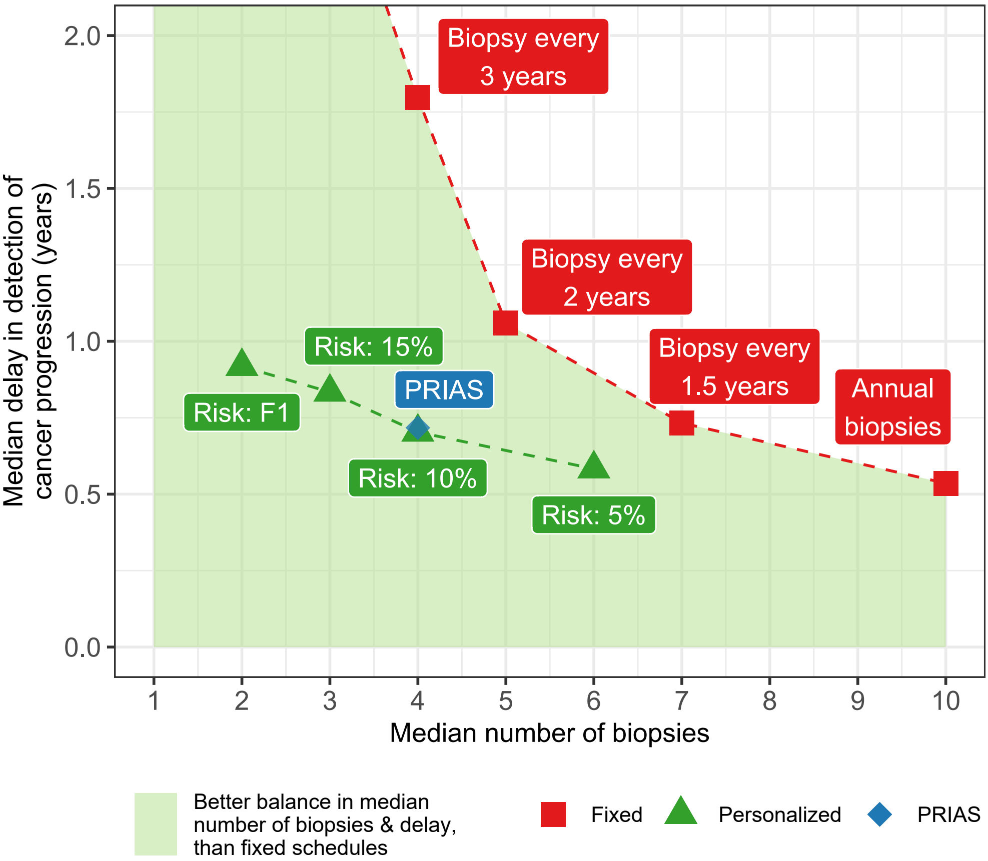

Figure 3.6: **Burden-biopsy frontier:** Median number of biopsies (X-axis), and median delay in detection of cancer progression, in years (Y-axis), estimated from the simulation study. **Personalized schedules:** Risk: 15%, Risk: 10%, and Risk: 5% approaches, schedule a biopsy if the cumulative-risk of cancer progression at a visit is more than 15%, 10%, and 5%, respectively. Risk: F1 works similarly, except that it utilizes a visit-specific threshold (see Section 3.2.3). The green shaded region depicts the region of better balance in the median number of biopsies and median delay than the currently practiced fixed/heuristic schedules.

patients. We assume *fast progressing* patients are the ones with an initially misdiagnosed state of cancer (Cooperberg et al., 2011) or high-risk patients who choose AS instead of immediate treatment upon diagnosis. These are roughly 30% of the population, having a cancer progression time less than 3.5 years. We label the remaining 20% patients as *intermediate progressing* patients.

For *fast progressing* patients (Panel A, Figure 3.7), we note that the personalized schedules with a fixed 10% risk threshold and visit-specific threshold chosen using F_1 score, reduce one biopsy for 50% of the patients, compared to PRIAS and annual schedule. Despite this, the delay (years) is similar for the personalized schedule with fixed 10% risk threshold (median: 0.7, IQR: 0.3–1.0), and the commonly used annual (median: 0.6, IQR: 0.3–0.9) and PRIAS (median: 0.7, IQR: 0.3–1.0) schedules.

For *intermediate progressing* patients (Panel A, Figure 3.7), we note that the delay (years) due to personalized schedule with fixed 5% risk threshold (median: 0.6, IQR: 0.3–0.9) is comparable to that of annual schedule (median 0.5, IQR: 0.2–0.7). However, it schedules fewer biopsies (median: 6, IQR: 5–7) than the annual schedule (median: 7, IQR: 5–8). The delay (years) for PRIAS (median: 0.7, IQR: 0.3–1.3) and personalized schedule with fixed 10% risk (median: 0.7, IQR: 0.4–1.3) are similar, but the personalized approach schedules one less biopsy for 50% of the patients. Although the approach with the visit-specific risk threshold chosen using F_1 score schedules fewer biopsies than the 10% fixed risk approach, it also has a higher delay.

The patients who are at the most advantage with the personalized schedules are the *slow progressing* patients. These are a total of 50% patients who did not progress during the entire study. Hence, the delay is not available for these patients (Panel C of Figure 3.7). For all of these patients, the annual schedule leads to 10 (unnecessary) biopsies. The schedule of the PRIAS program schedules a median of six biopsies (IQR: 4–8). In comparison, the biopsies scheduled by the personalized schedules using fixed 10% risk threshold (median: 4, IQR: 4–6) and visit-specific risk chosen using F_1 score (median: 2, IQR: 2–4), are much fewer.

Overall, we observed that the personalized schedule which uses a 10% risk

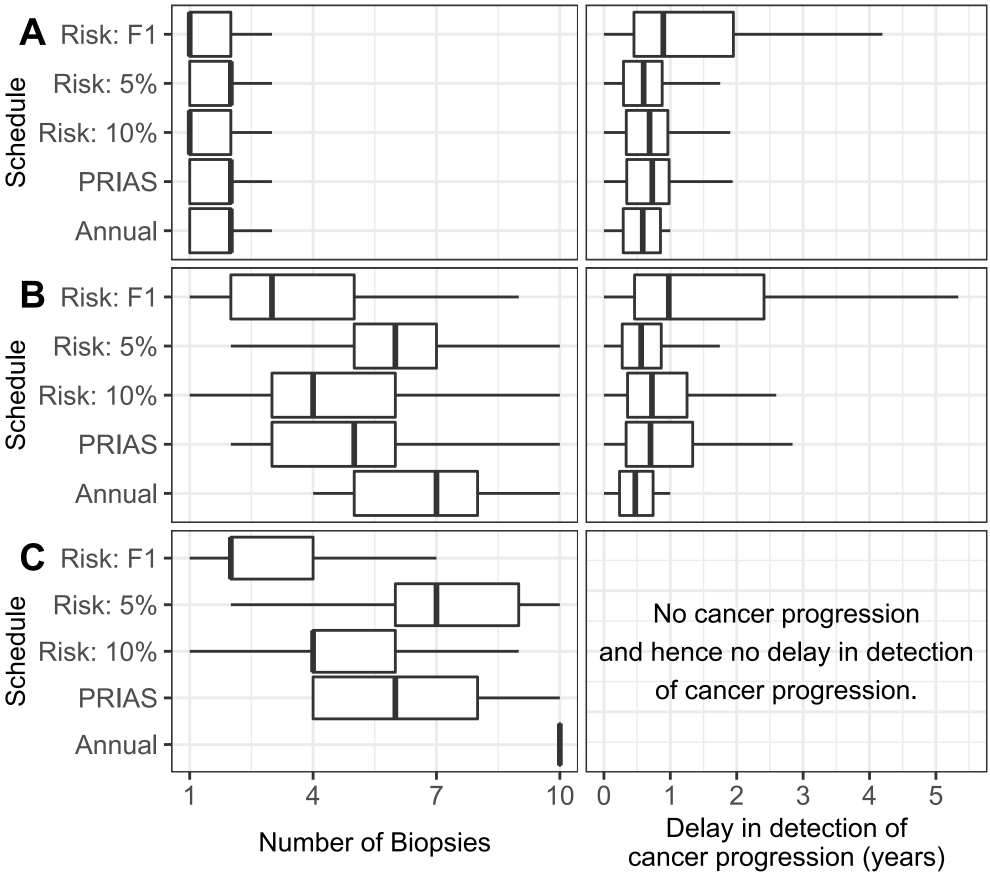

Figure 3.7: Variation in the number of biopsies, and the delay in detection of cancer progression, in years, for various biopsy schedules. **Panel A:** simulated patients with cancer progression times between 0 and 3.5 years (*fast progressing*). **Panel B:** simulated patients with progression times between 3.5 and 10 years (*intermediate progressing*). **Panel C:** simulated patients who did not have cancer progression in the ten years of follow-up (*slow progressing*). **Personalized schedules:** Risk: 10% approach schedules a biopsy at a visit if the corresponding cumulative-risk of cancer progression is more than 10%. Risk: 5% and Risk: F1 work similarly, except that a visit-specific threshold is used in the latter (see Section 3.2.3). **Annual:** Yearly biopsies, and **PRIAS:** biopsies as per PRIAS protocol (see Section 3.1).

threshold at all follow-up visits is dominant over the PRIAS schedule, biennial schedule of biopsies, and biopsies every one and a half years (see Table 3.4 for the latter two schedules). This personalized schedule not only schedules fewer biopsies than the aforementioned currently practiced schedules, but the delay in the detection of cancer progression is also either equal or less. The personalized schedule, which uses the risk threshold chosen based on classification accuracy (F_1 score) is dominant over the triennial schedule of biopsies. The personalized schedule which uses a 5% risk threshold schedules fewer biopsies than the annual schedule, while the delay is only trivially more than the annual schedule.

3.4 Discussion

We proposed a methodology which better balances the number of biopsies, and the delay in detection of cancer progression than the currently practiced biopsy schedules, for low-risk prostate cancer patients enrolled in active surveillance (AS) programs. The proposed methodology combines a patient's observed DRE and PSA measurements, and the time of the latest biopsy, into a personalized cancer progression risk function. If the cumulative-risk of cancer progression at a follow-up visit is above a certain threshold, then a biopsy is scheduled. We conducted an extensive simulation study, based on a replica of the patients from the PRIAS program, to compare this personalized approach for biopsies with the currently practiced biopsy schedules. We found personalized schedules to be dominant over many of the current biopsy schedules (see Section 3.3).

The main reason for the better performance of personalized schedules is that they account for the variation in cancer progression rate between patients, and also over time within the same patient. In contrast, the existing fixed/heuristic schedules ignore that roughly 50% of the patients never progress in the first ten years of follow-up (*slow progressing* patients) and do not require biopsies. The *fast progressing* patients require early detection. However, existing methods of identifying these patients, such as the use of

PSA doubling time in PRIAS, inappropriately assume that PSA evolves linearly over time. Thus, they may not correctly identify such patients. The personalized approach, however, models the PSA profiles non-linearly. Furthermore, it appends information from PSA with information from DRE and previous biopsy results and combines them into a single cancer progression risk function. The risk function is a finer quantitative measure than individual data measurements observed for the patients. In comparison to decision making with flowcharts, the risk as a single measure of a patient's underlying state of cancer may facilitate shared decision making for biopsies.

Existing work on reducing the burden of biopsies in AS primarily advocates less frequent heuristic schedules (e.g., biopsies biennially instead of annually) of biopsies (Inoue et al., 2018). To our knowledge, risk-based biopsy schedules have barely been explored yet in AS (Nieboer et al., 2018; Bruinsma et al., 2016). The part of our results pertaining to the fixed/heuristic schedules is comparable with corresponding results obtained in existing work (Inoue et al., 2018), even though the AS cohorts are not the same. Thus, we anticipate similar validity for the results pertaining to the personalized schedules.

A limitation of the personalized approach is that the choice of risk threshold is not straightforward, as different thresholds lead to different combinations of the number of biopsies and the delay in the detection of cancer progression. An approach is to choose a risk threshold that leads to personalized schedule dominant (e.g., 10% risk) over the currently practiced schedules, for a given delay. Since personalized biopsy schedules are less burdensome, they may lead to better compliance. A second limitation is that the results that we presented are valid only in a ten year follow-up period, whereas prostate cancer is a slowly progressing disease. Thus more detailed results, especially for *slow progressing* patients, cannot be estimated. However, very few AS cohorts have a longer follow-period than PRIAS (Bruinsma et al., 2016). In a screening setting, often the ethno-racial background of the patient, as well as the history of cancer in first degree relatives, are checked. Our model does not take into account either. The reason is that the history of cancer in relatives been found to be predictive of cancer progression only in African-American patients (Goh et al., 2013; Telang et al., 2017). This is also evident by the fact that PRIAS and many other surveillance programs do not utilize this information in their biopsy protocols (Bokhorst et al., 2016; Nieboer et al., 2018). In addition, patients who have a higher risk of an aggressive form of cancer are usually not recommended active surveillance. Hence the proposed model is relevant only for low-risk prostate cancer patients eligible for active surveillance. An exception is the active surveillance patients who are old and/or have comorbid illnesses. Currently, such patients may be removed from active surveillance and are instead offered the less intensive watchful waiting (Bokhorst et al., 2016) option. It is also possible to model watchful waiting as a competing risk in our model. However, this falls outside the scope of the current work because cancer progression, as detected via biopsy, is the standard trigger for treatment advice. Lastly, our results are not valid when the patient data is missing not at random (MNAR).

There are multiple ways to extend the personalized decision-making approach. For example, biopsy Gleason grading is susceptible to inter-observer variation (Coley et al., 2017). Thus accounting for it in our model will be interesting to investigate further. To improve the decision making methodology, future consequences of a biopsy can be accounted for in the model by combining Markov decision processes with joint models for time-to-event and longitudinal data. There is also a potential for including diagnostic information from magnetic resonance imaging (MRI), such as the volume of the prostate tumor as a longitudinal measurement in our model. The resulting predictions can be used to decide the time of the next MRI as well as to make a decision of biopsy. The same holds true for the quality of life measures as well. However, given the scarceness of both MRI and quality of life measurements in the dataset, including them in the current model may not be feasible. We intend further to validate our results in a multi-center AS cohort and subsequently develop a web application to assist in making shared decisions for biopsies.

Acknowledgements The first and last Authors would like to acknowledge support by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (the national research council of the Netherlands) VIDI grant nr. 016.146.301, and Erasmus University Medical Center funding. The Authors also thank the Erasmus University Medical Center's Cancer Computational Biology Center for giving access to their IT-infrastructure and software that was used for the computations and data analysis in this study. Lastly, we thank Joost Van Rosmalen from the Department of Biostatistics, Erasmus University Medical Center for feedback on the manuscript.

Conflict of Interest The Authors declare that there is no conflict of interest.

Appendix

3.A Parameter Estimation

We estimate the parameters of the joint model using Markov chain Monte Carlo (MCMC) methods under the Bayesian framework. Let $\boldsymbol{\theta}$ denote the vector of all of the parameters of the joint model. The joint model postulates that given the random effects, the time to cancer progression, and the DRE and PSA measurements taken over time are all mutually independent. Under this assumption the posterior distribution of the parameters is given by:

$$\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D}_n) &\propto \prod_{i=1}^{n} p(l_i, r_i, \boldsymbol{y}_{di}, \boldsymbol{y}_{pi}, \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})\\
&\propto \prod_{i=1}^{n} p(l_i, r_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{y}_{di} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{y}_{pi} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}),\\
p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) &= \frac{1}{\sqrt{(2\pi)^q \text{det}(\boldsymbol{D})}} \exp(\boldsymbol{b}_i^T \boldsymbol{D}^{-1} \boldsymbol{b}_i),
\end{aligned}$$

where, the likelihood contribution of the DRE outcome, conditional on the random effects is:

$$p(\boldsymbol{y}_{di} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \prod_{k=1}^{n_{di}} \frac{\exp\Big[-\textsf{logit}\Big\{\textsf{Pr}(y_{dik} > \textsf{T1c})\Big\} I(y_{dik} = \textsf{T1c})\Big]}{1 + \exp\Big[-\textsf{logit}\Big\{\textsf{Pr}(y_{dik} > \textsf{T1c})\Big\}\Big]},$$

where $I(\cdot)$ is an indicator function which takes the value 1 if the k-th repeated DRE measurement $y_{dik} =$ T1c, and takes the value 0 otherwise. The likelihood contribution of the PSA outcome, conditional on the random effects is:

$$p(\boldsymbol{y}_{pi} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^{n_{pi}}} \exp\left(-\frac{\|\boldsymbol{y}_{pi} - \boldsymbol{m}_{pi}\|^2}{\sigma^2}\right),$$

The likelihood contribution of the time to cancer progression outcome is given by:

$$p(l_i, r_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \exp\Big\{-\int_0^{l_i} h_i(s)\mathrm{d}s\Big\} - \exp\Big\{-\int_0^{r_i} h_i(s)\mathrm{d}s\Big\}. \quad (3.7)$$

The integral in (3.7) does not have a closed-form solution, and therefore we use a 15-point Gauss-Kronrod quadrature rule to approximate it.

We use independent normal priors with zero mean and variance 100 for the fixed effects $\{\beta_{0d}, \ldots, \beta_{3d}, \beta_{0p}, \ldots, \beta_{6p}\}$, and inverse Gamma prior with shape and rate both equal to 0.01 for the parameter σ^2. For the variance-covariance matrix $\boldsymbol{D}$ of the random effects we take inverse Wishart prior with an identity scale matrix and degrees of freedom equal to 7 (number of random effects). For the relative risk model's parameters $\{\gamma_1, \gamma_2\}$ and the association parameters $\{\alpha_{1d}, \alpha_{1p}, \alpha_{2p}\}$, we use independent normal priors with zero mean and variance 100.

3.A.1 Parameter Estimates

The longitudinal evolution of $\log_2$(PSA$+1$) is modeled with non-linear terms, and hence the interpretation of the coefficients in this model is not straightforward. In the case of the evolution of DRE, the coefficients in the model correspond to a patient with a random effect value equal to 0. That is, the coefficients do not describe the marginal evolution of DRE over time. To avoid these issues, instead of the parameter estimates, in Figure 3.8 and Figure 3.9 we present the fitted marginal evolution of probability of DRE $>$ T1c and $\log_2(\text{PSA}+1)$, respectively, over a period of 10 years for a hypothetical patient who is included in AS at the age of 70 years. In addition, we present plots of observed versus fitted DRE and PSA profiles for nine randomly selected PRIAS patients in Figure 3.10 and Figure 3.11, respectively. Lastly, the quantile-quantile plot of subject-specific residuals in Figure 3.12 shows that the assumption of t-distributed (df=3) errors is reasonably met by the fitted model.

For the relative risk sub-model, the parameter estimates in Table 3.2 show that both $\log_2(\text{PSA}+1)$ velocity, and the log odds of having DRE $>$ T1c were significantly associated with the hazard of cancer progression. It is important to note that since age, $\log_2(\text{PSA}+1)$ value and velocity, and log odds of DRE $>$ T1c are all measured on different scales, a comparison between the corresponding parameter estimates is not easy. To this end, in Table 3.3, we present the hazard (of cancer progression) ratio, for an increase in the aforementioned variables from their first to the third quartile. For example, an increase in log odds of DRE $>$ T1c, from -6.650 to -4.356 (fitted first and third quartiles) corresponds to a hazard ratio of 1.402. The interpretation of the rest is similar.

3.A.2 Simulation Study Results

In the simulation study, we evaluate the following in-practice fixed/heuristic approaches (Loeb et al., 2014; Inoue et al., 2018) for biopsies: biopsy every year, biopsy every one and a half years, biopsy every two years and biopsy every three years. For the personalized biopsy approach, we evaluate three fixed risk thresholds: 5%, 10%, and 15%, and a risk threshold was chosen using F_1 score. Lastly, we also evaluate the PRIAS schedule of biopsies. We compare all the aforementioned schedules on two criteria, namely the

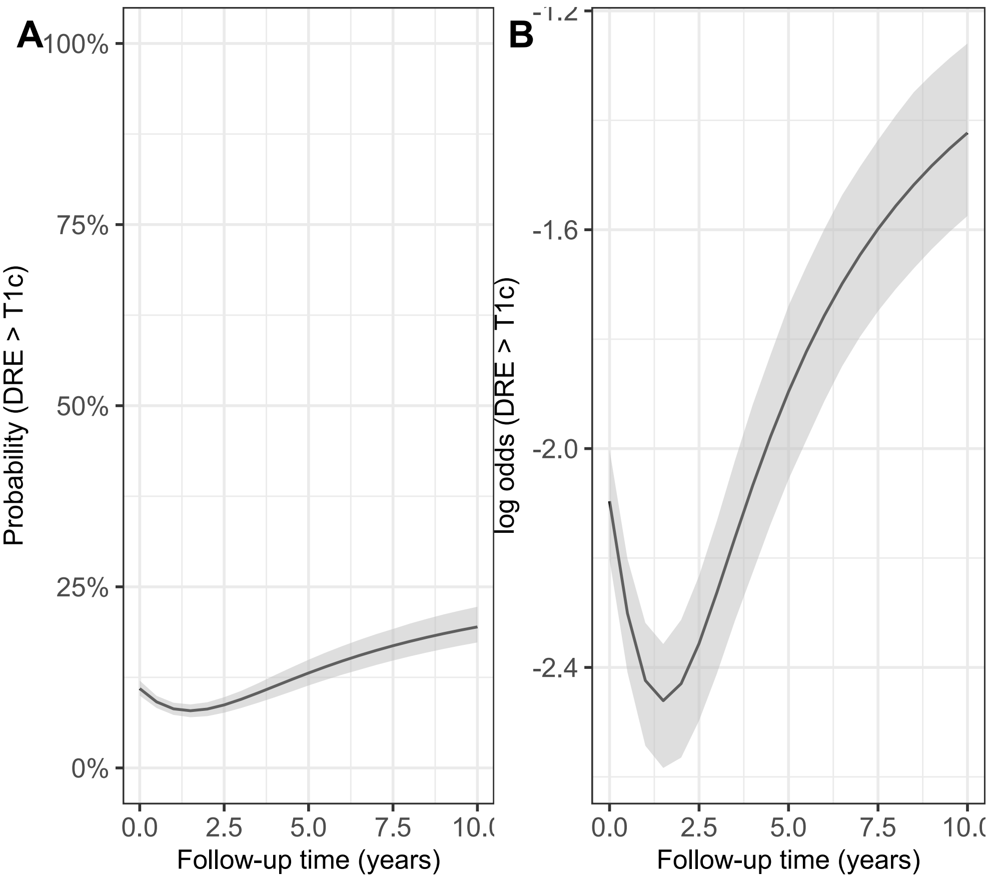

Figure 3.8: **Fitted marginal evolution** of the probability of obtaining a DRE larger than T1c, and the corresponding marginal log odds, with 95% credible interval. These results are for a hypothetical AS patient who is included in AS at the age of 70 years.

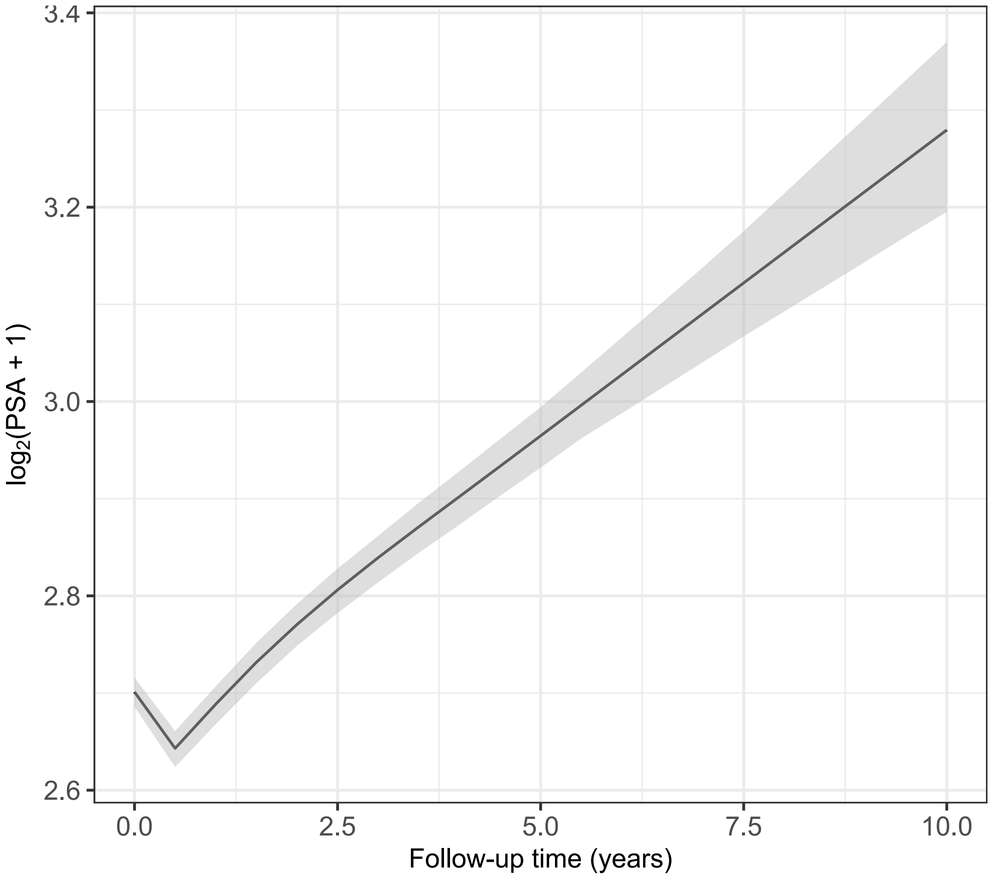

Figure 3.9: **Fitted marginal evolution** of $\log_2(\text{PSA}+1)$ measurements over a period of 10 years with 95% credible interval, for a hypothetical patient who is included in AS at the age of 70 years.

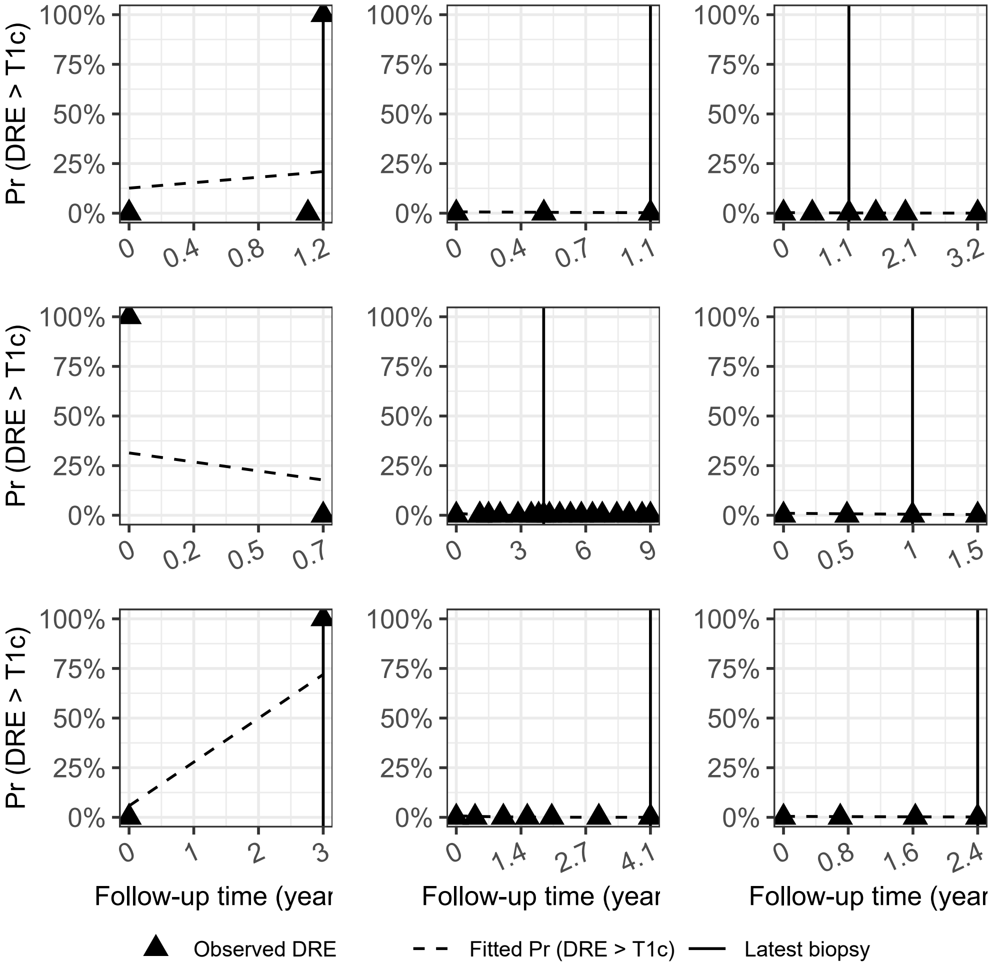

Figure 3.10: **Observed DRE versus fitted probabilities** of obtaining a DRE measurement larger than T1c, for nine randomly selected PRIAS patients. The fitted profiles utilize information from the observed DRE measurements, PSA measurements, and time of the latest biopsy. Observed DRE measurements plotted against 0% probability are equal to T1c. Observed DRE measurements plotted against 100% probability are larger than T1c.

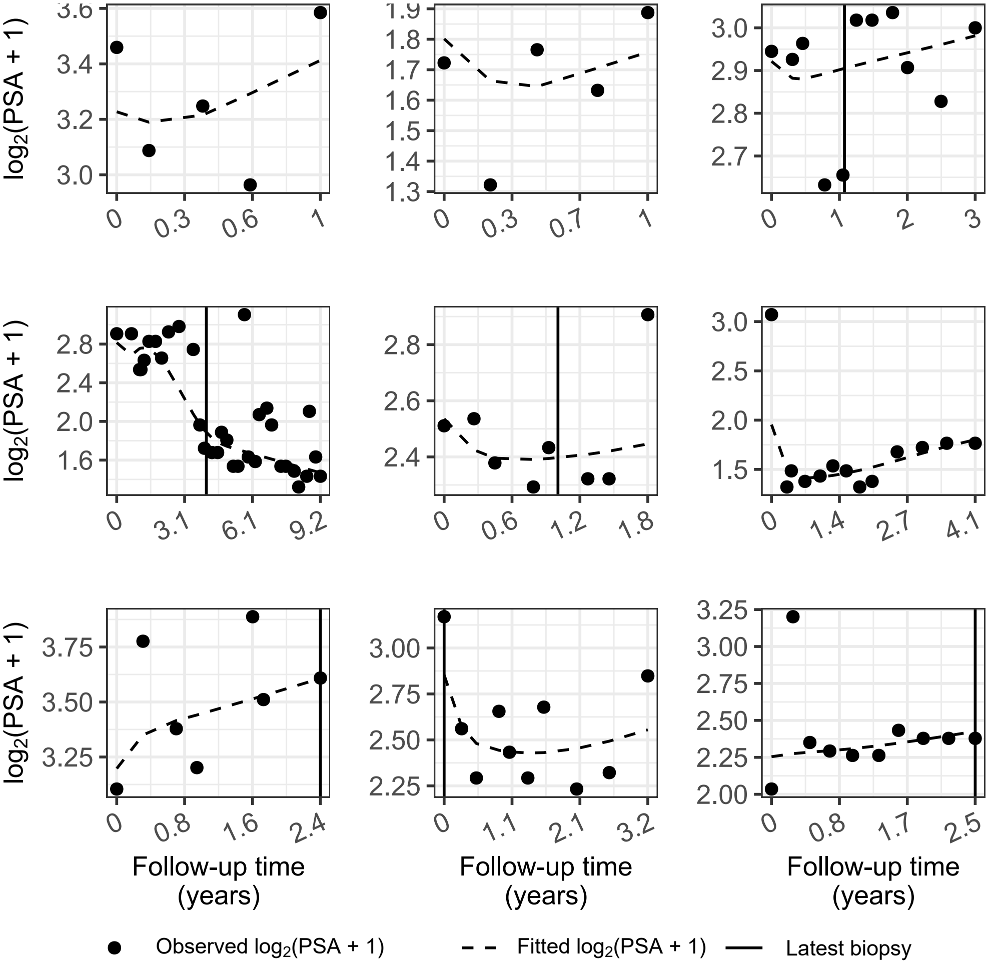

Figure 3.11: **Fitted versus observed** $\log_2(\text{PSA} + 1)$ profiles for nine randomly selected PRIAS patients. The fitted profiles utilize information from the observed PSA measurements, DRE measurements, and time of the latest biopsy.

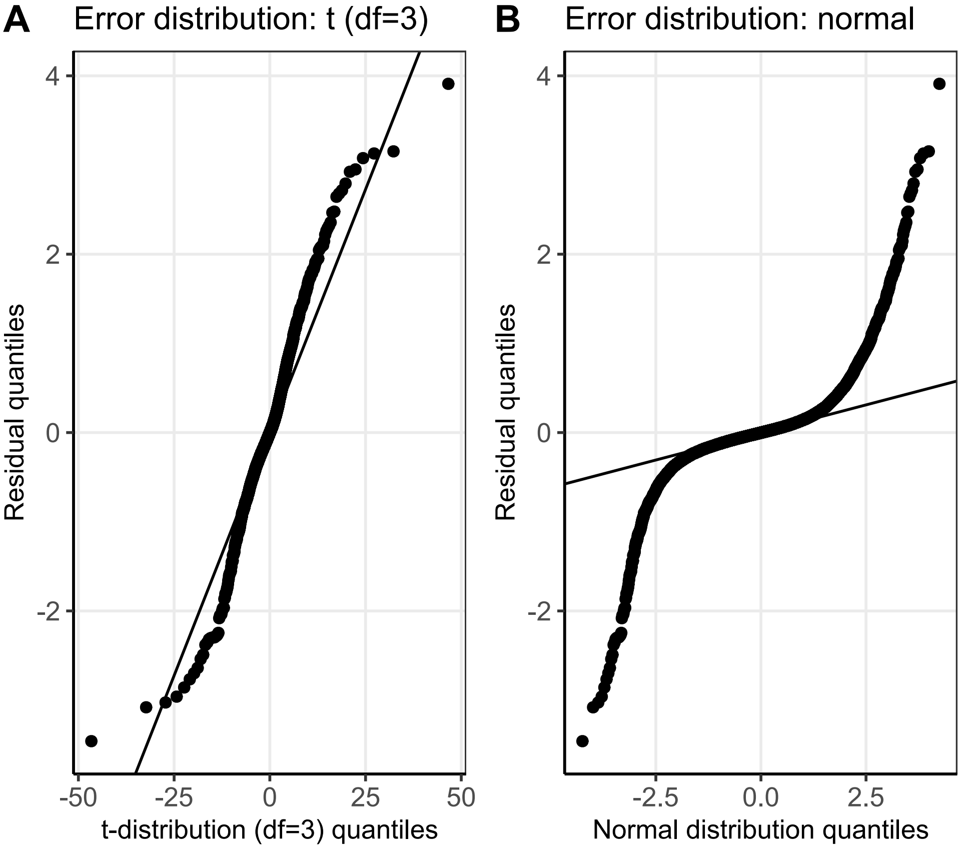

Figure 3.12: **Quantile-quantile plot** of the subject-specific residuals from different joint models fitted to the PRIAS dataset. **Panel A**: model assuming a t-distribution (df=3) for the error term ε_p **Panel B**: model assuming a normal distribution for the error term ε_p.

Table 3.2: **Parameters of the relative-risk sub-model**: Estimated mean and 95% credible interval. Age is median centered.

Variable	Mean	Std. Dev	2.5%	97.5%	P
(Age $-$ 70)	0.012	0.006	0.000	0.022	0.045
(Age $-$ 70)^2	-0.001	0.001	-0.002	0.000	0.095
logit{Pr(DRE $>$ T1c)}	0.147	0.017	0.115	0.183	<0.001
Fitted $\log_2$(PSA $+ 1$) value	0.104	0.078	-0.044	0.256	0.193
Fitted $\log_2$(PSA $+ 1$) velocity	3.396	0.564	2.376	4.475	<0.001

Table 3.3: **Hazard (of cancer progression) ratio and 95% credible interval (CI)**, for an increase in the variables of relative risk sub-model, from their first quartile (Q$_1$) to their third quartile (Q$_3$). Except for age, quartiles for all other variables are based on their fitted values obtained from the joint model fitted to the PRIAS dataset.

Variable	Q$_1$	Q$_3$	Hazard ratio [95% CI]
Age	65	75	1.129 [1.002, 1.251]
logit{Pr(DRE $>$ T1c)}	-6.650	-4.356	1.402 [1.301, 1.521]
$\log_2$(PSA $+ 1$) value	2.336	3.053	1.079 [0.969, 1.201]
$\log_2$(PSA $+ 1$) velocity	-0.032	0.161	1.938 [1.582, 2.372]

number of biopsies they schedule and the corresponding delay in detection of cancer progression in years (time of positive biopsy - the true time of cancer progression). The corresponding results, using 500 × 250 test patients are presented in Table 3.4.

3.B Source Code

The source code for fitting the joint model is available at `https://github.com/anirudhtomer/prias/blob/master/src/chapter3_mdmpaper/fittingModel jmFit.R`.

Table 3.4: **Simulation study results for all patients**: Estimated first, second (median), and third quartiles for number of biopsies (Q_1^{nb}, Q_2^{nb}, Q_3^{nb}) and for the delay in detection of cancer progression (Q_1^{delay}, Q_2^{delay}, Q_3^{delay}), in years, for various biopsy schedules. The delay is equal to the difference between the time of the positive biopsy and the unobserved true time of progression. The results in the table are obtained from test patients of our simulation study.

In-practice schedules	Q_1^{nb}	Q_2^{nb}	Q_3^{nb}	Q_1^{delay}	Q_2^{delay}	Q_3^{delay}
Every year (annual)	3	10	10	0.3	0.5	0.8
Every 1.5 years	2	7	7	0.4	0.7	1.1
Every 2 years	2	5	5	0.6	1.1	1.5
Every 3 years	1	4	4	1.1	1.8	2.3
PRIAS	2	4	6	0.3	0.7	1.0
Personalized approach						
Risk threshold: 5%	2	6	8	0.3	0.6	0.9
Risk threshold: 10%	2	4	5	0.3	0.7	1.0
Risk threshold: 15%	2	3	4	0.4	0.8	1.4
Risk using F_1 score	1	2	3	0.5	0.9	2.2

The code generating the simulation population is available at `https://github.com/anirudhtomer/prias/blob/master/src/chapter3_mdmpaper/simulationStudy/controller.R`.

The code for scheduling biopsies using fixed and risk based schedules is available at `https://github.com/anirudhtomer/prias/blob/master/src/chapter3_mdmpaper/simulationStudy/schedules.R`.

3.3 References

Bokhorst, L. P., Alberts, A. R., Rannikko, A., Valdagni, R., Pickles, T., Kakehi, Y., Bangma, C. H., Roobol, M. J., and PRIAS study group (2015). Compliance rates with the Prostate Cancer Research International Active Surveillance (PRIAS) protocol and disease reclassification in noncompliers. *European Urology*, 68(5):814–821.

Bokhorst, L. P., Valdagni, R., Rannikko, A., Kakehi, Y., Pickles, T., Bangma, C. H., Roobol, M. J., and PRIAS study group (2016). A decade of active surveillance in the PRIAS study: an update and evaluation of the criteria used to recommend a switch to active treatment. *European Urology*, 70(6):954–960.

Bruinsma, S. M., Bangma, C. H., Carroll, P. R., Leapman, M. S., Rannikko, A., Petrides, N., Weerakoon, M., Bokhorst, L. P., Roobol, M. J., Ehdaie, B., et al. (2016). Active surveillance for prostate cancer: a narrative review of clinical guidelines. *Nature Reviews Urology*, 13(3):151–167.

Coley, R. Y., Zeger, S. L., Mamawala, M., Pienta, K. J., and Carter, H. B. (2017). Prediction of the pathologic Gleason score to inform a personalized management program for prostate cancer. *European Urology*, 72(1):135–141.

Cooperberg, M. R., Cowan, J. E., Hilton, J. F., Reese, A. C., Zaid, H. B., Porten, S. P., Shinohara, K., Meng, M. V., Greene, K. L., and Carroll, P. R. (2011). Outcomes of active surveillance for men with intermediate-risk prostate cancer. *Journal of Clinical Oncology*, 29(2):228–234.

De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

de Carvalho, T. M., Heijnsdijk, E. A., and de Koning, H. J. (2017). Estimating the risks and benefits of active surveillance protocols for prostate cancer: a microsimulation study. *BJU International*, 119(4):560–566.

Ehdaie, B., Vertosick, E., Spaliviero, M., Giallo-Uvino, A., Taur, Y., O'sullivan, M., Livingston, J., Sogani, P., Eastham, J., Scardino, P., et al. (2014). The impact of repeat biopsies on infectious complications in men with prostate cancer on active surveillance. *The Journal of Urology*, 191(3):660–664.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Etzioni, R., Penson, D. F., Legler, J. M., Di Tommaso, D., Boer, R., Gann, P. H., and Feuer, E. J. (2002). Overdiagnosis due to prostate-specific antigen screening: lessons from US prostate cancer incidence trends. *Journal of the National Cancer Institute*, 94(13):981–990.

Fujita, K., Landis, P., McNeil, B. K., and Pavlovich, C. P. (2009). Serial prostate biopsies are associated with an increased risk of erectile dysfunction in men with prostate cancer on active surveillance. *The Journal of Urology*, 182(6):2664–2669.

Goh, C. L., Saunders, E. J., Leongamornlert, D. A., Tymrakiewicz, M., Thomas, K., Selvadurai, E. D., Woode-Amissah, R., Dadaev, T., Mahmud, N., Castro, E., et al. (2013). Clinical implications of family history of prostate cancer and genetic risk single nucleotide polymorphism (snp) profiles in an active surveillance cohort. *BJU International*, 112(5):666–673.

Inoue, L. Y., Lin, D. W., Newcomb, L. F., Leonardson, A. S., Ankerst, D., Gulati, R., Carter, H. B., Trock, B. J., Carroll, P. R., Cooperberg, M. R., et al. (2018). Comparative analysis of biopsy upgrading in four prostate cancer active surveillance cohorts. *Annals of Internal Medicine*, 168(1):1–9.

Lin, H., McCulloch, C. E., Turnbull, B. W., Slate, E. H., and Clark, L. C. (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine*, 19(10):1303–1318.

Loeb, S., Carter, H. B., Schwartz, M., Fagerlin, A., Braithwaite, R. S., and Lepor, H. (2014). Heterogeneity in active surveillance protocols worldwide. *Reviews in Urology*, 16(4):202–203.

Nieboer, D., Tomer, A., Rizopoulos, D., Roobol, M. J., and Steyerberg, E. W. (2018). Active surveillance: a review of risk-based, dynamic monitoring. *Translational Andrology and Urology*, 7(1):106–115.

Pearson, J. D., Morrell, C. H., Landis, P. K., Carter, H. B., and Brant, L. J. (1994). Mixed-effects regression models for studying the natural history of prostate disease. *Statistics in Medicine*, 13(5-7):587–601.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press.

Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276.

Schröder, F., Hermanek, P., Denis, L., Fair, W., Gospodarowicz, M., and Pavone-Macaluso, M. (1992). The TNM classification of prostate cancer. *The Prostate*, 21(S4):129–138.

Telang, J. M., Lane, B. R., Cher, M. L., Miller, D. C., and Dupree, J. M. (2017). Prostate cancer family history and eligibility for active surveillance: a systematic review of the literature. *BJU International*, 120(4):464–467.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, 65(2):87–108.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.

Chapter 4

Personalized Schedules for Burdensome Surveillance Tests

This chapter is based on the paper
Tomer, A., Nieboer, D., Roobol, M.J., Steyerberg, E.W., and Rizopoulos, D. (2020), Personalized schedules for burdensome surveillance tests. Submitted to *Journal of the American Statistical Association*

Abstract

Benchmark surveillance *tests* for diagnosing disease *progression* (e.g., biopsies, endoscopies) in early-stage chronic non-communicable diseases (e.g., cancer, lung diseases) are usually invasive. For detecting progression timely, patients undergo invasive tests planned in a fixed one-size-fits-all manner (e.g., annually). We present personalized test schedules based on progression-risk, that aim to optimize the number of tests (burden) and time delay in detecting progression (shorter is beneficial) better than fixed schedules. Our motivation comes from the problem of scheduling biopsies in prostate cancer surveillance.

Using joint models for time-to-event and longitudinal data, we consolidate patients' longitudinal data (e.g., biomarkers) and results of previous tests, into individualized future cumulative-risk of progression. We then create personalized schedules by planning tests on future visits where the predicted cumulative-risk is above a *threshold* (e.g., 5% risk). We update personalized schedules with data gathered over follow-up. To find the optimal risk threshold, we minimize a utility function of the expected number of tests (burden) and expected time delay in detecting progression (shorter is beneficial) for different thresholds. We estimate these two in a patient-specific manner for following any schedule, by utilizing a patient's predicted risk profile. Patients/doctors can employ these quantities to compare personalized and fixed schedules objectively.

4.1 Introduction

Chronic non-communicable diseases (e.g., cancer, lung, cardiovascular diseases) cause 60–70% of human deaths worldwide (WHO et al., 2014). Often patients diagnosed with an early-stage disease undergo surveillance *tests* to detect disease *progression* timely. A progression is a non-terminal event, and usually a trigger for treatment and/or removal from surveillance. Benchmark tests used for confirming progression are usually *invasive*, e.g., biopsies in prostate cancer surveillance (Bokhorst et al., 2015), endoscopies in Barrett's esophagus (Weusten et al., 2017), colonoscopies in colorectal cancer (Krist et al., 2007), and bronchoscopies in post lung transplant (McWilliams et al., 2008) surveillance.

Invasive tests are repeated until progression is observed, typically as per a one-size-fits-all *fixed schedule*, e.g., biannually, (Krist et al., 2007; McWilliams et al., 2008; Bokhorst et al., 2015). A time gap between tests causes a time delay in detecting progression (Figure 4.1). A shorter delay in detecting progression (*benefit*) can provide a larger window of opportunity for curative treatment. However, with fixed schedules, this means conducting tests frequently. Frequent tests are *burdensome* as they may cause pain and/or severe medical complications (Krist et al., 2007; Loeb et al., 2013). Consequently, patients may not always comply with frequent tests (Bokhorst et al., 2015; Le Clercq et al., 2015). In general, because fixed schedules do not differentiate between fast and slow/non-progressing patients, they impose disproportionate burden/benefits across the patient population.

The goal of this work (Figure 4.1) is to optimize the number of invasive tests (burden) and the time delay in detecting progression (shorter is beneficial) better than fixed schedules. Specifically, we intend to *personalize* test schedules using patients' clinical data accumulated over surveillance follow-up. This data includes baseline characteristics, previous test results, and longitudinal outcomes (e.g., biomarkers, medical imaging, physical examination). Many surveillance protocols currently personalize test schedules using heuristic methods such as decision flowcharts (Bokhorst et al., 2015; Weusten et al., 2017). However, flowcharts discretize continuous outcomes,

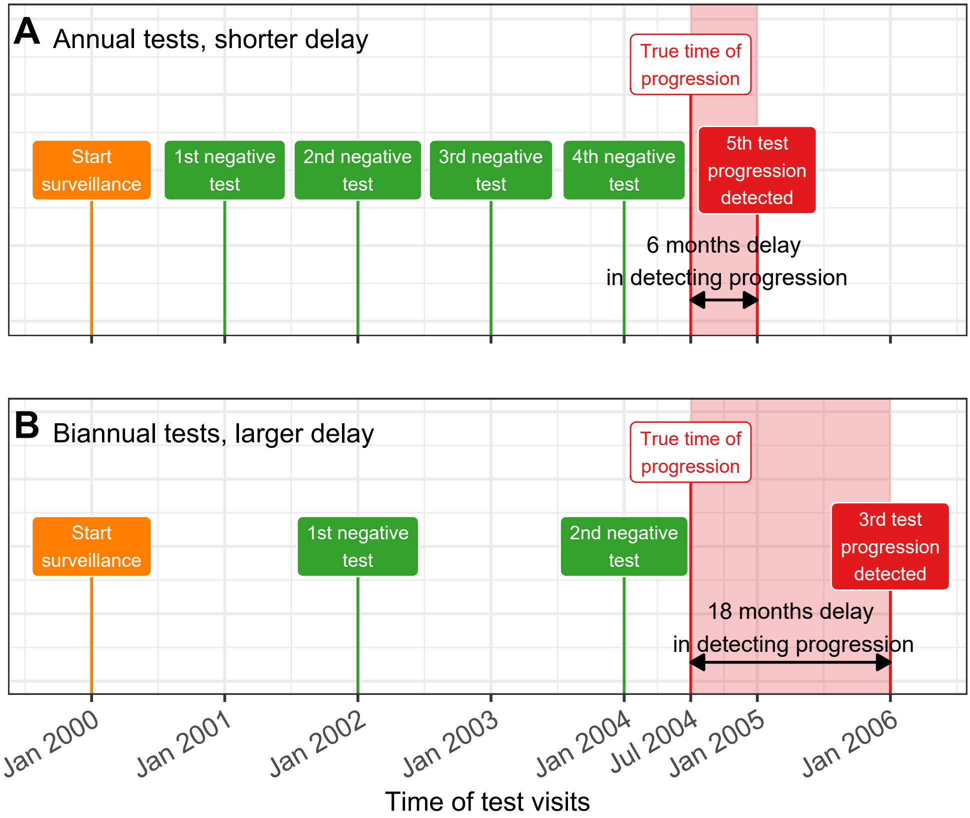

Figure 4.1: **Goal: Finding the optimal tradeoff between the number of invasive tests (burden) and time delay in detecting progression (shorter is beneficial)**. A progression is a non-terminal event in the surveillance of early-stage chronic non-communicable diseases. The true time of progression for the patient illustrated in this figure is July 2004. Since invasive tests are conducted repeatedly, progression is interval-censored and always observed with a delay. Frequent periodical invasive tests in **Panel A** lead to a shorter time delay in detecting progression than infrequent periodical invasive tests in **Panel B**. The interval-censored time of progression is Jan 2004–Jan 2005 in **Panel A** and between Jan 2004–Jan 2006 in **Panel B**.

often exploit only the last measurement, ignore the measurement error in observed data, and plan only one test at a time. Alternatively, a complete personalized schedule of tests can be obtained using partially observable Markov decision processes or POMDPs (Alagoz et al., 2010; Steimle and Denton, 2017). Although POMDPs typically discretize continuous longitudinal outcomes to avoid the curse of dimensionality. In scenarios such as ours, where decisions (test/no test) and disease state (low-grade disease/progressed) are both binary, POMDPs may not be necessary either. The reason is that such POMDPs give the same optimal schedule, which can be alternatively obtained by just planning a test when the probability of transition from non-progressed to progressed state is more than a certain threshold (see Vickers and Elkin, 2006, Equation 1).

Personalized schedules can also be obtained by optimizing an explicit utility function of the burden and/or benefit of a schedule. A challenge in this approach is quantifying burden and benefit. For a single test decision, Tomer et al. (2019a) quantify the burden and benefit as the time difference by which the test undershoots (unnecessary test) or overshoots (delayed detection) the true progression time of a patient, respectively. Whereas, for a complete test schedule, Bebu and Lachin (2017) quantify burden as the number of tests planned (or their cost), and benefit as short time delay in detecting progression. Although, unlike the number of tests, the costs of time delay in detecting progression are not always quantifiable. For this issue, Bebu and Lachin (2017), and Vickers and Elkin (2006) have proposed scheduling tests when the risk of progression is above a threshold. Risk-based methodologies has also been explored by Rizopoulos et al. (2015), and to evaluate the choice of risk thresholds Wang et al. (2019) and Tomer et al. (2019b) use measures of diagnostic accuracy (e.g., false-positive rate, true positive rate). However, a limitation of risk-based test decisions is that a single decision does not inform patients about the clinical consequences of continuing on surveillance. Also, measures of diagnostic accuracy are not personalized criteria for choosing risk thresholds.

We improve upon the works referenced above in many ways. Instead of a single risk-based test decision, we derive full risk-based test schedules that dynamically update with new clinical data over follow-up. Along with each schedule, we provide patients the clinical consequences of following it. Namely, the expected number of tests that will be required out of all planned tests to detect progression and the expected time delay in detecting progression. Unlike measures of diagnostic accuracy, we calculate these in a personalized manner. Also, these two are easily-quantifiable surrogates for important clinical aspects such as the window of opportunity for curative treatment, risk of adverse outcomes due to delayed detection of progression, financial costs of tests, risk of side-effects, and reduction in quality of life, etc. Our methodology is as follows. We first develop a full specification of the joint distribution of the patient-specific longitudinal outcomes and the time of progression. To this end, we utilize joint models for time-to-event and longitudinal data (Tsiatis and Davidian, 2004; Rizopoulos, 2012) because they are inherently personalized. Specifically, joint models utilize patient-specific random effects (McCulloch and Neuhaus, 2005) to model longitudinal outcomes without discretizing them. Subsequently, we input clinical data of a new patient into the fitted model to obtain their predicted patient-specific cumulative-risk of progression at future visits. We then create personalized schedules by planning tests on future visits where this predicted cumulative-risk is above a particular *threshold* (e.g., 5% risk). We automate the choice of this threshold and the resulting schedule. In particular, we optimize a utility function of the expected number of tests (burden) and time delay in detecting progression (shorter is beneficial) for personalized schedules. We estimate these two quantities for any given schedule in a patient-specific manner using the patient's predicted risk profile. Hence, patients/doctors can compare the consequences of opting for personalized versus fixed schedules objectively.

Our motivation comes from the problem of scheduling biopsies in the world's largest prostate cancer surveillance study, called Prostate Cancer Research International Active Surveillance (Bokhorst et al., 2015), or PRIAS. It has 7813 low/very-low grade cancer patients (1134 progressions, 104904 longitudinal measurements), many of whom are potentially over-diagnosed due to prostate-specific antigen (PSA) based screening (Loeb et al., 2014a). To reduce subsequent over-treatment, in surveillance, serious treatments (e.g., surgery, radiotherapy) are delayed until progression is observed. Surveillance involves regular monitoring of a patient's PSA (ng/mL), digital rectal examination or DRE (tumor shape/size), and biopsy Gleason grade group (Epstein et al., 2016). Among these, a biopsy Gleason grade group ≥ 2 is the reference test for confirming progression. Most often, biopsies are scheduled annually (Loeb et al., 2014b). However, such a frequent schedule can put an unnecessary burden on patients with slow/non-progressing cancers and cause non-compliance (Bokhorst et al., 2015). Since prostate cancer has the second-highest incidence among all cancers in males (Torre et al., 2015), individualized biopsy schedules can reduce the burden of biopsies in numerous patients worldwide.

The remaining paper is as follows. Section 4.2 introduces the joint modeling framework. We describe the personalized scheduling methodology in Section 4.3, and demonstrate them for prostate cancer surveillance patients in Section 4.4. In Section 4.5, we compare personalized and fixed schedules via a simulation study based on a joint model fitted to the PRIAS dataset.

4.2 Joint Model for Time-to-Progression and Longitudinal Outcomes

Let T_i^* denote the true time of disease progression for the i-th patient. Progression is always interval censored $l_i < T_i^* \leq r_i$ (Figure 4.1). Here, r_i and l_i denote the time of the last and second last invasive tests, respectively, when patients progress. In non-progressing patients, l_i denotes the time of the last test and $r_i = \infty$. Assuming K types of longitudinal outcomes, let $\boldsymbol{y}_{ki}$ denote the $n_{ki} \times 1$ longitudinal response vector of the k-th outcome, $k \in \{1, \ldots, K\}$. The observed data of all n patients is given by $\mathcal{A}_n = \{l_i, r_i, \boldsymbol{y}_{1i}, \ldots \boldsymbol{y}_{Ki}; i = 1, \ldots, n\}$.

4.2.1 Longitudinal Sub-process

To model multiple longitudinal outcomes in a unified framework, a joint model employs individual generalized linear mixed sub-models (McCulloch and Neuhaus, 2005). Specifically, the conditional distribution of the k-th outcome $\boldsymbol{y}_{ki}$ given a vector of patient-specific random effects $\boldsymbol{b}_{ki}$ is assumed to belong to the exponential family, with linear predictor given by,

$$g_k\Big[E\{y_{ki}(t) \mid \boldsymbol{b}_{ki}\}\Big] = m_{ki}(t) = \boldsymbol{x}_{ki}^\top(t)\boldsymbol{\beta}_k + \boldsymbol{z}_{ki}^\top(t)\boldsymbol{b}_{ki},$$

where $g_k(\cdot)$ denotes a known one-to-one monotonic link function, $y_{ki}(t)$ is the value of the k-th longitudinal outcome for the i-th patient at time t, and $\boldsymbol{x}_{ki}(t)$ and $\boldsymbol{z}_{ki}(t)$ are the time-dependent design vectors for the fixed $\boldsymbol{\beta}_k$ and random effects $\boldsymbol{b}_{ki}$, respectively. To model the correlation between different longitudinal outcomes, we link their corresponding random effects. Specifically, we assume that the vector of random effects $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}^\top, \ldots, \boldsymbol{b}_{Ki}^\top)^\top$ follows a multivariate normal distribution with mean zero and variance-covariance matrix W.

4.2.2 Survival Sub-process

In the survival sub-process, the hazard of progression $h_i(t)$ at a time t is assumed to depend on a function of patient and outcome-specific linear predictors $m_{ki}(t)$ and/or the random effects,

$$\begin{aligned} h_i\Big\{t \mid \mathcal{M}_i(t), \boldsymbol{w}_i(t)\Big\} = h_0(t) \exp \Bigg[&\boldsymbol{\gamma}^\top \boldsymbol{w}_i(t) \\ &+ \sum_{k=1}^{K} f_k\Big\{\mathcal{M}_{ki}(t), \boldsymbol{w}_i(t), \boldsymbol{b}_{ki}, \boldsymbol{\alpha}_k\Big\}\Bigg], \quad t > 0, \end{aligned}$$

where $h_0(\cdot)$ denotes the baseline hazard, $\mathcal{M}_{ki}(t) = \{m_{ki}(s) \mid 0 \leq s < t\}$ is the history of the k-th longitudinal process up to t, and $\boldsymbol{w}_i(t)$ is a vector of exogenous, possibly time-varying covariates with regression coefficients γ. Functions $f_k(\cdot)$, parameterized by vector of coefficients $\boldsymbol{\alpha_k}$, specify the features of each longitudinal outcome that are included in the linear predictor of the relative-risk model (Brown, 2009; Rizopoulos, 2012; Taylor et al., 2013). Some examples, motivated by the literature (subscripts k dropped for brevity), are,

$$\begin{cases} f\Big\{\mathcal{M}_i(t), \boldsymbol{w}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\Big\} = \alpha m_i(t), \\ f\Big\{\mathcal{M}_i(t), \boldsymbol{w}_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\Big\} = \alpha_1 m_i(t) + \alpha_2 m_i'(t), \quad \text{with } m_i'(t) = \frac{\mathrm{d}m_i(t)}{\mathrm{d}t}. \end{cases}$$

These formulations of $f(\cdot)$ postulate that the hazard of progression at time t may depend on underlying level $m_i(t)$ of the longitudinal outcome at t, or on both the level and velocity $m_i'(t)$ (e.g., PSA value and velocity in prostate cancer) of the outcome at t. Lastly, the baseline hazard $h_0(t)$ is modeled flexibly using P-splines (Eilers and Marx, 1996). Lastly, $h_0(t)$ is the baseline hazard at time t, and is modeled flexibly using P-splines (Eilers and Marx, 1996). More specifically:

$$\log h_0(t) = \gamma_{h_0,0} + \sum_{q=1}^{Q} \gamma_{h_0,q} B_q(t, \boldsymbol{v}),$$

where $B_q(t, \boldsymbol{v})$ denotes the q-th basis function of a B-spline with knots $\boldsymbol{v} = v_1, \ldots, v_Q$ and vector of spline coefficients γ_{h_0}. To avoid choosing the number and position of knots in the spline, a relatively high number of knots (e.g., 15 to 20) are chosen and the corresponding B-spline regression coefficients γ_{h_0} are penalized using a differences penalty (Eilers and Marx, 1996). The joint parameter estimation of the longitudinal and relative-risk sub-models using the Bayesian approach are presented in Appendix 4.A.

4.3 Personalized Schedule of Invasive Tests for Detecting Progression

4.3.1 Cumulative-risk of progression

Using the joint model fitted to the training data $\mathcal{A}_n$, we aim to derive a personalized schedule of invasive tests for a new patient j with true progression time T_j^*. To this end, our calculations exploit the *cumulative-risk* function. Let $t < T_j^*$ be the time of the last conducted test at which progression was not observed. Let $\{\mathcal{Y}_{1j}(v), \ldots, \mathcal{Y}_{Kj}(v)\}$ denote the history of observed longitudinal data up to the current visit time v. The current visit can be after the last negative test, i.e., $v \geq t$ (e.g., PSA after negative biopsy in prostate cancer). The cumulative-risk of progression for patient j at future time u is then given by,

$$\begin{aligned} R_j(u \mid t, v) &= \mathsf{Pr}\Big\{T_j^* \leq u \mid T_j^* > t, \mathcal{Y}_{1j}(v), \ldots, \mathcal{Y}_{Kj}(v), \mathcal{A}_n\Big\} \\ &= \int\int \mathsf{Pr}(T_j^* \leq u \mid T_j^* > t, \boldsymbol{b}_j, \boldsymbol{\theta}) \\ &\quad \times p\Big\{\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_{1j}(v), \ldots, \mathcal{Y}_{Kj}(v), \boldsymbol{\theta}\Big\} \\ &\quad \times p(\boldsymbol{\theta} \mid \mathcal{A}_n) \mathrm{d}\boldsymbol{b}_j \mathrm{d}\boldsymbol{\theta}, \quad u \geq t. \end{aligned} \tag{4.1}$$

The cumulative-risk function $R_j(\cdot)$ depends on patient-specific clinical data and the training dataset, via the posterior distribution of the random effects $\boldsymbol{b}_j$ and posterior distribution of the vector of all parameters $\boldsymbol{\theta}$ of the fitted joint model, respectively. This cumulative-risk function is dynamic, in the sense that it automatically updates over time as more longitudinal data become available (Figure 4.2).

4.3.2 Personalized Test Decision Rule

We intend to exploit the cumulative-risk function $R_j(\cdot)$ to develop a risk-based personalized schedule of invasive tests for the j-th patient. Typically, invasive tests are decided on the same visit times on which longitudinal data (e.g., biomarkers) are measured. Let $U = \{u_1, \ldots, u_L\}$ represent a schedule of such visits (e.g., biannual PSA measurement in prostate cancer). Here, $u_1 = v$ is also the current visit time. The maximum future visit time u_L can be chosen based on the available information in the training dataset $\mathcal{A}_n$. That is, tests for the new patient j are planned only up to a future visit time u_L at which a sufficient number of events in $\mathcal{A}_n$ are available for making

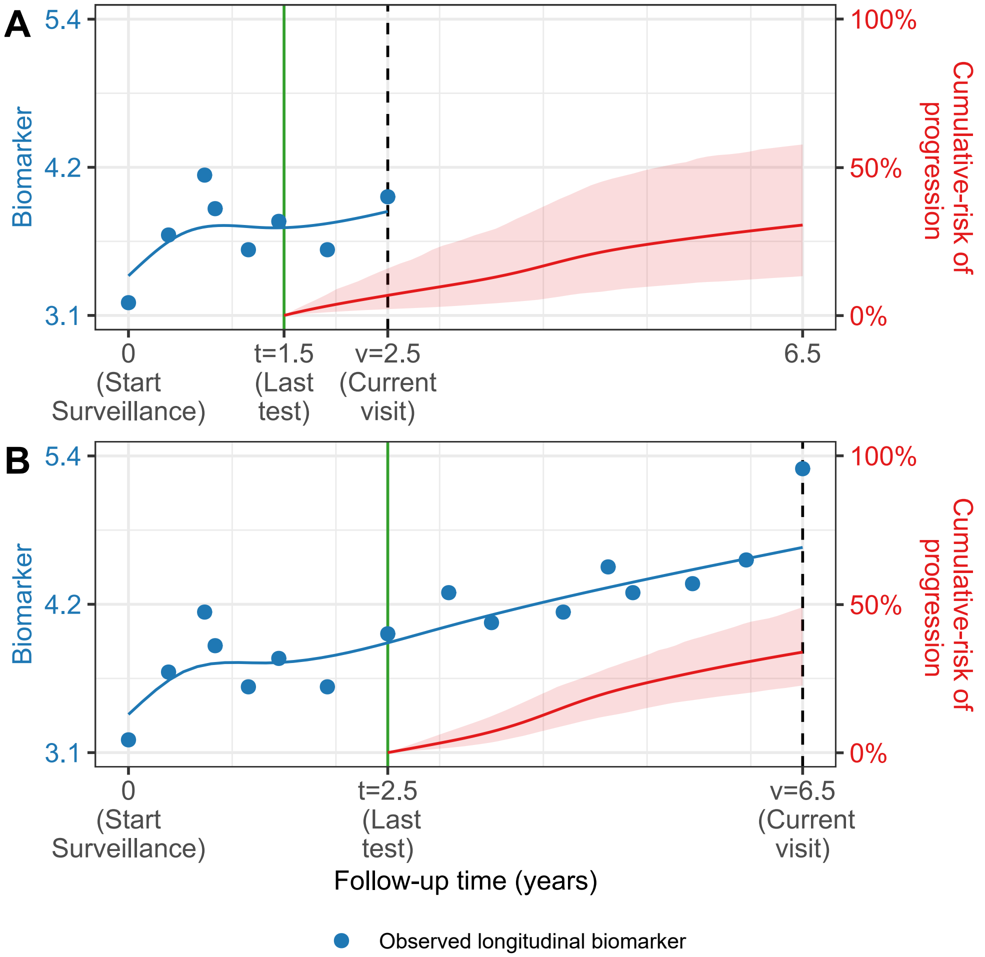

Figure 4.2: **Cumulative-risk of progression updated dynamically over follow-up** as more patient data is gathered. A single longitudinal outcome, namely, a continuous biomarker of disease progression, is used for illustration. **Panels A, B and C:** are ordered by the time of the current visit v (dashed vertical black line) of a new patient. At each of these visits, we combine the accumulated longitudinal data (shown in blue circles), and time of the last negative invasive test t (solid vertical green line) to obtain the updated cumulative-risk profile $R_j(u \mid t, v)$ (dotted red line with 95% credible interval shaded) of the patient defined in (4.1). All values are illustrative.

reliable risk predictions (e.g., up to the 80% or 90% percentile of progression times).

We propose to take the decision of conducting a test at a future visit time $u_l \in U$ if the cumulative-risk of progression at time u_l exceeds a certain risk threshold κ (Figure 4.3). In particular, the test decision at time u_l is given by,

$$Q_j^{\kappa}(u_l \mid t_l, v) = I\Big\{R_j(u_l \mid t_l, v) \geq \kappa\Big\}, \quad 0 \leq \kappa \leq 1, \tag{4.2}$$

where $I(\cdot)$ is the indicator function, $R_j(u_l \mid t_l, v)$ is the cumulative-risk of progression at the current decision time u_l, and $t_l < u_l$ is the time of the last test conducted before u_l. Thus, the future time at which a test will be planned, depends on both the threshold κ and the cumulative-risk of the patient. Moreover, when a test gets planned at time u_l, i.e., $Q_j^{\kappa}(u_l \mid t_l, v) = 1$, then the cumulative-risk profile is updated before making the next test decision at time u_{l+1} (Figure 4.3). Specifically, the cumulative-risk at time u_{l+1} is updated by setting the corresponding time of the last test $t_{l+1} = u_l$. This accounts for the possibility that progression may occur after time $u_l < T_j^*$. Hence, the time of last test t_l is defined as,

$$t_l = \begin{cases} t, & \text{if } l = 1,\\ u_{l-1}, & \text{if } l \geq 2 \text{ and } Q_j^{\kappa}(u_{l-1} \mid t_{l-1}, v) = 1,\\ t_{l-1}, & \text{if } l \geq 2 \text{ and } Q_j^{\kappa}(u_{l-1} \mid t_{l-1}, v) = 0. \end{cases}$$

We should note that in all future test decisions, we use only the observed longitudinal data up to the current visit time v, i.e., $\{\mathcal{Y}_{1j}(v), \ldots, Y_{Kj}(v)\}$.

4.3.3 Expected Number of Tests and Expected Time Delay in Detecting Progression

To facilitate shared-decision making of invasive tests, we translate our proposed decision rule, i.e., the choice of a specific risk threshold κ, into two clinically relevant quantities. First, the number of tests (burden) we expect

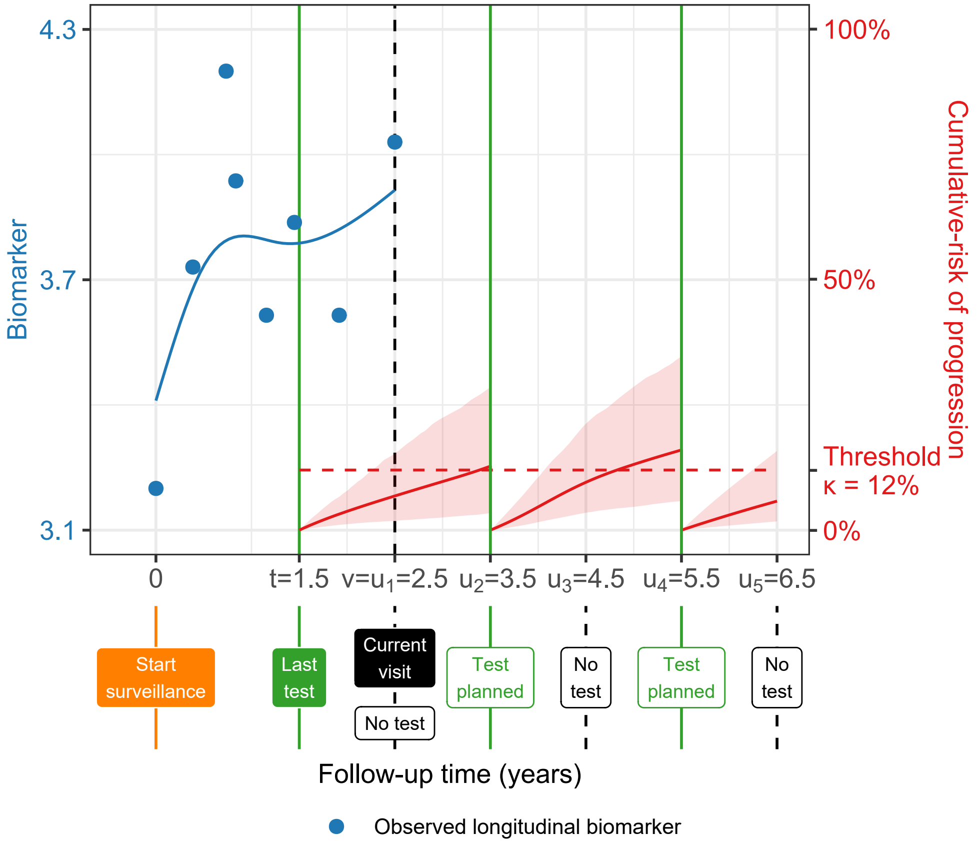

Figure 4.3: **Successive personalized test decisions based on patient-specific cumulative-risk of progression (4.2).** Time of current visit: $v = 2.5$ years (dashed vertical black line). Time of the last test on which progression was not observed: $t = 1.5$ years. Longitudinal data up to current visit: $\mathcal{Y}_j(v)$ is a continuous biomarker (blue circles). Example risk threshold: $\kappa = 0.12$ (12%). Grid of future visits on which future tests are planned: $U = \{2.5, 3.5, 4.5, 5.5, 6.5\}$ years. The cumulative-risk profiles $R_j(u_l \mid t_l, v)$ employed in (4.2) are shown with dotted red lines (95% credible intervals shaded), and are updated each time a test is planned (solid vertical green lines). Future test decisions $Q_j(u_l \mid t_l, v)$ defined in (4.2) are: $Q_j^\kappa(u_1 = 2.5 \mid t_1 = 1.5, v) = 0$, $Q_j^\kappa(u_2 = 3.5 \mid t_2 = 1.5, v) = 1$, $Q_j^\kappa(u_3 = 4.5 \mid t_2 = 3.5, v) = 0$, $Q_j^\kappa(u_4 = 5.5 \mid t_2 = 3.5, v) = 1$, and $Q_j^\kappa(u_5 = 6.5 \mid t_5 = 4.5, v) = 0$. All values are illustrative.

to perform for patient j, and second, if the patient progresses, the time delay (shorter is beneficial) expected in detecting progression. To calculate these two quantities, we first suppose that patient j does not progress between his last negative test at time t and the maximum future visit time u_L. Under this assumption, the subset of future visit times in U on which a test is planned using (4.2) results into a personalized schedule of future tests (Figure 4.3), given by:

$$\{s_1, \ldots, s_{N_j}\} = \Big\{u_l \in U : Q_j^\kappa(u_l \mid t_l, v) = 1\Big\}, \quad N_j \leq L. \tag{4.3}$$

If patient j never progressed in the period $[t, u_L]$, as we initially supposed, all N_j tests in $\{s_1, \ldots, s_{N_j}\}$ will be conducted. However, fewer tests will be performed if the patient did progress at some point $T_j^* < u_L$. We formally define the discrete random variable $\mathcal{N}_j$ denoting the number of performed tests in conjunction with the true progression time T_j^* as,

$$\mathcal{N}_j(S_j^\kappa) = \begin{cases} 1, & \text{if } \ t < T_j^* \leq s_1, \\ 2, & \text{if } \ s_1 < T_j^* \leq s_2, \\ \vdots & \\ N_j, & \text{if } \ s_{N_j - 1} < T_j^* \leq s_{N_j}, \end{cases}$$

where $S_j^\kappa = \{s_1, \ldots, s_{N_j}\}$ is the schedule of planned future tests. The expected number of future tests for patient j will be the expected value $E\Big\{\mathcal{N}_j(S_j^\kappa)\Big\}$, given by the expression,

$$E\Big\{\mathcal{N}_j(S_j^\kappa)\Big\} = \sum_{n=1}^{N_j} n \times \text{Pr}(s_{n-1} < T_j^* \leq s_n \mid T_j^* \leq s_{N_j}), \quad s_0 = t,$$

where

$$\text{Pr}(s_{n-1} < T_j^* \leq s_n \mid T_j^* \leq s_{N_j}) = \frac{R_j(s_n \mid t, v) - R_j(s_{n-1} \mid t, v)}{R_j(s_{N_j} \mid t, v)}.$$

Similarly, we can define the expected time delay in detecting progression, under the assumption that progression occurs before u_L. Specifically, the random variable time delay is equal to the difference between the time of the test at which progression is observed and the true time of progression T_j^*, and is given by,

$$\mathcal{D}_j(S_j^\kappa) = \begin{cases} s_1 - T_j^*, & \text{if } t < T_j^* \leq s_1, \\ s_2 - T_j^*, & \text{if } s_1 < T_j^* \leq s_2, \\ \vdots & \\ s_{N_j} - T_j^*, & \text{if } s_{N_j - 1} < T_j^* \leq s_{N_j}, \end{cases}$$

The expected time delay in detecting progression is the expected value of $\mathcal{D}_j(S_j^\kappa)$, given by the expression,

$$E\Big\{\mathcal{D}_j(S_j^\kappa)\Big\} = \sum_{n=1}^{N_j} \Big\{s_n - E(T_j^* \mid s_{n-1}, s_n, v)\Big\} \times \text{Pr}(s_{n-1} < T_j^* \leq s_n \mid T_j^* \leq s_N),$$

where $E(T_j^* \mid s_{n-1}, s_n, v)$ denotes the conditional expected time of progression for the scenario $s_{n-1} < T_j^* \leq s_n$ and is calculated as the area under the corresponding survival curve,

$$E(T_j^* \mid s_{n-1}, s_n, v) = s_{n-1} + \int_{s_{n-1}}^{s_n} \text{Pr}\Big\{T_j^* \geq u \mid s_{n-1} < T_j^* \leq s_n, \\ \mathcal{Y}_{1j}(v), \ldots, \mathcal{Y}_{Kj}(v), \mathcal{A}_n\Big\} \mathrm{d}u.$$

The personalized schedule in (4.3), and the corresponding personalized expected number of tests and time delay, have the advantage of getting updated with newly collected data over follow-up. Also, the expected number of tests and time delay can be calculated for any schedule, fixed or personalized. Hence, patients/doctors can use them to compare different schedules. Although, a fair comparison of time delays between different schedules for the same patient, requires a compulsory test at a common horizon time point in all schedules.

4.3.4 How to Select the Risk Threshold κ

The risk threshold κ controls the timing and the total number of invasive tests in the personalized schedule S_j^κ. Through the timing and the total number of planned tests, κ also indirectly affects the potential time delay (Figure 4.1) in detecting progression if a particular schedule is followed. Hence, κ should be chosen while balancing both the number of invasive tests (burden) and the time delay in detecting progression (shorter is beneficial).

To facilitate the choice of κ in practice, following our developments in the previous section, we translate the different choices for threshold κ into the expected number of tests and time delay. In particular, for a patient j having data available up to his current visit time v, we can construct a bi-dimensional Euclidean space of his expected total number of tests (x-axis) and expected time delay in detecting progression (y-axis), for different personalized test schedules obtained by varying κ in $[0, 1]$, e.g., Figure 4.4.

The ideal schedule for j-th patient is the one in which only one test is conducted, at exactly the true time of progression T_j^*. In other words, the time delay will be zero. If we weigh the expected number of tests and time delay as equally important, then we can select as the optimal threshold at current visit time v, the threshold $\kappa^*(v)$ which minimizes the Euclidean distance between the ideal schedule, i.e., point (1, 0) and the set of points representing the different personalized schedules S_j^κ corresponding to various $\kappa \in [0, 1]$, i.e.,

$$\kappa^*(v) = \arg\min_{0 \leq \kappa \leq 1} \sqrt{\Big[E\big\{\mathcal{N}_j(S_j^\kappa)\big\} - 1\Big]^2 + \Big[E\big\{\mathcal{D}_j(S^\kappa)\big\} - 0\Big]^2}. \quad (4.4)$$

In certain scenarios, patients/doctors may be apprehensive about undergoing more than a maximum expected number of future tests, or having an expected time delay higher than certain months. For such purposes, the Euclidean distance in (4.4) can be optimized under constraints on the expected number of tests or expected time delay (Figure 4.4). Doing so alleviates two problems, namely, that the time delay and the number of tests have different

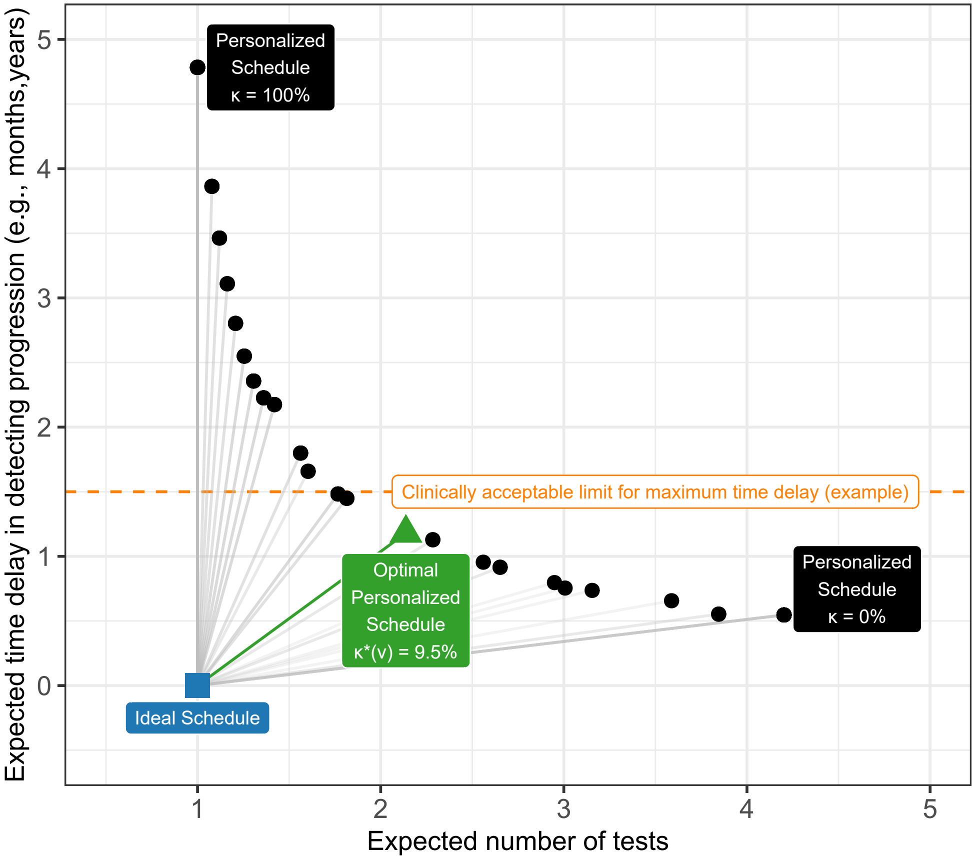

Figure 4.4: **Optimal current-visit time v specific risk threshold $\kappa^*(v)$ obtained using (4.4)** for the patient shown in Figure 4.3. Ideal schedule of tests: point (1,0) shown as a blue square. It plans exactly one invasive test at the true time of progression T_j^* of a patient. Hence, the time delay in detecting progression is zero. Various personalized schedules based on a grid of thresholds κ in $[0, 1]$ are shown with black circles. Higher thresholds lead to fewer tests, but also higher expected time delay. The personalized schedule based on $\kappa^*(v) = 9.5\%$ threshold (green triangle) has the least Euclidean distance (solid green line) to the ideal schedule. It is also possible to optimize the least distance under a certain clinically acceptable limit on the time delay (dotted horizontal orange line).

units of measurement, and that in (4.4) they are weighted equally (Cook and Wong, 1994).

We considered shorter delays in detecting progression as the benefit of repeated tests. However, it is also common to describe the benefit of testing in terms of decision-theoretic measures such as quality-adjusted life-years/expectancy (QALY/QALE) gained (Sassi, 2006). Optimizing (4.4) with QALE needs, setting the optimal point in a Euclidean space with QALE as a dimension, and obtaining expected QALEs for different schedules. For estimating the expected QALE in a personalized manner, a mathematical definition of QALE in terms of time delay $\mathcal{D}_j$ in detecting progression (de Carvalho et al., 2017b) is required.

4.4 Application of Personalized Schedules in Prostate Cancer Surveillance

We next demonstrate personalized schedules for scheduling biopsies in prostate cancer active surveillance. To this end, we use results from a joint model fitted to the PRIAS dataset introduced in Section 4.1. The model definition (Appendix 4.B.1) utilized a linear mixed sub-model for biannually measured PSA (continuous: log-transformed from ng/mL), and a logistic mixed sub-model for biannually measured DRE (binary: tumor palpable or not). In the survival sub-model, fitted PSA value, fitted instantaneous PSA velocity (defined in Section 4.2.2), and log-odds of having a DRE indicating a palpable tumor, were included as time-dependent predictors. The model parameters were estimated under the Bayesian framework using the R package **JMbayes** (Rizopoulos, 2016), and are presented in Appendix 4.B.1. We next briefly present the key results relevant for personalized scheduling.

First, the cause-specific cumulative-risk of cancer progression at the maximum study period of ten years was 50% (Figure 4.7). This indicates that many patients may not require all of the yearly biopsies they are usually prescribed. Since personalized schedules are risk-based, their overall performance is dependent on the predictive accuracy and discrimination capacity of the fitted model. In this regard, the model had a moderate time-dependent area under the receiver operating characteristic curve or AUC (Rizopoulos et al., 2017) over the follow-up period (between 0.61 and 0.68). The time-dependent mean absolute prediction error or MAPE (Rizopoulos et al., 2017) was moderate to large (between 0.08 and 0.24) and decreased rapidly after year one of the follow-up. Thus, personalized schedules based on this model may work better after year one with more follow-up data. Details on AUC and MAPE are provided in Appendix 5.C.1.

4.4.1 Personalized Biopsy Schedules for a Demonstration Prostate Cancer Patient

We utilized the joint model fitted to the PRIAS dataset to schedule biopsies in a demonstration prostate cancer patient shown in Figure 4.5. The time of his last negative biopsy was $t = 3.5$ years, and the time of the current visit was $v = 5$ years. We made biopsy decisions over his future visits for PSA measurement $U = \{u_1 = 5, u_2 = 5.5, \ldots, u_L = 10\}$ years using four different schedules. Two of the fixed schedules are annual biopsy schedule and the PRIAS schedule. The PRIAS schedule has compulsory biopsies at year one, four, seven, and ten of follow-up, and additional annual biopsies if PSA doubling-time (Bokhorst et al., 2015) is high. Remaining two schedules are personalized, namely, with a fixed threshold $\kappa = 10\%$ risk, and an automatically chosen current visit time v specific risk $\kappa^*(v)$ (Section 4.3.4). Since the demonstration patient's time of last negative biopsy $t = 3.5$ is after year one of follow-up, a time delay in detecting progression up to three years may not lead to adverse downstream outcomes (de Carvalho et al., 2017a).

The cumulative-risk of progression of the demonstration patient increases 3% yearly on average, up to 19% at the maximum study period of ten years. Hence, the patient may progress slowly. Consequently, risk-based personalized approaches plan fewer biopsies than the annual schedule (Panel B,

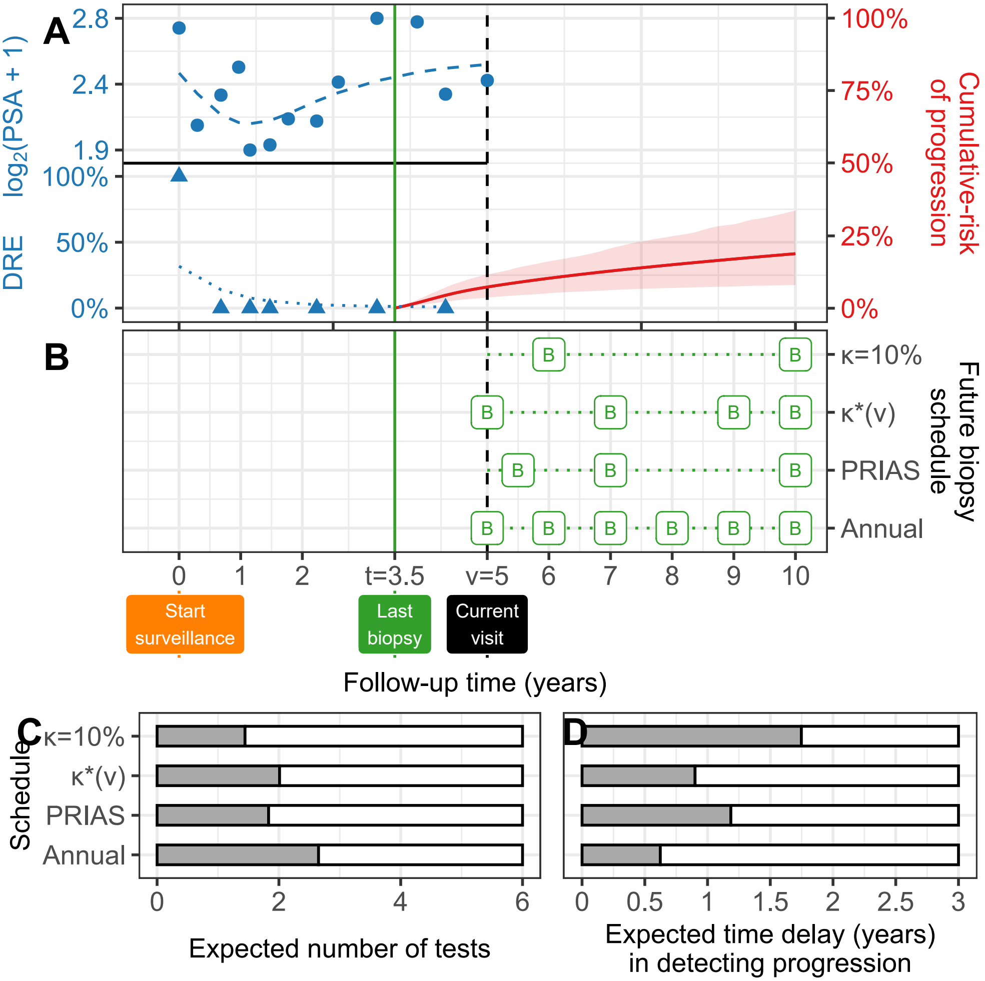

Figure 4.5: **Personalized schedules for a demonstration prostate cancer patient**. **Panel A**: Current visit: $v = 5$ years. Last negative biopsy: $t = 3.5$ years. Longitudinal data: $\log_2(\text{PSA} + 1)$ transformed (Tomer et al., 2019b) PSA (observed: blue dots, fitted: dashed blue line), binary DRE (observed: blue triangles, fitted probability: dotted blue line). Cumulative-risk profile: solid red line (95% credible interval shaded). **Panel B**: 'B' indicates a planned biopsy. $\kappa = 10\%$ and $\kappa^*(v)$ are personalized biopsy schedules using a risk threshold of 10%, and a visit time v specific automatic threshold (4.4), respectively. PRIAS biopsy schedule is defined in Section 4.4.1. **Panel C,D**: For all schedules we calculate the expected number of tests and expected time delay in detecting progression if the patient progresses before year ten. With a recommended minimum gap of one year between biopsies, maximum possible number of tests are six.

Figure 4.5). Also, the time delay in detecting progression for personalized schedules (Panel D, Figure 4.5) is below the safe limit of three years mentioned earlier. Thus, personalized schedules can be a suitable alternative to the annual schedule.

4.5 Simulation Study

Although we evaluated personalized schedules for a demonstration patient, we also intend to analyze and compare personalized and fixed schedules in a full cohort. Our criteria for comparison of schedules are the total number of invasive tests planned (burden), and the actual time delay in detecting progression (shorter is beneficial) for each schedule. Due to the periodical nature of schedules, the actual time delay in detecting progression cannot be observed in real-world surveillance. Hence, instead, we compare personalized versus fixed schedules via an extensive simulated randomized clinical trial in which each hypothetical patient undergoes each schedule. To keep our simulation study realistic, we employ the prostate cancer active surveillance scenario. Specifically, our simulated population is generated using the joint model fitted to the PRIAS cohort (Appendix 4.B.2).

4.5.1 Simulation Setup

From the simulation population, we first sample 500 datasets, each representing a hypothetical prostate cancer surveillance program with 1000 patients in it. We generate a true cancer progression time for each of the 500×1000 patients, and then sample longitudinal DRE and PSA measurements biannually (PRIAS protocol) for them. We split each dataset into training (750 patients) and test (250 patients) parts, and generate a random and non-informative censoring time for the training patients. All training and test patients also observe Type-I censoring at year ten of follow-up (current study period of PRIAS). We next fit a joint model of the same specification as the model fitted to PRIAS (Appendix 4.B.2), to each of the 500 training datasets and retrieve MCMC samples from the 500 sets of the posterior distribution of the parameters. In each of the 500 hypothetical surveillance programs, we utilize the corresponding fitted joint models to obtain the cumulative-risk of progression in each of the 500 × 250 test patients. These cumulative-risk profiles are further used to create personalized biopsy schedules for the test patients.

For each test patient, we conduct hypothetical biopsies using two fixed (PRIAS and annual schedule) and three personalized biopsy schedules. Personalized schedules are based on, a fixed risk threshold $\kappa = 10\%$, an optimal current visit time v specific threshold $\kappa^*(v)$ chosen via (4.4), and an optimal threshold obtained under the constraint that expected time delay in detecting progression is less than 0.75 years (9 months), denoted $\kappa^*\{v \mid E(\mathcal{D}) \leq 0.75\}$. The choice of 0.75 years delay constraint is arbitrary and is only used to illustrate that applying the constraint limits the average delay at 0.75 years. Successive personalized biopsy decisions are made only on the standard PSA follow-up visits, utilizing clinical data accumulated only until the corresponding current visit time (4.2). We maintain a minimum recommended gap of one year between consecutive prostate biopsies (Bokhorst et al., 2015) as well. Biopsies are conducted until progression is detected, or the maximum follow-up period at year ten (horizon) is reached. The actual time delay in detecting progression is equal to the difference in time at which progression is detected and the actual (simulated) time of progression of a patient.

4.5.2 Simulation Results

In the simulation study, nearly 50% of the patients observed progression during the ten year study period (*progressing*) and 50% did not (*non-progressing*). While we can calculate the total number of biopsies scheduled in all 500×250 test patients, the actual time delay in detecting progression is available only for progressing patients. Hence, we show the simulation results separately for progressing and non-progressing patients (Figure 4.6).

Before discussing delay in detecting progression (Panel A, Figure 4.6), we note that mean delay up to 1.7 years in all patients (Inoue et al., 2018), and up to three years in patients who progress after year one of follow-up (de Carvalho et al., 2017a), may not increase risks of adverse outcomes later. In this regard, the annual biopsies guarantee a maximum delay of one year in all patients. However, they also schedule the highest number of biopsies (Median 3, Inter-quartile range or IQR: 1–6). Much fewer biopsies are planned by the PRIAS schedule (Median 2, IQR: 1–4), but it also has a higher time delay (Median 0.74, IQR: 0.38–1.00 years). The personalized schedule based on optimal risk threshold $\kappa^*(v)$ schedules fewer biopsies than PRIAS and has a delay (Median 0.86, IQR: 0.46–1.26 years) slightly higher than PRIAS. The expected delay for risk threshold optimized with a constraint on expected delay $\kappa^*\{v \mid E(D) \leq 0.75\}$ is equal to 0.61 years, i.e., the constraint works as expected.

The simulated non-progressing patients (Panel B, Figure 4.6) gained the most with personalized schedules. The annual schedule plans 10 (unnecessary) biopsies for each such patient, and the PRIAS schedule plans a median of 6 (IQR: 4–8) biopsies. In contrast, the personalized schedule based on optimized risk threshold $\kappa^*(v)$ plans fewer biopsies consistently (Median 6, IQR: 6–7). The 10% threshold based schedule plans even fewer biopsies (Median 5, IQR: 4–6).

4.6 Discussion

In this paper, we presented a methodology to create personalized schedules for burdensome diagnostic *tests* used to detect disease *progression* in early-stage chronic non-communicable disease *surveillance*. For this purpose, we utilized joint models for time-to-event and longitudinal data. Our approach first combines a patient's clinical data (e.g., longitudinal biomarkers) and previous invasive test results to estimate patient-specific cumulative-risk of disease progression over their current and future follow-up visits. We then plan future invasive tests whenever this cumulative-risk of progression is

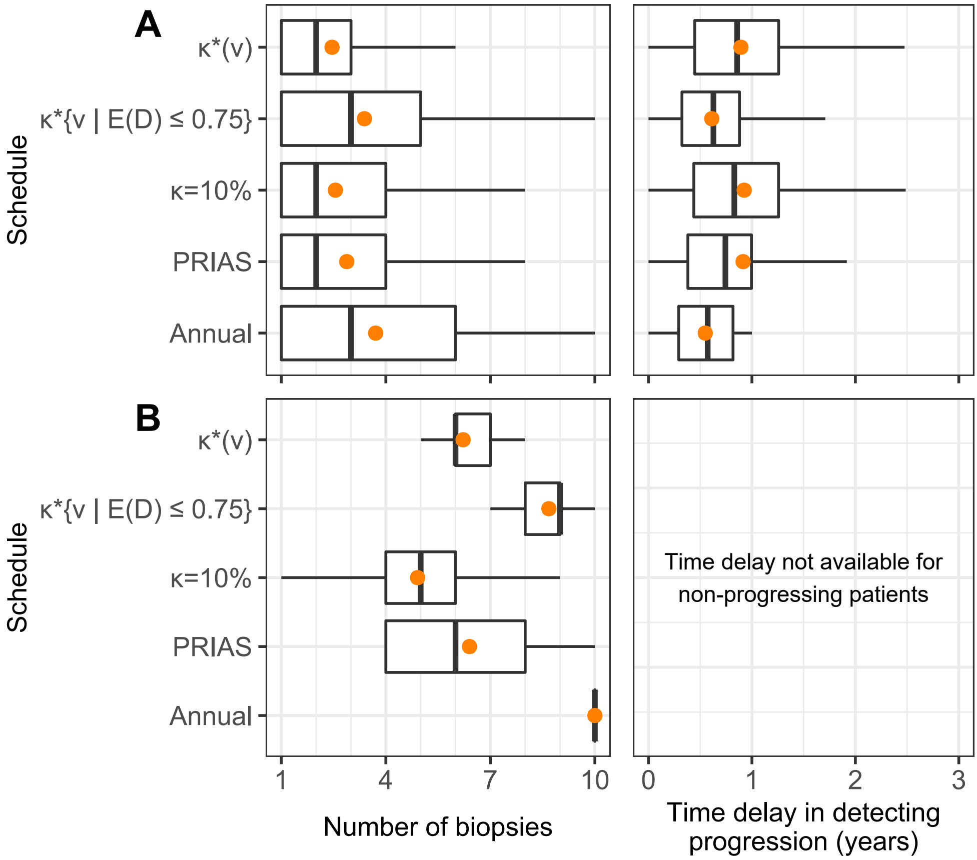

Figure 4.6: **Number of biopsies and the time delay in detecting cancer progression for various biopsy schedules** obtained via a simulation study. **Mean** is indicated by the orange circle. Time delay (years) is calculated as (time of positive biopsy - the actual simulated time of cancer progression). Biopsies are conducted until cancer progression is detected. **Panel A:** simulated patients who obtained cancer progression in the ten year study period (progressing). **Panel B:** simulated patients who did not obtain cancer progression in the ten year study period (non-progressing). Types of schedules: $\kappa = 10\%$ and $\kappa^*(v)$ schedule a biopsy if the cumulative-risk of cancer progression at the current visit time v is more than 10%, and an automatically chosen threshold (4.4), respectively. Schedule $\kappa^*\{v \mid E(\mathcal{D}) \leq 0.75\}$ is similar to $\kappa^*(v)$ except that the euclidean distance in (4.4) is minimized under the constraint that expected delay in detecting progression is at most 9 months (0.75 years). Annual corresponds to a schedule of yearly biopsies, and PRIAS corresponds to biopsies as per PRIAS protocol (Section 4.4).

predicted to be above a certain threshold. We select the risk threshold automatically in a personalized manner, by optimizing a utility function of the patient-specific consequences of choosing a particular risk threshold based schedule. These consequences are, namely, the number of invasive tests (burden) planned in a schedule, and the expected time delay in detection of progression (shorter is beneficial) if the patient progresses. Last, we calculate this expected time delay in a personalized manner for both personalized and fixed schedules to assist patients/doctors in making a more informed decision of choosing a test schedule.

Using joint models gives us certain advantages. First, since joint models employ random-effects, the corresponding risk-based schedules are inherently personalized. Second, to predict this patient-specific risk of progression, joint models utilize all observed longitudinal measurements of a patient. Also, the continuous longitudinal outcomes are not discretized, which is commonly a case in Markov Decision Process and flowchart-based test schedules. Third, personalized schedules update automatically with more patient data over follow-up. Fourth, we calculated the expected number of tests (burden) and expected time delay in detecting progression (shorter is beneficial) in a patient-specific manner. Using our methodology, these can be calculated for both personalized and fixed schedules. Thus, patients/doctors can compare risk-based and fixed schedules and choose one according to their preferences for the expected burden-benefit ratio. Last, although this work concerns invasive test schedules in disease surveillance, the methodology is generic for use under a screening setting as well.

Personalized schedules that we proposed require a risk threshold. We optimized the threshold choice using a generic utility function based on the expected number of biopsies and time delay in detecting progression. We used only these two measures because they are easy to interpret but simultaneously critical for deciding the timing of invasive tests. Also, the time delay in detecting progression is an easily-quantifiable surrogate for the window of opportunity for curative treatment and additional benefits of observing progression early. Practitioners may extend/modify our utility function by adding to/replacing time delay with commonly used decision-theoretic measures such as quality-adjusted life-years/expectancy (QALY/QALE).

We evaluated personalized schedules in a full cohort via a realistic simulation of a randomized clinical trial for prostate cancer surveillance patients. We observed that personalized schedules reduced many unnecessary biopsies for non-progressing patients compared to the widely used annual schedule. This happened at the cost of simultaneously having a slightly longer time delay in detecting progression. Although, this delay should still be safe because it was almost equal to the delay of the world's largest prostate cancer active surveillance program PRIAS's schedule. The simulation study results are by no means the performance-limit of the personalized schedules. Instead, models with higher predictive accuracy and discrimination capacity than the PRIAS based model may lead to an even better balance between the number of tests and the time delay in detecting progression. As for the practical usability of the PRIAS based model in prostate cancer surveillance, despite the moderate predictive performance, we expect this model's overall impact to be positive. There are two reasons for this. First, the risk of adverse outcomes because of personalized schedules is quite low because of the low rate of metastases and prostate cancer specific mortality in prostate cancer patients (Bokhorst et al., 2015). Second, studies (de Carvalho et al., 2017a; Inoue et al., 2018) have suggested that after the confirmatory biopsy at year one of follow-up, biopsies may be done as infrequently as every two to three years, with limited adverse consequences. In other words, longer delays in detecting progression may be acceptable after the first negative biopsy.

There are certain limitations to this work. First, in practice, most cohorts have a limited study period. Hence, the cumulative-risk profiles of patients and resulting personalized schedules can only be created up to the maximum study period. For this problem, the risk prediction model should be updated with more follow-up data over time. The proposed joint model assumed all events other than progression to be non-informative censoring. Alternative models that account for competing risks may lead to better results as they estimate absolute and not the cause-specific risk of progression. The detection of progression is susceptible to inter-observer variation, e.g., pathologists may grade the same biopsy differently. Progression is sometimes obscured due to sampling error, e.g., biopsy results vary based on location and number of biopsy cores. Although models that account for inter-observer variation (Balasubramanian and Lagakos, 2003) and sampling error (Coley et al., 2017) will provide better risk estimates, the methodology for obtained personalized schedules can remain the same.

Acknowledgements The first and last authors would like to acknowledge support by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (the national research council of the Netherlands) VIDI grant nr. 016.146.301, and Erasmus University Medical Center funding. Part of this work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. The authors also thank the Erasmus University Medical Center's Cancer Computational Biology Center for giving access to their IT-infrastructure and software that was used for the computations and data analysis in this study. Last, we would like to thank the PRIAS consortium for enabling this research project.

Appendix

4.A Parameter Estimation

We estimate the parameters of the joint model using Markov chain Monte Carlo (MCMC) methods under the Bayesian framework. Let $\boldsymbol{\theta}$ denote the vector of all of the parameters of the joint model. The joint model postulates that given the random effects, the time to progression, and all of the longitudinal measurements taken over time are all mutually independent. Under this assumption the posterior distribution of the parameters is given by:

$$\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D}_n) &\propto \prod_{i=1}^{n} p(l_i, r_i, \boldsymbol{y}_{1i}, \dots \boldsymbol{y}_{Ki}, \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\
&\propto \prod_{i=1}^{n} \prod_{k=1}^{K} p(l_i, r_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{y}_{ki} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}), \\
p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) &= \frac{1}{\sqrt{(2\pi)^{|W|} \text{det}(\boldsymbol{D})}} \exp(\boldsymbol{b}_i^\top \boldsymbol{D}^{-1} \boldsymbol{b}_i),
\end{aligned}$$

where, the likelihood contribution of the k-th longitudinal outcome vector $\boldsymbol{y}_{ki}$ for the i-th patient, conditional on the random effects is:

$$p(\boldsymbol{y}_{ki} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \prod_{j=1}^{n_{ki}} \exp\left[\frac{y_{kij}\psi_{kij}(\boldsymbol{b}_{ki}) - c_k\big\{\psi_{kij}(\boldsymbol{b}_{ki})\big\}}{a_k(\varphi)} - d_k(y_{kij}, \varphi)\right],$$

where n_{ki} are the total number of longitudinal measurements of type k for patient i. The natural and dispersion parameters of the exponential family are denoted by $\psi_{kij}(\boldsymbol{b}_{ki}$ and φ, respectively. In addition, $c_k(\cdot), a_k(\cdot), d_k(\cdot)$ are known functions specifying the member of the exponential family. The likelihood contribution of the time to progression outcome is given by:

$$\begin{aligned}
p(l_i, r_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \exp\Big[&-\int_0^{l_i} h_i\big\{s \mid \mathcal{M}_i(t), \boldsymbol{w}_i(t)\big\} \mathrm{d}s\Big] \\
&- \exp\Big[-\int_0^{r_i} h_i\big\{s \mid \mathcal{M}_i(t), \boldsymbol{w}_i(t)\big\} \mathrm{d}s\Big]. \qquad (4.5)
\end{aligned}$$

The integral in (4.5) does not have a closed-form solution, and therefore we use a 15-point Gauss-Kronrod quadrature rule to approximate it.

We use independent normal priors with zero mean and variance 100 for the fixed effect parameters of the longitudinal model. For scale parameters we inverse Gamma priors. For the variance-covariance matrix $\boldsymbol{D}$ of the random effects we take inverse Wishart prior with an identity scale matrix and degrees of freedom equal to the total number of random effects. For the relative risk model's parameters $\boldsymbol{\gamma}$ and the association parameters $\boldsymbol{\alpha}$, we use independent normal priors with zero mean and variance 100. However, when $\boldsymbol{\alpha}$ becomes high dimensional (e.g., when several functional forms are considered per longitudinal outcome), we opt for a global-local ridge-type shrinkage prior, i.e., for the s-th element of $\boldsymbol{\alpha}$ we assume:

$$\alpha_s \sim \mathcal{N}(0, \tau\psi_s),\ \tau^{-1} \sim \text{Gamma}(0.1, 0.1),\ \psi_s^{-1} \sim \text{Gamma}(1, 0.01). \quad (4.6)$$

The global smoothing parameter τ has sufficiently mass near zero to ensure shrinkage, while the local smoothing parameter ψ_s allows individual coefficients to attain large values. Other options of shrinkage or variable-selection priors could be used as well (Andrinopoulou and Rizopoulos, 2016). Finally, the penalized version of the B-spline approximation to the baseline hazard is specified using the following hierarchical prior for γ_{h_0} (Lang and Brezger, 2004):

$$p(\gamma_{h_0} \mid \tau_h) \propto \tau_h^{\rho(\boldsymbol{K})/2} \exp\Big(-\frac{\tau_h}{2}\gamma_{h_0}^\top \boldsymbol{K} \gamma_{h_0}\Big) \quad (4.7)$$

where τ_h is the smoothing parameter that takes a Gamma$(1, \tau_{h\delta})$ prior distribution, with a hyper-prior $\tau_{h\delta} \sim$ Gamma$(10^{-3}, 10^{-3})$, which ensures a proper posterior distribution for γ_{h_0} (Jullion and Lambert, 2007), $\boldsymbol{K} = \Delta_r^\top \Delta_r + 10^{-6}\boldsymbol{I}$, with Δ_r denoting the r-th difference penalty matrix, and $\rho(\boldsymbol{K})$ denotes the rank of $\boldsymbol{K}$.

4.B Joint Model for the PRIAS Dataset Used in Simulation Study

In this work, we reused a joint model we previously fitted to the PRIAS dataset (Tomer et al., 2019b, 2020). The PRIAS database is not openly accessible. However, access to the database can be requested on the basis of a study proposal approved by the PRIAS steering committee. The website of the PRIAS program is `www.prias-project.org`. For the sake of completeness and reproducibility of results, we have presented the PRIAS based model's definition and parameter estimates below. Figure 4.7 shows the cumulative-risk of progression over the follow-up period.

4.B.1 Model Specification

Let T_i^* denote the true progression time of the i-th patient included in PRIAS. Since biopsies are conducted periodically, T_i^* is observed with interval censoring $l_i < T_i^* \leq r_i$. When progression is observed for the patient at his latest biopsy time r_i, then l_i denotes the time of the second latest biopsy. Otherwise, l_i denotes the time of the latest biopsy and $r_i = \infty$. Let $\boldsymbol{y}_{di}$ and $\boldsymbol{y}_{pi}$ denote his observed DRE (digital rectal examination) and PSA (prostate-specific antigen) longitudinal measurements, respectively. The observed data of all n patients is denoted by $\mathcal{D}_n = \{l_i, r_i, \boldsymbol{y}_{di}, \boldsymbol{y}_{pi}; i = 1, \ldots, n\}$.

The patient-specific DRE and PSA measurements over time are modeled using a bivariate generalized linear mixed effects sub-model. The sub-model for DRE is given by:

$$\begin{aligned}\text{logit}\Big[\text{Pr}\{y_{di}(t) > \text{T1c}\}\Big] &= \beta_{0d} + b_{0di} + (\beta_{1d} + b_{1di})t\\ &\quad + \beta_{2d}(\text{Age}_i - 65) + \beta_{3d}(\text{Age}_i - 65)^2 \end{aligned} \tag{4.8}$$

where, t denotes the follow-up visit time, and Age_i is the age of the i-th patient at the time of inclusion in AS. The fixed effect parameters are denoted by $\{\beta_{0d}, \ldots, \beta_{3d}\}$, and $\{b_{0di}, b_{1di}\}$ are the patient specific random effects. With this definition, we assume that the patient-specific log odds of obtaining a DRE measurement larger than T1c (palpable tumor) remain linear over time.

The mixed effects sub-model for PSA is given by:

$$\begin{aligned}\log_2 \Big\{y_{pi}(t) + 1\Big\} &= m_{pi}(t) + \varepsilon_{pi}(t),\\ m_{pi}(t) &= \beta_{0p} + b_{0pi} + \sum_{k=1}^{3}(\beta_{kp} + b_{kpi})B_k(t, \mathcal{K})\\ &\quad + \beta_{4p}(\text{Age}_i - 65) + \beta_{5p}(\text{Age}_i - 65)^2, \end{aligned} \tag{4.9}$$

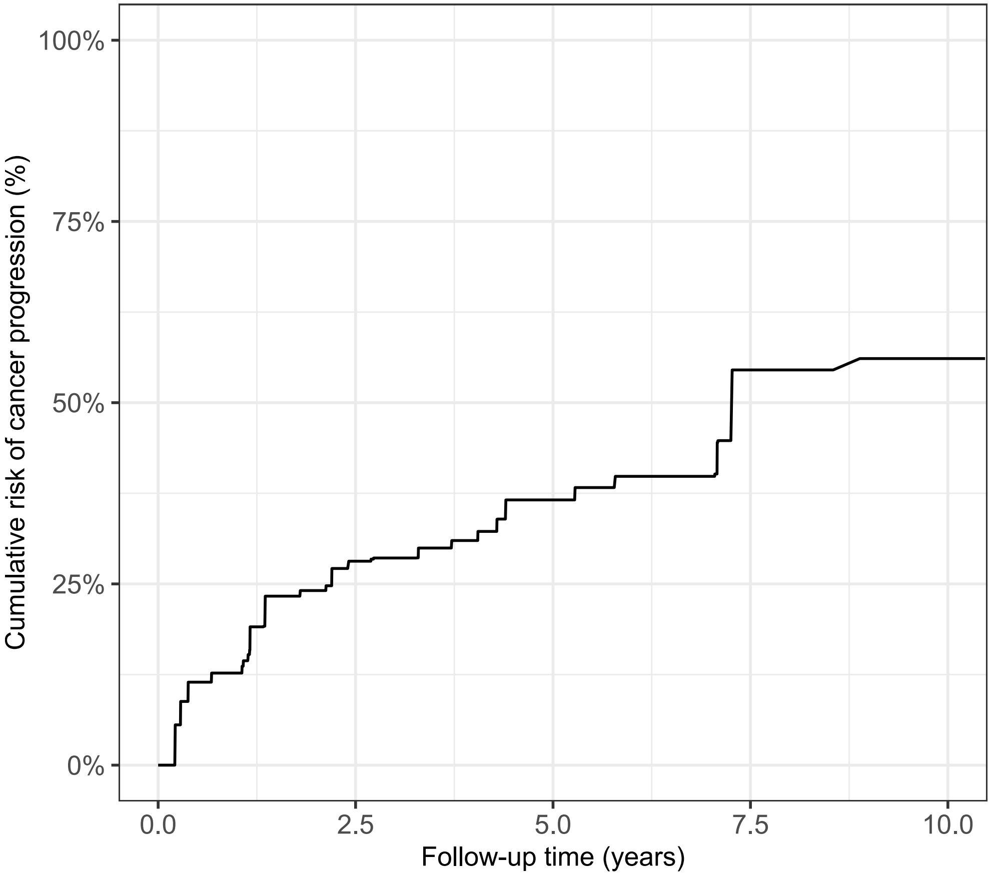

Figure 4.7: **Estimated cumulative-risk of cancer progression (Tomer et al., 2020)** for patients in the Prostate Cancer Research International Active Surveillance (PRIAS) dataset. Nearly 50% patients (*slow progressing*) do not progress in the ten year follow-up period. Cumulative-risk is estimated using nonparametric maximum likelihood estimation (Turnbull, 1976), to account for interval censored progression times observed in the PRIAS dataset. Censoring includes death, removal from surveillance on the basis of observed longitudinal data, and patient dropout.

where, $m_{pi}(t)$ denotes the underlying measurement error free value of the $\log_2(\text{PSA}+1)$ transformed (Tomer et al., 2019b) measurements at time t. We model it non-linearly over time using B-splines (De Boor, 1978). To this end, the B-spline basis function $B_k(t, \mathcal{K})$ has two internal knots at $\mathcal{K} = \{0.75, 2.12\}$ years (33-rd and 66-th percentile of observed follow-up times), and boundary knots at 0 and 6.4 years (95-th percentile of the observed follow-up times). The fixed effect parameters are denoted by $\{\beta_{0p}, \ldots, \beta_{5p}\}$, and $\{b_{0pi}, \ldots, b_{3pi}\}$ are the patient specific random effects. The error $\varepsilon_{pi}(t)$ is assumed to be t-distributed with three degrees of freedom (Tomer et al., 2019b) and scale σ, and is independent of the random effects.

To account for the correlation between the DRE and PSA measurements of a patient, link their corresponding random effects are linked. Specifically, the complete vector of random effects $\boldsymbol{b}_i = (b_{0di}, b_{1di}, b_{0pi}, \ldots, b_{3pi})^\top$ is assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{W}$.

To model the impact of DRE and PSA measurements on the risk of progression, the joint model uses a relative risk sub-model. More specifically, the hazard of progression $h_i(t)$ at a time t is given by:

$$\begin{aligned} h_i(t) = h_0(t) \exp \Big(&\gamma_1(\text{Age}_i - 65) + \gamma_2(\text{Age}_i - 65)^2 \\ &+ \alpha_{1d}\text{logit}\Big[\text{Pr}\{y_{di}(t) > \text{T1c}\}\Big] + \alpha_{1p} m_{pi}(t) + \alpha_{2p} \frac{\partial m_{pi}(t)}{\partial t}\Big), \end{aligned} \tag{4.10}$$

where, γ_1, γ_2 are the parameters for the effect of age. The parameter α_{1d} models the impact of log odds of obtaining a DRE $>$ T1c on the hazard of progression. The impact of PSA on the hazard of progression is modeled in two ways: a) the impact of the error free underlying PSA value $m_{pi}(t)$, and b) the impact of the underlying PSA velocity $\partial m_{pi}(t)/\partial t$. The corresponding parameters are α_{1p} and α_{2p}, respectively. Lastly, $h_0(t)$ is the baseline hazard at time t, and is modeled flexibly using P-splines (Eilers and Marx, 1996).

Table 4.1: Estimated variance-covariance matrix $\boldsymbol{W}$ of the random effects $\boldsymbol{b} = (b_{0d}, b_{1d}, b_{0p}, b_{1p}, b_{2p}, b_{3p})$ from the joint model fitted to the PRIAS dataset.

Random Effects	b_{0d}	b_{1d}	b_{0p}	b_{1p}	b_{2p}	b_{3p}
b_{0d}	9.233	-0.183	-0.213	0.082	0.058	0.023
b_{1d}	-0.183	1.259	0.091	0.079	0.145	0.109
b_{0p}	-0.213	0.091	0.247	0.007	0.067	0.018
b_{1p}	0.082	0.079	0.007	0.248	0.264	0.189
b_{2p}	0.058	0.145	0.067	0.264	0.511	0.327
b_{3p}	0.023	0.109	0.018	0.189	0.327	0.380

Table 4.2: Estimated mean and 95% credible interval for the parameters of the longitudinal sub-model (4.8) for the DRE outcome.

Variable	Mean	Std. Dev	2.5%	97.5%
(Intercept)	-4.407	0.151	-4.716	-4.113
(Age $- 65$)	0.057	0.009	0.039	0.075
(Age $- 65)^2$	-0.002	0.001	-0.004	0.000
visitTimeYears	-1.089	0.113	-1.292	-0.866

4.B.2 Parameter Estimates

The posterior parameter estimates for the PRIAS based joint model are shown in Table 4.2 (longitudinal sub-model for DRE outcome), Table 4.3 (longitudinal sub-model for PSA outcome) and Table 4.4 (relative risk sub-model). The parameter estimates for the variance-covariance matrix $\boldsymbol{W}$ from the longitudinal sub-model are shown in the following Table 4.1:

For the relative risk sub-model (4.10), the parameter estimates in Table 4.4 show that both $\log_2(\text{PSA} + 1)$ velocity, and the log odds of having DRE $>$ T1c were significantly associated with the hazard of progression.

Table 4.3: Estimated mean and 95% credible interval for the parameters of the longitudinal sub-model (4.9) for the PSA outcome.

Variable	Mean	Std. Dev	2.5%	97.5%
(Intercept)	2.687	0.007	2.674	2.701
(Age $-$ 65)	0.008	0.001	0.006	0.010
(Age $-$ $65)^2$	-0.001	0.000	-0.001	0.000
Spline: [0.00, 0.75] years	0.199	0.009	0.181	0.217
Spline: [0.75, 2.12] years	0.293	0.012	0.269	0.316
Spline: [2.12, 6.4] years	0.379	0.014	0.352	0.406
σ	0.144	0.001	0.142	0.145

Table 4.4: Estimated mean and 95% credible interval for the parameters of the relative risk sub-model (4.10) of the joint model fitted to the PRIAS dataset.

Variable	Mean	Std. Dev	2.5%	97.5%
(Age $-$ 65)	0.034	0.005	0.025	0.043
(Age $-$ $65)^2$	0.000	0.001	-0.001	0.001
logit$\{$Pr(DRE $>$ T1c)$\}$	0.047	0.014	0.018	0.073
Fitted $\log_2$(PSA $+ 1)$ value	0.024	0.076	-0.125	0.170
Fitted $\log_2$(PSA $+ 1)$ velocity	2.656	0.291	2.090	3.236

As described in Section 4.A the baseline hazard of the joint model model utilized a cubic P-spline. The knots of this P-spline were placed at the following time points: 0.000, 0.000, 0.000, 0.000, 0.401, 0.801, 1.202, 1.603, 2.003, 2.404, 2.805, 3.205, 3.606, 4.007, 4.407, 4.808, 5.209, 12.542, 12.542, 12.542, 12.542 The parameters of the fitted spline function are given in Table 4.5.

Data of the demonstration patient in Figure 4.5 is available in Table 4.6.

Table 4.5: Estimated parameters of the P-spline function utilized to model the baseline hazard $h_0(t)$ in joint model fitted to the PRIAS dataset. Parameters are named with the prefix 'ps' indicating P-spline parameter.

Variable	Mean	Std. Dev	2.5%	97.5%
ps1	-1.091	0.535	-2.286	-0.235
ps2	-2.113	0.271	-2.638	-1.591
ps3	-2.486	0.308	-3.095	-1.883
ps4	-2.083	0.311	-2.740	-1.483
ps5	-1.918	0.279	-2.460	-1.388
ps6	-2.620	0.265	-3.138	-2.140
ps7	-3.169	0.303	-3.796	-2.580
ps8	-3.416	0.340	-4.075	-2.823
ps9	-3.432	0.345	-4.103	-2.796
ps10	-3.223	0.352	-3.997	-2.573
ps11	-2.840	0.349	-3.577	-2.214
ps12	-2.481	0.350	-3.148	-1.762
ps13	-2.540	0.352	-3.206	-1.840
ps14	-2.841	0.321	-3.447	-2.212
ps15	-3.046	0.381	-3.853	-2.328
ps16	-3.113	0.701	-4.533	-1.796
ps17	-3.195	1.232	-5.894	-0.978

4.C Risk Based Schedules Versus All Possible Schedules

In Section 4.3.2, we let $U = \{u_1, \ldots, u_L\}$ represent a schedule of pre-fixed future visits (e.g., biannual PSA measurement in prostate cancer) on which we wanted to decide for conducting future invasive test. Since each test decision is binary, given L visits in U, a total of 2^L test schedules can be created. Risk-based personalized schedules obtained via (4.2) constitute a small subset of these 2^L schedules. In search for an optimal schedule

Table 4.6: Data of the demonstration patient in Figure 4.5. Age of the patient at baseline was 60 years and time of last negative biopsy was 3.5 years. DRE: digital rectal examination.

Visit time (years)	PSA	$\log_2(\text{PSA} + 1)$	DRE $>$ T1c
0.00	5.7	2.77	1
0.30	3.2	2.09	-
0.68	4.0	2.30	0
0.97	4.6	2.50	-
1.15	2.9	1.92	0
1.47	3.0	1.95	0
1.77	3.3	2.14	-
2.23	3.5	2.12	0
2.58	4.4	2.39	-
3.21	6.1	2.84	0
3.86	5.9	2.81	-
4.32	3.9	2.31	0
5.00	4.4	2.41	-

(Section 4.3.4), we create a Euclidean space of all possible 2^L schedules and not just risk-based schedules (Figure 4.8).

4.D Simulation Study Extended Results

In the simulation study, we evaluated the following biopsy schedules (Loeb et al., 2014b; Inoue et al., 2018): biopsy every year (annual), biopsy according to the PRIAS schedule (PRIAS), personalized biopsy schedules based on two fixed risk thresholds, namely, $\kappa = 10\%$, and automatically chosen optimal $\kappa^*(v)$ (Section 4.3), and automatically chosen optimal $\kappa^*\{v \mid E(D) \leq 0.75\}$ with a constraint of 9 months (0.75 years) on expected delay in detecting progression. Lastly, we added two more optimal schedules to this list. Specifically, $S^*(v)$ denotes an optimal schedule among all possible schedules (Sec-

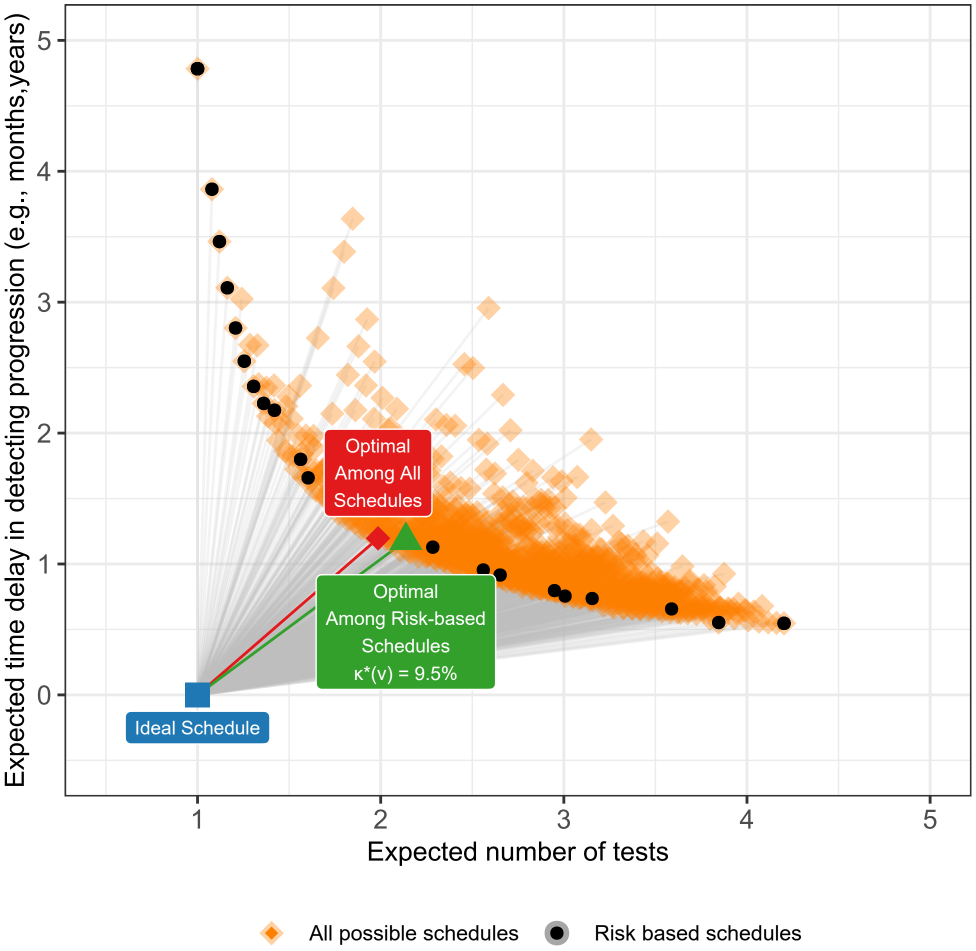

Figure 4.8: **Extension of Figure 4.4, by optimizing the Euclidean distance (4.4) among all possible schedules**. Let $U = \{u_1, \dots, u_L\}$ represent a schedule of pre-fixed future visits (e.g., biannual PSA measurement in prostate cancer) on which we wanted to decide for conducting future invasive test. Since each test decision is binary, given L visits in U, a total of 2^L test schedules can be created (orange Rhombus). Risk-based personalized schedules obtained via (4.2) and shown by black circles, constitute a small subset of these 2^L schedules. Ideal schedule of tests: point (1,0) shown as a blue square. It plans exactly one invasive test at the true time of progression T_j^* of a patient. That is, zero time delay in detecting progression.

tion 4.C), and $S^*\{v \mid E(D) \leq 0.75\}$ is extension of $S^*(v)$ with a constraint of 9 months (0.75 years) on expected delay in detecting progression.

We compare all the aforementioned schedules on two criteria, namely the number of biopsies they schedule and the corresponding time delay in detection of cancer progression in years (time of positive biopsy - true time of cancer progression). The corresponding results, using 500 × 250 test patients are presented in Table 4.7. Since the simulated cohorts are based on PRIAS, roughly only 50% of the patients progress in the ten year study period. While we are able to calculate the total number of biopsies scheduled in all 500×250 test patients, but the time delay in detection of progression is available only for those patients who progress in ten years (*progressing*). Hence, we show the simulation results separately for *progressing* and *non-progressing* patients.

4.E Partially Observable Markov Decision Processes

Partially observable Markov decision processes or POMDPs have been utilized in numerous optimal screening and surveillance test schedules for chronic diseases (Steimle and Denton, 2017), and especially for nearly all types of cancers (Alagoz et al., 2010). A notable advantage of POMDPs is that they find an optimal schedule from all schedules possible over a set of follow-up visits. In our case, this means all 2^L possible schedules given visit schedule $U = \{u_1, \dots, u_L\}$. To our knowledge, POMDPs and joint models have not been integrated yet. Thus, our aim is to integrate them to make the definition of POMDPs personalized, and then evaluate their strengths and limitations. The components of our discrete-time space POMDP are as follows (subscript j denotes the subject).

Decision epochs: The decision epoch $u \in U$ is the time at which we want to take a decision of an invasive test. These are typically pre-fixed future

Table 4.7: **Simulation study results for all patients**: Estimated mean (μ), median (Med), first quartile Q_1, and third quartile Q_3 for number of biopsies (nb) and for the time delay (d) in detection of cancer progression in years, for various biopsy schedules. The delay is equal to the difference between the time of the positive biopsy and the simulated true time of progression. Types of schedules: $\kappa = 10\%$ and $\kappa^*(v)$ schedule a biopsy if the cumulative-risk of cancer progression at a visit is more than 10%, and an automatically chosen threshold, respectively. Schedule $\kappa^*\{v \mid E(D) \leq 0.75\}$ is an extension of $\kappa^*(v)$ with a constraint of 9 months (0.75 years) on expected delay in detecting progression. $S^*(v)$ denotes an optimal schedule among all possible schedules (Section 4.C), and $S^*\{v \mid E(D) \leq 0.75\}$ is an extension of $S^*(v)$ with a constraint of 9 months (0.75 years) on expected delay in detecting progression. Annual corresponds to a schedule of yearly biopsies, and PRIAS corresponds to biopsies as per PRIAS protocol.

Schedule	Q_1^{nb}	μ^{nb}	Med^{nb}	Q_3^{nb}	Q_1^{d}	μ^{d}	Med^{d}	Q_3^{d}
Progressing patients (50%)								
Annual	1	3.71	3	6	0.29	0.55	0.57	0.82
PRIAS	1	2.88	2	4	0.38	0.92	0.74	1.00
$\kappa = 10\%$	1	2.55	2	4	0.45	1.00	0.85	1.33
$\kappa^*(v)$	1	2.46	2	3	0.45	0.89	0.86	1.26
$\kappa^*\{v \mid E(D) \leq 0.75\}$	1	3.39	3	5	0.32	0.61	0.63	0.88
$S^*(v)$	1	2.07	2	3	0.55	1.06	1.01	1.49
$S^*\{v \mid E(D) \leq 0.75\}$	1	2.79	2	4	0.39	0.75	0.76	1.06
Non-progressing patients (50%)								
Annual	10	10.00	10	10	-	-	-	-
PRIAS	4	6.40	6	8	-	-	-	-
$\kappa = 10\%$	4	4.91	5	6	-	-	-	-
$\kappa^*(v)$	6	6.22	6	7	-	-	-	-
$\kappa^*\{v \mid E(D) \leq 0.75\}$	8	8.68	9	9	-	-	-	-
$S^*(v)$	5	6.49	5	6	-	-	-	-
$S^*\{v \mid E(D) \leq 0.75\}$	7	7.22	7	7	-	-	-	-

follow-up visits for biomarker measurements (Section 4.3.2).

Actions: Two types actions can be taken at each decision epoch u, namely, an invasive test IT or waiting until the next decision epoch W. The action taken at time u is denoted by $q(u) \in \{IT, W\}$. The history of all actions taken until time u is $Q(u) = \{q(0), \dots q(u)\}$.

(Disease) States: At decision epoch u, the disease state of the patient is denoted by $s_j(u) \in S$. The vector of all states $S_j = \{P, NP, R\}$, where P denotes that the patient has obtained disease progression (event of interest), and NP denotes that patient has not obtained progression. Unlike progression P and not progression NP, the third state R called removal from surveillance, is observable. Removal of a patient from surveillance occurs only after progression P has been observed. The state R is also an absorbing state, and hence it always transitions to itself, irrespective of the action taken.

In joint modeling terms, $s_j(u) = NP$ is equivalent to $T_j^* > u$ (right-censored), and $s_j(u) = P$ means $u^- < T_j^* \leq u$ (interval-censored), where $u^- = \max\{v \mid s_j(v) = NP, v < u\}$ is the time of the last visit on which an invasive test was conducted to confirm that the patient had not progressed.

Observations: We cannot observe the underlying disease states P and NP unless we take the action invasive test IT. However, the disease state is manifested by observable clinical data, e.g., PSA and DRE in prostate cancer. Specifically, we can observe a K-tuple of clinical data $\boldsymbol{y}_j(u) = \{y_{1j}(u), \dots y_{Kj}(u)\}$ on each decision epoch u to guide our actions. When the patient is in state R (removed from surveillance) a special observation tuple $\boldsymbol{y}_j(u) = (\phi, \dots, \phi)$, denoting empty data, is observed. The observation history at time u is given by $\mathcal{Y}_j(u) = \{\boldsymbol{y}_j(0) \dots \boldsymbol{y}_j(u)\}$.

Typically POMDPs make two assumptions about clinical observations. First, that observations are categorical in nature. This is done to avoid the curse of dimensionality. Second, at any time $u - 1$ the probability distribution of future observations $p\{\boldsymbol{y}_j(u) \mid s_j(u)\}$ is assumed independent of the observation history. This means that probability distribution of future observations adds unique information over observed data. Conversely, in the joint modeling framework continuous observations are allowed, and probability distribution of future observation $p\{\boldsymbol{y}_j(u) \mid \boldsymbol{b}_j\}$ depends entirely on the patient-specific random effects $\boldsymbol{b}_j$ that are estimated from the observed data $p\{\boldsymbol{b}_j \mid s_j(u), \mathcal{Y}_j(u-1)\}$. That is, the probability distribution of future observations adds no extra value over observed data and current disease state. Hence, hereafter we denote longitudinal data history as $\mathcal{Y}_j$ and do not specify the time up to which it is observed.

Belief: The states P, NP cannot be observed directly. In this regard, our belief regarding what state the patient is in, is given by the corresponding probability of being in a certain state. The vector of these probabilities is called the belief vector. It is given by,

$$\begin{aligned}\pi_j(u) = \Big[&\text{Pr}\Big\{s_j(u) = P \mid \mathcal{Y}_j, Q(u-1)\Big\},\\ &\text{Pr}\Big\{s_j(u) = NP \mid \mathcal{Y}_j, Q(u-1)\Big\},\\ &\text{Pr}\Big\{s_j(u) = R \mid \mathcal{Y}_j, Q(u-1)\Big\}\Big].\end{aligned}$$

The sum $\sum_{s_j(u) \in S} \text{Pr}\{s_j(u) \mid \mathcal{Y}_j, Q(u-1)\} = 1$ of these probabilities is always equal to one. Since the state R removed from surveillance can be observed directly, $\text{Pr}\{s_j(u) = R \mid \mathcal{Y}_j, Q_j(u-1)\} \in \{0, 1\}$.

The belief vector is calculated on the basis of both the current observation, previous belief, latest action, and transition probabilities from previous state to current state. To this end, POMDPs utilize the Bayes rule (Steimle and Denton, 2017). In contrast, in joint modeling framework the disease state distribution is estimated as random-effects, and subsequently the belief is expressed as the probability distribution of the time to event outcome. Specifically,

$$\begin{aligned}\text{Pr}\Big\{s_j(u) = NP \mid \mathcal{Y}_j, Q(u-1)\Big\} &= \text{Pr}(T_j^* > u \mid T_j^* > u^-, \mathcal{Y}_j),\\ \text{Pr}\Big\{s_j(u) = P \mid \mathcal{Y}_j, Q(u-1)\Big\} &= \text{Pr}(T_j^* \leq u \mid T_j^* > u^-, \mathcal{Y}_j). \qquad (4.11)\end{aligned}$$

Transition Probabilities: Patient's disease state changes over follow-up, and also with actions. For example, if an action $q(u) = IT$ is taken when the patient is in state $s_j(u) = P$, then the state at next decision time $u+1$ is removal from surveillance, i.e, $s_j(u+1) = R$. However, if $s_j(u) = NP$, then invasive test action IT or waiting W do not change state. This is because, disease transition from not progressed to progressed is a natural process, and not altered by invasive tests. Transition from one state to another happens with a certain probability. This probability can be obtained from the joint model, as shown in Table 4.8.

Table 4.8: State transition matrix for a POMDP

Action: Invasive Test $q(u) = IT$			
$\pi_j(u+1)$	$s_j(u) = NP$	$s_j(u) = P$	$s_j(u) = R$
$s_j(u+1) = P$	$1 - \text{Pr}(T_j^* > u+1 \mid T_j^* > u, \mathcal{Y}_j)$	0	0
$s_j(u+1) = NP$	$\text{Pr}(T_j^* > u+1 \mid T_j^* > u, \mathcal{Y}_j)$	0	0
$s_j(u+1) = R$	0	1	1
Action: Waiting $q(u) = W$			
$\pi_j(u+1)$	$s_j(u) = NP$	$s_j(u) = P$	$s_j(u) = R$
$s_j(u+1) = P$	$1 - \text{Pr}(T_j^* > u+1 \mid T_j^* > u, \mathcal{Y}_j)$	1	0
$s_j(u+1) = NP$	$\text{Pr}(T_j^* > u+1 \mid T_j^* > u, \mathcal{Y}_j)$	0	0
$s_j(u+1) = R$	0	0	1

The state transition matrix in Table 4.8 can also be seen as a belief transition matrix. Specifically, given the current state $s_j(u)$ and history of actions $q(u)$ we can obtain a new belief. For example, if $s_j(u) = NP$, and $q(u) = IT$ then belief vector at next epoch $u + 1$ is given by:

$$\pi_j\Big\{u + 1 \mid s_j(u) = NP, q(u) = IT\Big\} = \Big[1 - \mathsf{Pr}(T_j^* > u + 1 \mid T_j^* > u, \mathcal{Y}_j), \\ \mathsf{Pr}(T_j^* > u + 1 \mid T_j^* > u, \mathcal{Y}_j), 0\Big]. \tag{4.12}$$

Reward: The criterion of optimality in POMDPs is the weighted cumulative reward. A reward is a number that is chosen manually for four possible outcomes (true-positive, false-positive, true-negative, and false-negative) of a binary test/no test decision in a schedule. The weighted cumulative reward of a schedule is the weighted sum of all rewards possible with all sequential test decisions in a schedule. Let us assume the following immediate rewards for action $q(u)$, conditional on knowing the state of the patient and the data. These are denoted by $B\{q(u) \mid s_j(u)\}$ and exemplified in Table 4.9. Since the state of the patient is unobservable, the weighted reward of an action at time u is:

Table 4.9: Reward matrix for a POMDP

	$q(u) = IT$	$q(u) = W$
$s_j(u) = NP$	a (unnecessary test)	b (saved unnecessary test)
$s_j(u) = P$	c (correct test)	d (skipped necessary test)
$s_j(u) = R$	0	0

$$B\Big\{q(u) \mid \pi_j(u)\Big\} = \sum_{s \in S} B\Big\{q(u) \mid s_j(u) = s\Big\} \mathsf{Pr}\Big\{s_j(u) = s \mid \mathcal{Y}_j, Q(u - 1)\Big\}.$$

The weights $\mathsf{Pr}\Big\{s_j(u) = s \mid \mathcal{Y}_j, Q(u - 1)\Big\}$ are defined in (4.11).

Dynamic Programming Equations: The dynamic programming equations for our POMDP, starting from a belief $\pi_j(u)$ at time u is given by (condition on observed data $\mathcal{Y}_j(u)$ and action history $Q(u-1)$ is dropped for brevity, but assumed):

$$
\begin{aligned}
V\Big\{\pi_j(u)\Big\} &= B\Big\{q^*(u) \mid \pi_j(u)\Big\} + \rho \sum_{s \in S} \text{Pr}\Big\{s_j(u) = s \mid \mathcal{Y}_j(u), Q(u-1)\Big\} \\
&\qquad \times V\Big[\pi_j\{u+1 \mid s_j(u) = s, q^*(u)\}\Big], \\
q^*(u) &= \underset{q(u) \in \{IT, W\}}{\text{arg max}} \Big\{B\Big\{q(u) \mid \pi_j(u)\Big\} + \rho \sum_{s \in S} \text{Pr}\Big\{s_j(u) = s \mid \mathcal{Y}_j(u), Q(u-1)\Big\} \\
&\qquad \times V\Big[\pi_j\{u+1 \mid s_j(u) = s, q(u)\}\Big]\Big\}
\end{aligned}
$$

where $0 \leq \rho \leq 1$ is the discount factor with the interpretation that it is the probability of rewards at and after time u being useful, and future belief $\pi_j\{u+1 \mid s_j(u) = s, q(u)\}$ is defined in (4.12).

4.E.1 Choice of Reward Function for POMDPs

Consider the scenario that patient has not been detected in progressed P state yet. That is, there is zero probability of being the third state R called removed from surveillance. In this scenario, the belief vector at any time u is given by $\pi_j(u) = (p, 1-p, 0)$, where $p = \text{Pr}\Big\{s_j(u) = P \mid \mathcal{Y}_j(u), Q(u-1)\Big\}$ is the probability that the patient is currently in state P. If we calculate the weighted reward of a single action (that is not looking ahead in time), it is given by,

$$
\begin{aligned}
B\Big\{q(u) = IT\Big\} &= c \times p + a \times (1-p), \\
B\Big\{q(u) = W\Big\} &= d \times p + b \times (1-p). \qquad (4.13)
\end{aligned}
$$

The action invasive test IT will be taken if reward of test is more than reward of waiting, i.e., $B\Big\{q(u) = IT\Big\} > B\Big\{q(u) = W\Big\}$. Thus using (4.13), we can say that if $p > (b-a)/(c-d+b-a)$, then action IT will be taken. The right hand side $(b-a)/(c-d+b-a)$ is a constant. Infinite combinations of rewards a, b, c, d can satisfy the condition $p > (b-a)/(c-d+b-a)$. Typically POMDP rewards are chosen based on survey results (Steimle and Denton, 2017) and translated as quality-adjusted life-years (QALY) saved. However, the main concern is that with infinite optimal reward sets, any reward set can be cherry-picked, including those that correspond to (improbable) thousands of quality-adjusted life-years saved. Besides the estimates for QALYs are usually not personalized before use in the model.

4.F Source Code

The source code for reproducing results of this chapter is available at `https://github.com/anirudhtomer/PersonalizedSchedules`.

4.7 References

Alagoz, O., Ayer, T., and Erenay, F. S. (2010). Operations research models for cancer screening. *Wiley Encyclopedia of Operations Research and Management Science*.

Andrinopoulou, E.-R. and Rizopoulos, D. (2016). Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures. *Statistics in Medicine*, 35(26):4813–4823.

Balasubramanian, R. and Lagakos, S. W. (2003). Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika*, 90(1):171–182.

Bebu, I. and Lachin, J. M. (2017). Optimal screening schedules for disease progression with application to diabetic retinopathy. *Biostatistics*, 19(1):1–13.

Bokhorst, L. P., Alberts, A. R., Rannikko, A., Valdagni, R., Pickles, T., Kakehi, Y., Bangma, C. H., Roobol, M. J., and PRIAS study group (2015). Compliance rates with the Prostate Cancer Research International Active Surveillance (PRIAS) protocol and disease reclassification in noncompliers. *European Urology*, 68(5):814–821.

Brown, E. R. (2009). Assessing the association between trends in a biomarker and risk of event with an application in pediatric HIV/AIDS. *The Annals of Applied Statistics*, 3(3):1163–1182.

Coley, R. Y., Zeger, S. L., Mamawala, M., Pienta, K. J., and Carter, H. B. (2017). Prediction of the pathologic Gleason score to inform a personalized management program for prostate cancer. *European Urology*, 72(1):135–141.

Cook, R. D. and Wong, W. K. (1994). On the equivalence of constrained and compound optimal designs. *Journal of the American Statistical Association*, 89(426):687–692.

De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

de Carvalho, T. M., Heijnsdijk, E. A., and de Koning, H. J. (2017a). Estimating the risks and benefits of active surveillance protocols for prostate cancer: a microsimulation study. *BJU International*, 119(4):560–566.

de Carvalho, T. M., Heijnsdijk, E. A., and de Koning, H. J. (2017b). When should active surveillance for prostate cancer stop if no progression is detected? *The Prostate*, 77(9):962–969.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., and Humphrey, P. A. (2016). The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 40(2):244–252.

Inoue, L. Y., Lin, D. W., Newcomb, L. F., Leonardson, A. S., Ankerst, D., Gulati, R., Carter, H. B., Trock, B. J., Carroll, P. R., Cooperberg, M. R., et al. (2018). Comparative analysis of biopsy upgrading in four prostate cancer active surveillance cohorts. *Annals of Internal Medicine*, 168(1):1–9.

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive bayesian p-splines models. *Computational Statistics & Data Analysis*, 51(5):2542–2558.

Krist, A. H., Jones, R. M., Woolf, S. H., Woessner, S. E., Merenstein, D., Kerns, J. W., Foliaco, W., and Jackson, P. (2007). Timing of repeat colonoscopy: disparity between guidelines and endoscopists' recommendation. *American Journal of Preventive Medicine*, 33(6):471–478.

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

Le Clercq, C., Winkens, B., Bakker, C., Keulen, E., Beets, G., Masclee, A., and Sanduleanu, S. (2015). Metachronous colorectal cancers result from missed lesions and non-compliance with surveillance. *Gastrointestinal Endoscopy*, 82(2):325–333.e2.

Loeb, S., Bjurlin, M. A., Nicholson, J., Tammela, T. L., Penson, D. F., Carter, H. B., Carroll, P., and Etzioni, R. (2014a). Overdiagnosis and overtreatment of prostate cancer. *European Urology*, 65(6):1046–1055.

Loeb, S., Carter, H. B., Schwartz, M., Fagerlin, A., Braithwaite, R. S., and Lepor, H. (2014b). Heterogeneity in active surveillance protocols worldwide. *Reviews in Urology*, 16(4):202–203.

Loeb, S., Vellekoop, A., Ahmed, H. U., Catto, J., Emberton, M., Nam, R., Rosario, D. J., Scattoni, V., and Lotan, Y. (2013). Systematic review of complications of prostate biopsy. *European Urology*, 64(6):876–892.

McCulloch, C. E. and Neuhaus, J. M. (2005). Generalized linear mixed models. *Encyclopedia of Biostatistics*, 4.

McWilliams, T. J., Williams, T. J., Whitford, H. M., and Snell, G. I. (2008). Surveillance bronchoscopy in lung transplant recipients: risk versus benefit. *The Journal of Heart and Lung Transplantation*, 27(11):1203–1209.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software*, 72(7):1–46.

Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276.

Rizopoulos, D., Taylor, J. M., Van Rosmalen, J., Steyerberg, E. W., and Takkenberg, J. J. (2015). Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics*, 17(1):149–164.

Sassi, F. (2006). Calculating QALYs, comparing QALY and DALY calculations. *Health Policy and Planning*, 21(5):402–408.

Steimle, L. N. and Denton, B. T. (2017). Markov decision processes for screening and treatment of chronic diseases. In *Markov Decision Processes in Practice*, pages 189–222. Springer.

Taylor, J. M., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T., and Sandler, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1):206–213.

Tomer, A., Nieboer, D., Roobol, M. J., Bjartell, A., Steyerberg, E. W., Rizopoulos, D., and et al. (2020). Risk of upgrading based personalized biopsy schedules for prostate cancer active surveillance patients. -, manuscript submitted to British Journal of Urology International(-):–.

Tomer, A., Nieboer, D., Roobol, M. J., Steyerberg, E. W., and Rizopoulos, D. (2019a). Personalized schedules for surveillance of low-risk prostate cancer patients. *Biometrics*, 75(1):153–162.

Tomer, A., Rizopoulos, D., Nieboer, D., Drost, F.-J., Roobol, M. J., and Steyerberg, E. W. (2019b). Personalized decision making for biopsies in prostate cancer active surveillance programs. *Medical Decision Making*, 39(5):499–508.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, 65(2):87–108.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.

Wang, Y., Zhao, Y.-Q., and Zheng, Y. (2019). Learning-based biomarker-assisted rules for optimized clinical benefit under a risk-constraint. *Biometrics*.

Weusten, B., Bisschops, R., Coron, E., Dinis-Ribeiro, M., Dumonceau, J.-M., Esteban, J.-M., Hassan, C., Pech, O., Repici, A., Bergman, J., et al. (2017). Endoscopic management of Barrett's esophagus: European Society of Gastrointestinal Endoscopy (ESGE) position statement. *Endoscopy*, 49(02):191–198.

WHO, W. H. O. et al. (2014). *Global status report on noncommunicable diseases 2014*. Number WHO/NMH/NVI/15.1. World Health Organization.

Part II

Application

Chapter 5

Personalized Biopsy Schedules Based on Risk of Gleason Upgrading for Low-Risk Prostate Cancer Active Surveillance Patients

This chapter is based on the paper
Tomer, A., Nieboer, D., Roobol, M.J., Bjartell, A., Steyerberg, E.W., and Rizopoulos, D. (2020), Personalized Biopsy Schedules Based on Risk of Gleason Upgrading for Low-Risk Prostate Cancer Active Surveillance Patients. *BJU International*. Advance online publication. doi:10.1111/bju.15136

Abstract

Objective: To develop a model and methodology for predicting the risk of Gleason *upgrading* in prostate cancer active surveillance (AS) patients, and using the predicted risks to create risk-based *personalized* biopsy schedules as an alternative to one-size-fits-all schedules (e.g., annually). Furthermore, to assist patients and doctors in making shared decisions of biopsy schedules, by providing them quantitative estimates of the *burden* and *benefit* of opting for personalized versus any other schedule in AS. Last, to externally validate our model and implement it along with personalized schedules in a ready to use web-application.

Materials and Methods: Repeat prostate-specific antigen (PSA) measurements, timing and results of previous biopsies, and age at baseline from the world's largest AS study, Prostate Cancer Research International Active Surveillance or PRIAS (7813 patients, 1134 experienced upgrading). We fitted a Bayesian joint model for time-to-event and longitudinal data to this dataset. We then validated our model externally in the largest six AS cohorts of the Movember Foundation's Global Action Plan (GAP3) database ($> 20,000$ patients, 27 centers worldwide). Using the model predicted upgrading-risks, we scheduled biopsies whenever a patient's upgrading-risk was above a certain threshold. To assist patients/doctors in choice of this threshold, and to compare the resulting personalized schedule with currently practiced schedules, along with the timing and the total number of biopsies (burden) planned, for each schedule we provided them the time delay expected in detecting upgrading (shorter is better).

Results: The cause-specific cumulative upgrading-risk at year five of follow-up was 35% in PRIAS, and at most 50% in GAP3 cohorts. In the PRIAS based model, PSA velocity was a stronger predictor of upgrading (Hazard Ratio: 2.47, 95%CI: 1.93–2.99) than PSA value (Hazard Ratio: 0.99, 95%CI: 0.89–1.11). Our model had a moderate area under the receiver operating characteristic curve (0.6–0.7) in validation cohorts. The prediction error was moderate (0.1–0.2) in validation cohorts where the impact of PSA value and velocity on upgrading-risk was similar to PRIAS, but large (0.2–0.3) otherwise. Our model required recalibration of baseline upgrading-risk in validation cohorts. We implemented the validated models and the methodology for personalized schedules in a web-application (`http://tiny.cc/biopsy`).

Conclusions: We successfully developed and validated a model for predicting upgrading-risk, and providing risk-based personalized biopsy decisions, in prostate cancer AS. Personalized prostate biopsies are a novel alternative to fixed one-size-fits-all schedules that may help to reduce unnecessary prostate biopsies while maintaining cancer control. The model and schedules made available via a web-application enable shared decision making of biopsy schedules by comparing fixed and personalized schedules on total biopsies and expected time delay in detecting upgrading.

5.1 Introduction

Patients with low- and very low-risk screening-detected localized prostate cancer are recommended active surveillance (AS) usually, instead of immediate radical treatment (Briganti et al., 2018). In AS, cancer progression is monitored routinely via prostate-specific antigen (PSA), digital rectal examination (DRE), repeat biopsies, and recently, magnetic resonance imaging (MRI). Among these, the strongest indicator of cancer-related outcomes is the biopsy Gleason grade group (Epstein et al., 2016). When it increases from group 1 (Gleason 3+3) to 2 (Gleason 3+4) or higher, it is called *upgrading* (Bruinsma et al., 2017). Upgrading is an important endpoint in AS upon which patients are commonly advised curative treatment (Bul et al., 2013).

Biopsies in AS are always conducted with a time gap between them. Consequently, upgrading is always detected with a time delay (Figure 5.1) that cannot be measured directly. In this regard, to detect upgrading timely, many patients are prescribed fixed and frequent biopsies, most often annually Loeb et al. (2014). However, such one-size-fits-all schedules lead to unnecessary biopsies in slow/non-progressing patients. Biopsies are invasive, may be painful, and are prone to medical complications such as bleeding and septicemia(Loeb et al., 2013). Thus, biopsy burden and patient non-compliance to frequent biopsies (Bokhorst et al., 2015) have raised concerns regarding the optimal biopsy schedule (Inoue et al., 2018; Bratt et al., 2013) in AS.

Except for the confirmatory biopsy at year one of AS (Bokhorst et al., 2015), opinions and practice regarding the timing of remaining biopsies lack agreement (Nieboer et al., 2018). Some AS programs utilize patients' observed PSA, DRE, previous biopsy Gleason grade, and lately, MRI results to decide biopsies (Kasivisvanathan et al., 2020; Bul et al., 2013; Nieboer et al., 2018). In contrast, others discourage schedules based on clinical data and MRI results (Chesnut et al., 2020; Loeb et al., 2014), and instead support periodical one-size-fits-all biopsy schedules. Furthermore, some suggest replacing frequent periodical schedules with infrequent ones (e.g., bienni-

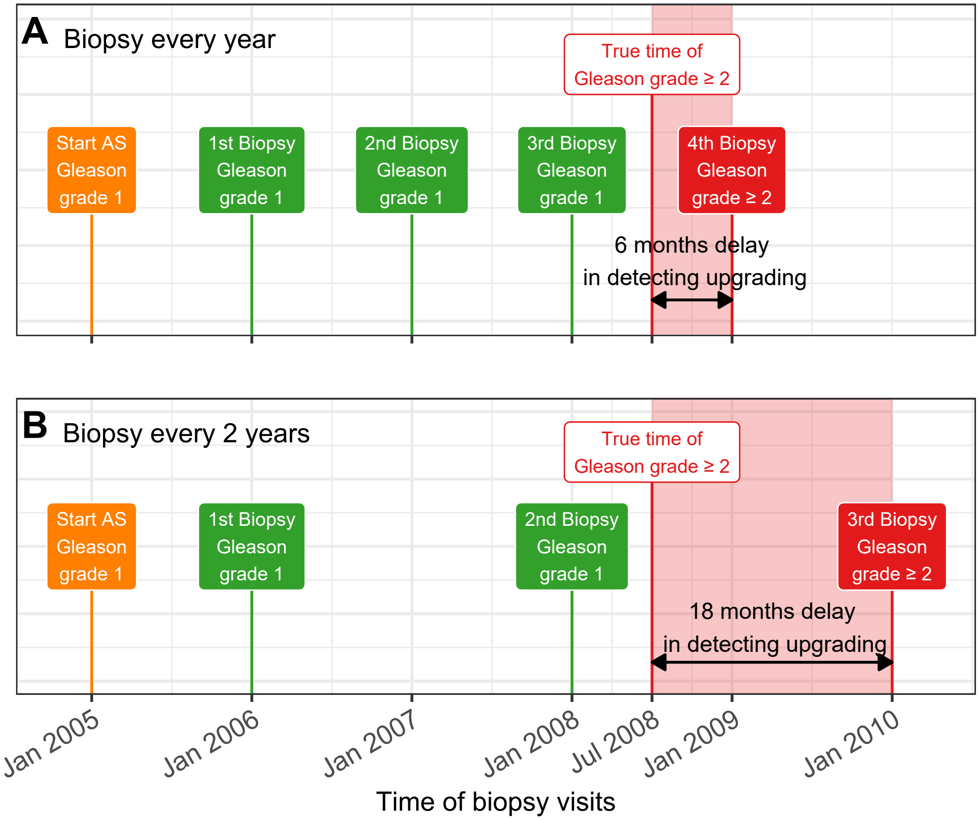
A Biopsy every year
True time of Gleason grade ≥ 2
Start AS Gleason grade 1
1st Biopsy Gleason grade 1
2nd Biopsy Gleason grade 1
3rd Biopsy Gleason grade 1
4th Biopsy Gleason grade ≥ 2
6 months delay in detecting upgrading
B Biopsy every 2 years
True time of Gleason grade ≥ 2
Start AS Gleason grade 1
1st Biopsy Gleason grade 1
2nd Biopsy Gleason grade 1
3rd Biopsy Gleason grade ≥ 2
18 months delay in detecting upgrading
Jan 2005
Jan 2006
Jan 2007
Jan 2008
Jul 2008
Jan 2009
Jan 2010
Time of biopsy visits

Figure 5.1: **Trade-off between the timing and number of biopsies (burden) and time delay in detecting Gleason upgrading (shorter is better):** The true time of Gleason upgrading (increase in Gleason grade group from group 1 to 2 or higher) for the patient in this figure is July 2008. When biopsies are scheduled annually (**Panel A**), upgrading is detected in January 2009 with a time delay of six months, and a total of four biopsies are scheduled. When biopsies are scheduled biennially (**Panel B**), upgrading is detected in January 2010 with a time delay of 18 months, and a total of three biopsies are scheduled. Since biopsies are conducted periodically, the time of upgrading is observed as an interval. For example, between Jan 2008–Jan 2009 in **Panel A** and between Jan 2008–Jan 2010 in **Panel B**. The phrase 'Gleason grade group' is shortened to 'Gleason grade' for brevity.

ally) (Inoue et al., 2018; de Carvalho et al., 2017a). Each of these approaches has limitations. For example, one-size-fits-all schedules can lead to many unnecessary biopsies because of differences in baseline *upgrading-risk* across cohorts (Inoue et al., 2018). Whereas, since observed clinical data has measurement error (e.g., PSA fluctuations), a flaw of using it directly is that it may lead to poor decisions. Also, decisions based on clinical data typically rely only on the latest data point and ignore previous repeated measurements. A novel alternative that counters these drawbacks is first processing patient data via a statistical model, and subsequently using model predicted upgrading-risks to create *personalized* biopsy schedules (Nieboer et al., 2018) (Figure 5.2). While, upgrading-risk calculators are not new (Coley et al., 2017; Ankerst et al., 2015; Partin et al., 1993; Makarov et al., 2007), not all are personalized either. Besides, they do not specify how risk predictions can be exploited to create a schedule.

This work is motivated by the problem of scheduling biopsies in AS. We have two goals. First, we want to assist practitioners in using clinical data in biopsy decisions in a statistically sound manner. To this end, we plan to develop a robust, generalizable statistical model that provides reliable individual upgrading-risk in AS. Subsequently, we will employ these predictions to derive risk-based personalized biopsy schedules. Our second goal is to enable shared decision making of biopsy schedules. We intend to achieve this by allowing patients and doctors to compare the *burden* and *benefit* (Figure 5.1) of opting for personalized schedules versus periodical schedules versus schedules based on clinical data. Specifically, we propose timing and number of planned biopsies (more/frequent are burdensome), and the expected time delay in detecting upgrading (shorter is beneficial) for any given schedule. While fulfilling our goals, we want to capture the maximum possible information from the available data. Hence, we will use all repeated measurements of patients, previous biopsy results, baseline characteristics, and keep our model flexible to accommodate future novel biomarkers. To fit this model, we will utilize data of the world's largest AS study, Prostate Cancer Research International Active Surveillance (PRIAS). To evaluate our model, we will externally validate it in the largest six AS cohorts from the

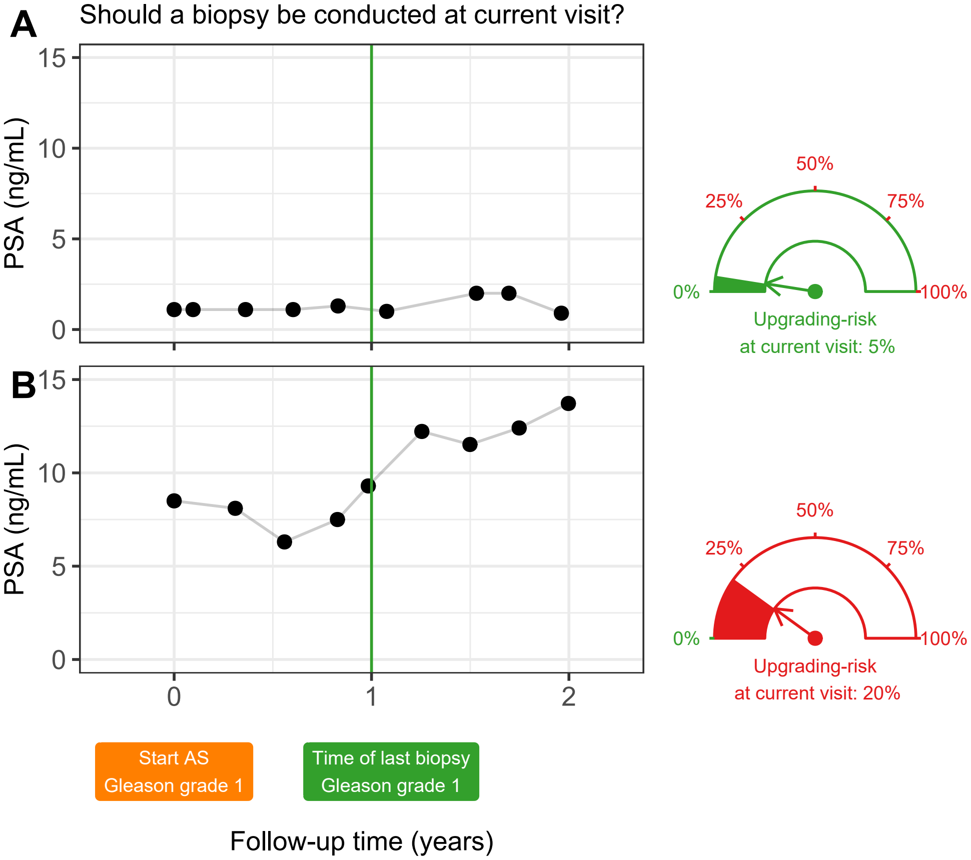

Figure 5.2: **Motivation for upgrading-risk based personalized biopsy decisions**: To utilize patients' complete longitudinal data and results from previous biopsies in making biopsy decisions. For this purpose, we first process data using a statistical model and then utilize the patient-specific predictions for risk of Gleason upgrading to schedule biopsies. For example, Patient A (**Panel A**) and B (**Panel B**) had their latest biopsy at year one of follow-up (green vertical line). Patient A's prostate-specific antigen (PSA) profile remained stable until his current visit at year two, whereas patient B's profile has shown a rise. Consequently, patient B's upgrading-risk at the current visit (year two) is higher than that of patient A. This makes patient B a more suitable candidate for biopsy than Patient A. Risk estimates in this figure are only illustrative.

Movember Foundation's Global Action Plan (GAP3) database (Bruinsma et al., 2018). Last, we aim to implement the validated model and methodology in a web-application.

5.2 Patients and Methods

5.2.1 Study Cohort

For developing a statistical model to predict upgrading-risk, we used the world's largest AS dataset, Prostate Cancer International Active Surveillance or PRIAS (Bul et al., 2013), dated April 2019 (Table 5.1). In PRIAS, biopsies were scheduled at year one, four, seven, ten, and additional yearly biopsies were scheduled when PSA doubling time was between zero and ten years. We selected all 7813 patients who had Gleason grade group 1 at inclusion in AS. Our primary event of interest is an increase in this Gleason grade group observed upon repeat biopsy, called *upgrading* (1134 patients). Upgrading is a trigger for treatment advice in PRIAS. Some examples of treatment options in active surveillance are radical prostatectomy, brachytherapy, definitive radiation therapy, and other alternative local treatments such as cryosurgery, High Intensity Focused Ultrasound, and External Beam Radiation Therapy. Comprehensive details on treatment options and their side effects are available in EAU-ESTRO-SIOG guidelines on prostate cancer (Mottet et al., 2017). In PRIAS 2250 patients were provided treatment based on their PSA, the number of biopsy cores with cancer, or anxiety/other reasons. However, our reasons for focusing solely on upgrading are that upgrading is strongly associated with cancer-related outcomes, and other treatment triggers vary between cohorts (Nieboer et al., 2018).

For externally validating our model's predictions, we selected the following largest (by the number of repeated measurements) six cohorts from Movember Foundation's GAP3 database (Bruinsma et al., 2018) version 3.1, covering nearly 73% of the GAP3 patients: the University of Toronto AS (Toronto), Johns Hopkins AS (Hopkins), Memorial Sloan Kettering Cancer

Table 5.1: **Summary of the PRIAS dataset as of April 2019**. The primary event of interest is upgrading, that is, increase in Gleason grade group from group 1 (Epstein et al., 2016) to 2 or higher. IQR: interquartile range, PSA: prostate-specific antigen. Study protocol URL: `https://www.prias-project.org`

Characteristic	**Value**
Total patients	7813
Upgrading (primary event)	1134
Treatment	2250
Watchful waiting	334
Loss to follow-up	249
Death (unrelated to prostate cancer)	95
Death (related to prostate cancer)	2
Median age at diagnosis (years)	66 (IQR: 61–71)
Median maximum follow-up per patient (years)	1.8 (IQR: 0.9–4.0)
Total PSA measurements	67578
Median number of PSA measurements per patient	6 (IQR: 4–12)
Median PSA value (ng/mL)	5.7 (IQR: 4.1–7.7)
Total biopsies	15686
Median number of biopsies per patient	2 (IQR: 1–2)

Center AS (MSKCC), King's College London AS (KCL), Michigan Urological Surgery Improvement Collaborative AS (MUSIC), and University of California San Francisco AS (UCSF, version 3.2). Only patients with a Gleason grade group 1 at the time of inclusion in these cohorts were selected. Summary statistics are presented in Section 5.B.

Choice of predictors: In our model, we used all repeated PSA measurements, the timing of the previous biopsy and Gleason grade, and age at inclusion in AS. Other predictors such as prostate volume, MRI results can also be important. MRI is utilized already for targeting biopsies, but regarding its use in deciding the time of biopsies, there are arguments both for and against it (Kasivisvanathan et al., 2020; Chesnut et al., 2020; Schoots et al., 2015). MRI is still a recent addition in most AS protocols. Consequently, repeated MRI data is very sparsely available in both PRIAS and GAP3 databases to make a stable prediction model. Prostate volume data is also sparsely available, especially in validation cohorts. Based on these reasons, we did not include them in our model. However, the model we propose next is extendable to include MRI and other novel biomarkers in the future.

5.2.2 Statistical Model

Modeling an AS dataset such as PRIAS, posed certain challenges. First, PSA was measured longitudinally, and over follow-up time it did not always increase linearly. Consequently, we expect that PSA measurements of a patient are more similar to each other than of another patient. In other words, we need to accommodate the within-patient correlation for PSA. Second, PSA was available only until a patient observed upgrading. Thus, we also need to model the association between the Gleason grades and PSA profiles of a patient, and handle missing PSA measurements after a patient experienced upgrading. Third, since the PRIAS biopsy schedule uses PSA, a patient's observed time of upgrading was also dependent on their PSA. Thus, the effect of PSA on the upgrading-risk need to be adjusted for the effect of PSA on the biopsy schedule. Fourth, many patients obtained treatment and watchful waiting before observing upgrading. Since we considered events other than upgrading as censoring, the model needs to account for patients' reasons for treatment or watchful waiting (e.g., age, treatment based on observed data). A model that handles these challenges in a statistically sound manner is the joint model for time-to-event and longitudinal data (Tomer et al., 2019a; Coley et al., 2017; Rizopoulos, 2012).

Our joint model consisted of two sub-models. Namely, a linear mixed-effects sub-model (Laird et al., 1982) for longitudinally measured PSA (log-transformed), and a relative-risk sub-model (similar to the Cox model) for the interval-censored time of upgrading. Patient age was used in both sub-models. Results and timing of the previous negative biopsies were used only in the risk sub-model. To account for PSA fluctuations (Nixon et al., 1997), we assumed t-distributed PSA measurement errors. The correlation between PSA measurements of the same patient was established using patient-specific random-effects. We fitted a unique curve to the PSA measurements of each patient (Panel A, Figure 5.3). Subsequently, we calculated the mathematical derivative of the patient's fitted PSA profile (5.2), to obtain his follow-up time specific instantaneous PSA velocity (Panel B, Figure 5.3). This instantaneous velocity is a stronger predictor of upgrading than the widely used average PSA velocity (Cooperberg et al., 2018). We modeled the impact of PSA on upgrading-risk by employing fitted PSA value and instantaneous velocity as predictors in the risk sub-model (Panel C, Figure 5.3). We adjusted the effect of PSA on upgrading-risk for the PSA dependent PRIAS biopsy schedule by estimating parameters using a full likelihood method (proof in Chapter 2.B). This approach also accommodates watchful waiting and treatment protocols that are also based on patient data. Specifically, the parameters (Section 5.A) of our two sub-models were estimated jointly under the Bayesian paradigm using the R package **JMbayes** (Rizopoulos, 2016).

5.2.3 Risk Prediction and Model Validation

Our model provides predictions for upgrading-risk over the entire future follow-up period of a patient (Panel C, Figure 5.3). However, we recommend using predictions only after year one. This is because most AS programs recommend a confirmatory biopsy at year one, especially to detect patients who may be misdiagnosed as low-grade at inclusion in AS. The model also automatically updates risk-predictions over follow-up as more patient data becomes available (Figure 4.2). We validated our model internally in the PRIAS cohort, and externally in the largest six GAP3 database cohorts. We employed calibration plots (Royston and Altman, 2013; Steyerberg et al., 2010) and follow-up *time-dependent* mean absolute risk prediction error or MAPE (Rizopoulos et al., 2017) to graphically and quantitatively evaluate

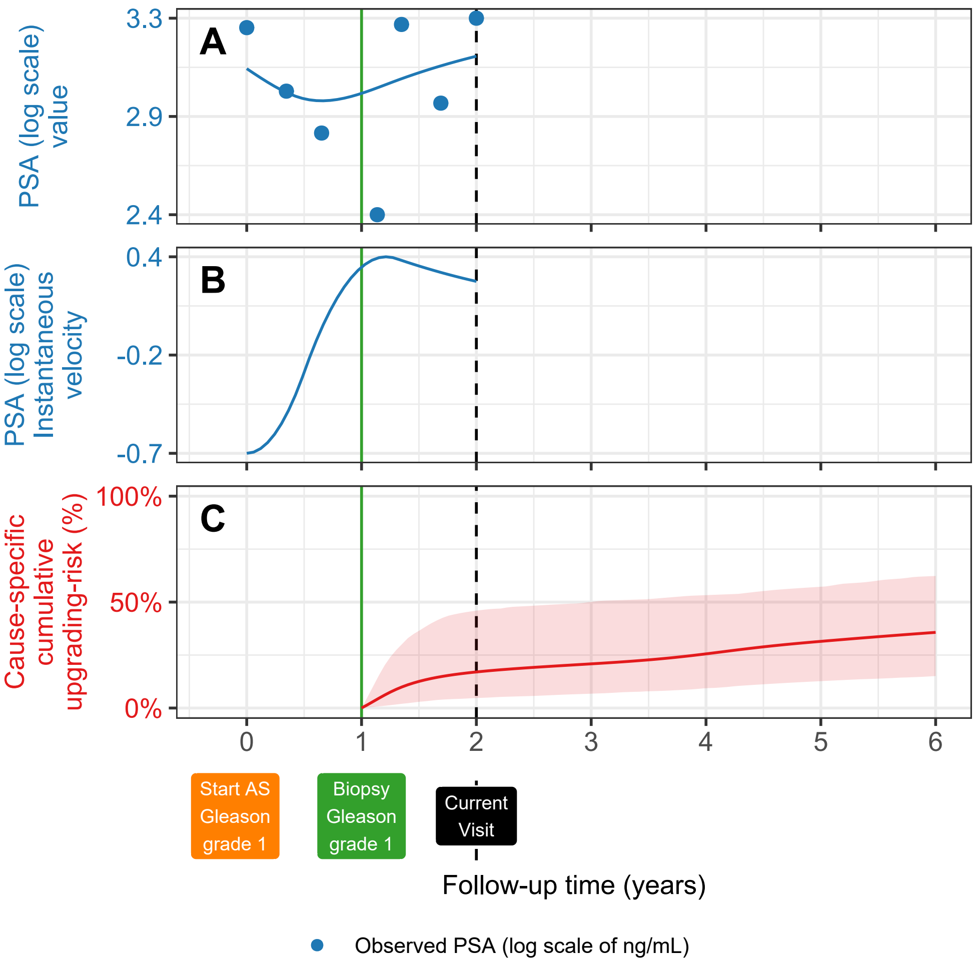

Figure 5.3: **Illustration of the joint model on a real PRIAS patient**. **Panel A:** Observed PSA (blue dots) and fitted PSA (solid blue line), log-transformed from ng/mL. **Panel B:** Estimated instantaneous velocity of PSA (log-transformed). **Panel C**: Predicted cause-specific cumulative upgrading-risk (95% credible interval shaded). Upgrading is defined as an increase in the Gleason grade group from group 1 (Epstein et al., 2016) to 2 or higher. This upgrading-risk is calculated starting from the time of the latest negative biopsy (vertical green line at year one of follow-up). The joint model estimated it by combining the fitted PSA (log scale) value and instantaneous velocity, and time of the latest negative biopsy. Black dashed line at year two denotes the time of current visit.

our model's risk prediction accuracy, respectively. We assessed our model's ability to discriminate between patients who experience/do not experience upgrading via the time-dependent area under the receiver operating characteristic curve or AUC (Rizopoulos et al., 2017).

The aforementioned *time-dependent* AUC and MAPE (Rizopoulos et al., 2017) are temporal extensions of their standard versions (Steyerberg et al., 2010) in a longitudinal setting. Specifically, at every six months of follow-up, we calculated a unique AUC and MAPE for predicting upgrading-risk in the subsequent one year (Appendix 5.C.1). For emulating a realistic situation, we calculated the AUC and MAPE at each follow-up using only the validation data available until that follow-up. Last, to resolve any potential model miscalibration in validation cohorts, we aimed to recalibrate our model's baseline hazard of upgrading (Appendix 5.C.1), individually for each cohort.

5.3 Results

The cause-specific cumulative upgrading-risk at year five of follow-up was 35% in PRIAS and at most 50% in validation cohorts (Panel B, Figure 5.4). In the fitted PRIAS model, the adjusted hazard ratio (aHR) of upgrading for an increase in patient age from 61 to 71 years (25-th to 75-th percentile) was 1.45 (95%CI: 1.30–1.63). For an increase in fitted PSA value from 2.36 to 3.07 (25-th to 75-th percentile, log scale), the aHR was 0.99 (95%CI: 0.89–1.11). The strongest predictor of upgrading-risk was instantaneous PSA velocity, with an increase from -0.09 to 0.31 (25-th to 75-th percentile), giving an aHR of 2.47 (95%CI: 1.93–2.99). The aHR for PSA value and velocity was different in each GAP3 cohort (Table 5.7).

The time-dependent AUC, calibration plot, and time-dependent MAPE of our model are shown in Figure 5.4, and Figure 5.9. In all cohorts, time-dependent AUC was moderate (0.6 to 0.7) over the whole follow-up period. Time-dependent MAPE was moderate (0.1 to 0.2) in those cohorts where the impact of PSA on upgrading-risk was similar to PRIAS (e.g., Hopkins cohort, Table 5.7), and large (0.2 to 0.3) otherwise. Our model was miscalibrated for validation cohorts (Panel B, Figure 5.4), because cohorts had differences in inclusion criteria (e.g., PSA density) and follow-up protocols (Bruinsma et al., 2018) which were not accounted in our model. Consequently, the PRIAS based model's fitted baseline hazard did not correspond to the baseline hazard in validation cohorts. To solve this problem, we recalibrated the baseline hazard of upgrading in validation cohorts (Figure 5.7). We compared risk predictions from the recalibrated models, with predictions from separately fitted cohort-specific joint models (Figure 5.8). The difference in predictions was lowest in the Johns Hopkins cohort (impact of PSA on upgrading-risk similar to PRIAS). Comprehensive results are in Appendix 5.B and Appendix 5.C.

5.3.1 Personalized Biopsy Schedules

We employed the PRIAS based fitted model to create personalized biopsy schedules for real PRIAS patients. Particularly, first using the model and patient's observed data, we predicted his cumulative upgrading-risk (Figure 5.5) on all of his future follow-up visits (biannually in PRIAS). Subsequently, we planned biopsies on those future visits where his conditional cumulative upgrading-risk was more than a certain threshold (see Chapter 4.3 for mathematical details). The choice of this threshold dictates the timing of biopsies in a risk-based personalized schedule. For example, personalized schedules based on 5% and 10% risk thresholds are shown in Figure 5.5.

To facilitate the choice of a risk-threshold, and for comparing the consequences of opting for a risk-based schedule versus any other schedule (e.g., annual, PRIAS), we predict expected time delay in detecting upgrading for following a schedule. We are able to predict this delay for any schedule. For example, in Panel C of Figure 5.5, the annual schedule has the least expected delay. In contrast, a personalized schedule based on a 10% risk threshold has a slightly larger expected delay, but it also schedules much fewer biopsies. An important aspect of this delay is that it is personalized as well. That is, even if two different patients are prescribed the same biopsy schedule, their expected delays will be different. This is because delay is estimated using

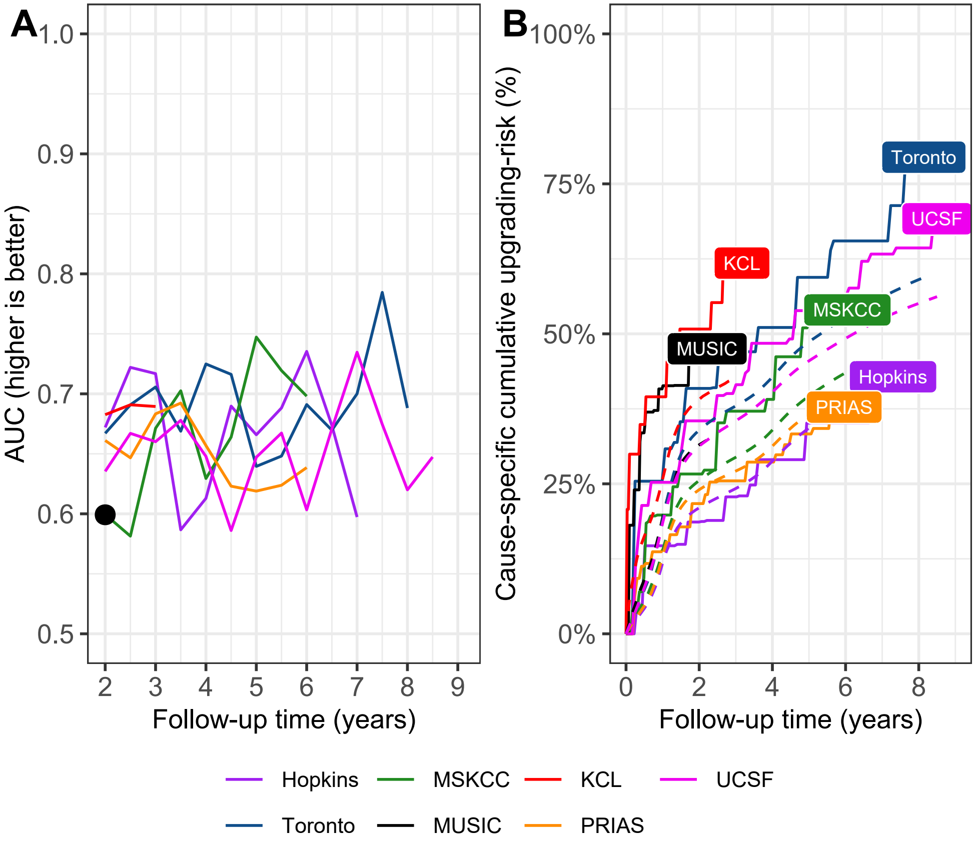

Figure 5.4: **Model Validation Results**. **Panel A**: time-dependent area under the receiver operating characteristic curve or AUC (measure of discrimination). AUC at year one is not shown because we do not intend to replace the confirmatory biopsy at year one. **Panel B**: calibration-at-large indicates model miscalibration. This is because solid lines depicting the non-parametric estimate of the cause-specific cumulative upgrading-risk (Turnbull, 1976), and dashed lines showing the average cause-specific cumulative upgrading-risk obtained using the joint model fitted to the PRIAS dataset, are not overlapping. Recalibrating the baseline hazard of upgrading resolved this issue (Figure 5.7). Full names of Cohorts are *PRIAS*: Prostate Cancer International Active Surveillance, *Toronto*: University of Toronto Active Surveillance, *Hopkins*: Johns Hopkins Active Surveillance, *MSKCC*: Memorial Sloan Kettering Cancer Center Active Surveillance, *KCL*: King's College London Active Surveillance, *MUSIC*: Michigan Urological Surgery Improvement Collaborative Active Surveillance, *UCSF*: University of California San Francisco AS.

all available clinical data of the patient (Chapter 4.3.3). While the timing and the total number of planned biopsies denote the burden of a schedule, a shorter expected time delay in detecting upgrading can be a benefit. These two, along with other measures such as a patient's comorbidities, anxiety, etc., can help to make an informed biopsy decision.

5.3.2 Web-Application

We implemented the PRIAS based model, recalibrated models for GAP3 cohorts, and personalized schedules in a user-friendly web-application `https://emcbiostatistics.shinyapps.io/prias_biopsy_recommender/`. This application works on both desktop and mobile devices. Patient data can be entered in Microsoft Excel format. The maximum follow-up time up to which predictions can be obtained depends on each cohort (Table 5.8). The web-application supports personalized, annual, and PRIAS schedules. For personalized schedules, users can control the choice of risk-threshold. The web-application also compares the resulting risk-based schedule's timing of biopsies, and expected time delay in detecting upgrading, with annual and PRIAS schedules, to enable sharing biopsy decision making.

5.4 Discussion

We successfully developed and externally validated a statistical model for predicting upgrading-risk (Bruinsma et al., 2017) in prostate cancer AS, and providing risk-based personalized biopsy decisions. Our work has four novel features over earlier risk calculators (Coley et al., 2017; Ankerst et al., 2015). First, our model was fitted to the world's largest AS dataset PRIAS and externally validated in the largest six cohorts of the Movember Foundation's GAP3 database (Bruinsma et al., 2018). Second, the model predicts a patient's current and future upgrading-risk in a personalized manner. Third, using the predicted risks, we created personalized biopsy schedules. We also calculated the expected time delay in detecting upgrading (less is beneficial) for fol-

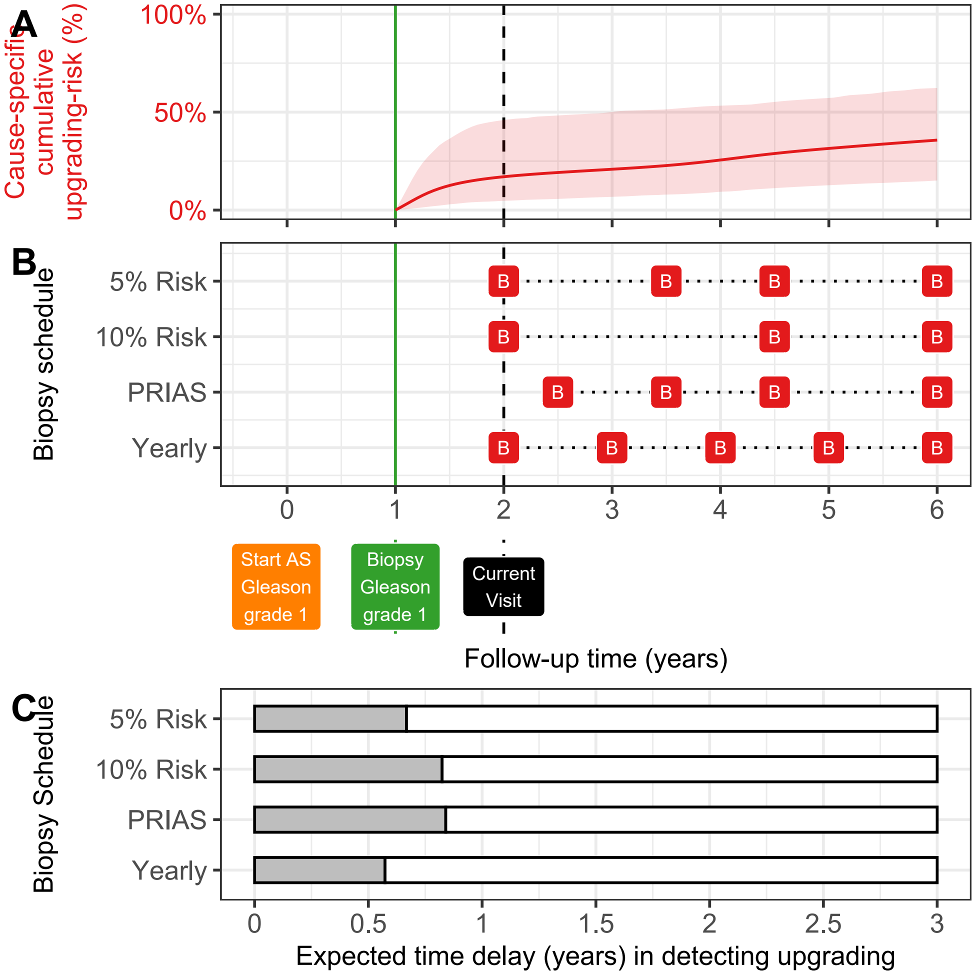

Figure 5.5: **Illustration of personalized and fixed schedules of biopsies for patient from Figure 5.3**. **Panel A:** Predicted cumulative upgrading-risk (95% credible interval shaded). **Panel B:** Different biopsy schedules with a red 'B' indicating a future biopsy. Risk: 5% and Risk: 10% are personalized schedules in which a biopsy is planned whenever the conditional cause-specific cumulative upgrading-risk is above 5% or 10% risk, respectively. Green vertical line at year one is the time of the latest negative biopsy. Black dashed line at year two denotes the time of the current visit. **Panel C:** Expected time delay in detecting upgrading (years) if patient progresses before year six. A compulsory biopsy was scheduled at year six (maximum biopsy scheduling time in PRIAS, Table 5.8) in all schedules for a meaningful comparison between them.

lowing any schedule. Thus, patients/doctors can compare schedules before making a choice. Fourth, we implemented our methodology in a user-friendly web-application (`https://emcbiostatistics.shinyapps.io/prias_biopsy_recommender/`) for both PRIAS and validated cohorts.

Our model and methods can be useful for numerous patients from PRIAS and the validated GAP3 cohorts (nearly 73% of all GAP3 patients). The model utilizes all repeated PSA measurements, results of previous biopsies, and baseline characteristics of a patient. We could not include MRI and PSA density because of sparsely available data in both PRIAS and GAP3 databases. But, our model is extendable to include them in the near future. The current discrimination ability of our model, exhibited by the *time-dependent* AUC, was between 0.6 and 0.7 over-follow. While this is moderate, it is also so because unlike the standard AUC (Steyerberg et al., 2010) the time-dependent AUC is more conservative as it utilizes only the validation data available until the time at which it is calculated. The same holds for the time-dependent MAPE (mean absolute prediction error). Although, MAPE varied much more between cohorts than AUC. In cohorts where the effect size for the impact of PSA value and velocity on upgrading-risk was similar to that for PRIAS (e.g., Hopkins cohort), MAPE was moderate. Otherwise, MAPE was large (e.g., KCL and MUSIC cohorts). We required recalibration of our model's baseline hazard of upgrading for all validation cohorts.

The clinical implications of our work are as follows. First, the cause-specific cumulative upgrading-risk at year five of follow-up was at most 50% in all cohorts (Panel B, Figure 5.4). That is, many patients may not require some of the biopsies planned in the first five years of AS. Given the non-compliance and burden of frequent biopsies (Bokhorst et al., 2015), the availability of our methodology as a web-application may encourage patients/doctors to consider upgrading-risk based personalized schedules instead. An additional advantage of personalized schedules is that they update as more patient data becomes available over follow-up. Despite the moderate predictive performance, we expect the overall impact of our model to be positive. There are two reasons for this. First, the risk of adverse outcomes because of the use of personalized schedules is quite low because of the low rate of metastases and prostate cancer specific mortality in AS patients (Table 5.1). Second, studies (de Carvalho et al., 2017b; Inoue et al., 2018) have suggested that after the confirmatory biopsy at year one of follow-up, biopsies may be done as infrequently as every two to three years, with limited adverse consequences. In other words, longer delays in detecting upgrading may be acceptable after the first negative biopsy. To evaluate the potential harm of personalized schedules, we compared them with fixed schedules in a realistic and extensive simulation study (Tomer et al., 2019b). We concluded that personalized schedules plan, on average, six fewer biopsies compared to annual schedule and two fewer biopsies than the PRIAS schedule in slow/non-progressing AS patients, while maintaining almost the same time delay in detecting upgrading as PRIAS schedule. Personalized schedules with different risk thresholds indeed have different performances across cohorts. Thus, to assist patients/doctors in choosing between fixed schedules and personalized schedules based on different risk thresholds, the web-application provides a patient-specific estimate of the expected time delay in detecting upgrading, for both personalized and fixed schedules. We hope that access to these estimates will objectively address patient apprehensions regarding adverse outcomes in AS. Last, we note that our web-application should only be used to decide biopsies after the compulsory confirmatory biopsy at year one of follow-up.

This work has certain limitations. Predictions for upgrading-risk and personalized schedules are available only for a currently limited, cohort-specific, follow-up period (Table 5.8). This problem can be mitigated by refitting the model with new follow-up data in the future. Recently, some cohorts started utilizing MRI to explore the possibility of targeting visible lesions by biopsy. Presently, the GAP3 database has limited PSA density and MRI follow-up data available. Since PSA density is used as an entry criterion in some active surveillance studies, including it as a predictor can improve the model. Although, the current model can be extended to include both MRI and PSA density data as predictors when they become available in future. We scheduled biopsies using cause-specific cumulative upgrading-risk, which ignores competing events such as treatment based on the number of positive biopsy cores. Employing a competing-risk model may lead to improved personalized schedules. Upgrading is susceptible to inter-observer variation too. Models which account for this variation (Coley et al., 2017; Balasubramanian and Lagakos, 2003) will be interesting to investigate further. Even with an enhanced risk prediction model, the methodology for personalized scheduling and calculation of expected time delay (Chapter 4.3.3) need not change. Last, our web-application only allows uploading patient data in Microsoft Excel format. Connecting it with patient databases can increase usability.

5.5 Conclusions

We successfully developed a statistical model and methodology for predicting upgrading-risk, and providing risk-based personalized biopsy decisions, in prostate cancer AS. We externally validated our model, covering nearly 73% patients from the Movember Foundations' GAP3 database. The model made available via a user-friendly web-application (`https://emcbiostatistics.shinyapps.io/prias_biopsy_recommender/`) enables shared decision making of biopsy schedules by comparing fixed and personalized schedules on total biopsies and expected time delay in detecting upgrading. Novel biomarkers and MRI data can be added as predictors in the model to improve predictions in the future. Recalibration of baseline upgrading-risk is advised for cohorts not validated in this work.

Acknowledgements We thank Jozien Helleman from the Department of Urology, Erasmus University Medical Center, for coordinating the project. The first and last authors would like to acknowledge support by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (the national research council of the Netherlands) VIDI grant nr. 016.146.301, and Erasmus University Medical Center funding. Part of this work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. The authors also thank the Erasmus University Medical Center's Cancer Computational Biology Center for giving access to their IT-infrastructure and software that was used for the computations and data analysis in this study. The PRIAS website is funded by the Prostate Cancer Research Foundation, Rotterdam (SWOP). We would like to thank the PRIAS consortium for enabling this research project.

This work was supported by the Movember Foundation. The funder did not play any role in the study design, collection, analysis or interpretation of data, or in the drafting of this paper.

Conflicts of Interest The authors do not report any conflict of interest, and have nothing to disclose.

Members of The Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium

Principle Investigators: Bruce Trock (Johns Hopkins University, The James Buchanan Brady Urological Institute, Baltimore, USA), Behfar Ehdaie (Memorial Sloan Kettering Cancer Center, New York, USA), Peter Carroll (University of California San Francisco, San Francisco, USA), Christopher Filson (Emory University School of Medicine, Winship Cancer Institute, Atlanta, USA), Jeri Kim / Christopher Logothetis (MD Anderson Cancer Centre, Houston, USA), Todd Morgan (University of Michigan and Michigan Urological Surgery Improvement Collaborative (MUSIC), Michigan, USA), Laurence Klotz (University of Toronto, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada), Tom Pickles (University of British Columbia, BC Cancer Agency, Vancouver, Canada), Eric Hyndman (University of Calgary, Southern Alberta Institute of Urology, Calgary, Canada), Caroline Moore (University College London & University College London Hospital Trust, London, UK), Vincent Gnanapragasam (University of Cambridge & Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK), Mieke Van Hemelrijck (King's College London, London, UK & Guy's and St Thomas' NHS

Foundation Trust, London, UK), Prokar Dasgupta (Guy's and St Thomas' NHS Foundation Trust, London, UK), Chris Bangma (Erasmus Medical Center, Rotterdam, The Netherlands/ representative of Prostate cancer Research International Active Surveillance (PRIAS) consortium), Monique Roobol (Erasmus Medical Center, Rotterdam, The Netherlands/ representative of Prostate cancer Research International Active Surveillance (PRIAS) consortium), The PRIAS study group, Arnauld Villers (Lille University Hospital Center, Lille, France), Antti Rannikko (Helsinki University and Helsinki University Hospital, Helsinki, Finland), Riccardo Valdagni (Department of Oncology and Hemato-oncology, Università degli Studi di Milano, Radiation Oncology 1 and Prostate Cancer Program, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy), Antoinette Perry (University College Dublin, Dublin, Ireland), Jonas Hugosson (Sahlgrenska University Hospital, Göteborg, Sweden), Jose Rubio-Briones (Instituto Valenciano de Oncología, Valencia, Spain), Anders Bjartell (Skåne University Hospital, Malmö, Sweden), Lukas Hefermehl (Kantonsspital Baden, Baden, Switzerland), Lee Lui Shiong (Singapore General Hospital, Singapore, Singapore), Mark Frydenberg (Monash Health; Monash University, Melbourne, Australia), Yoshiyuki Kakehi / Mikio Sugimoto (Kagawa University Faculty of Medicine, Kagawa, Japan), Byung Ha Chung (Gangnam Severance Hospital, Yonsei University Health System, Seoul, Republic of Korea)

Pathologist: Theo van der Kwast (Princess Margaret Cancer Centre, Toronto, Canada). Technology Research Partners: Henk Obbink (Royal Philips, Eindhoven, the Netherlands), Wim van der Linden (Royal Philips, Eindhoven, the Netherlands), Tim Hulsen (Royal Philips, Eindhoven, the Netherlands), Cees de Jonge (Royal Philips, Eindhoven, the Netherlands).

Advisory Regional statisticians: Mike Kattan (Cleveland Clinic, Cleveland, Ohio, USA), Ji Xinge (Cleveland Clinic, Cleveland, Ohio, USA), Kenneth Muir (University of Manchester, Manchester, UK), Artitaya Lophatananon (University of Manchester, Manchester, UK), Michael Fahey (Epworth HealthCare, Melbourne, Australia), Ewout Steyerberg (Erasmus Medical Center, Rotterdam, The Netherlands), Daan Nieboer (Erasmus Medical Center, Rotterdam, The Netherlands); Liying Zhang (University of Toronto, Sunnybrook

Health Sciences Centre, Toronto, Ontario, Canada)

Executive Regional statisticians: Ewout Steyerberg (Erasmus Medical Center, Rotterdam, The Netherlands), Daan Nieboer (Erasmus Medical Center, Rotterdam, The Netherlands); Kerri Beckmann (King's College London, London, UK & Guy's and St Thomas' NHS Foundation Trust, London, UK), Brian Denton (University of Michigan, Michigan, USA), Andrew Hayen (University of Technology Sydney, Australia), Paul Boutros (Ontario Institute of Cancer Research, Toronto, Ontario, Canada).

Clinical Research Partners' IT Experts: Wei Guo (Johns Hopkins University, The James Buchanan Brady Urological Institute, Baltimore, USA), Nicole Benfante (Memorial Sloan Kettering Cancer Center, New York, USA), Janet Cowan (University of California San Francisco, San Francisco, USA), Dattatraya Patil (Emory University School of Medicine, Winship Cancer Institute, Atlanta, USA), Emily Tolosa (MD Anderson Cancer Centre, Houston, Texas, USA), Tae-Kyung Kim (University of Michigan and Michigan Urological Surgery Improvement Collaborative, Ann Arbor, Michigan, USA), Alexandre Mamedov (University of Toronto, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada), Vincent LaPointe (University of British Columbia, BC Cancer Agency, Vancouver, Canada), Trafford Crump (University of Calgary, Southern Alberta Institute of Urology, Calgary, Canada), Vasilis Stavrinides (University College London & University College London Hospital Trust, London, UK), Jenna Kimberly-Duffell (University of Cambridge & Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK), Aida Santaolalla (King's College London, London, UK & Guy's and St Thomas' NHS Foundation Trust, London, UK), Daan Nieboer (Erasmus Medical Center, Rotterdam, The Netherlands), Jonathan Olivier (Lille University Hospital Center, Lille, France), Tiziana Rancati (Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy), Helén Ahlgren (Sahlgrenska University Hospital, Göteborg, Sweden), Juanma Mascarós (Instituto Valenciano de Oncología, Valencia, Spain), Annica Löfgren (Skåne University Hospital, Malmö, Sweden), Kurt Lehmann (Kantonsspital Baden, Baden, Switzerland), Catherine Han Lin (Monash University and Epworth HealthCare, Melbourne, Australia), Hiromi Hirama (Kagawa University, Kagawa, Japan), Kwang Suk Lee (Yonsei University College of Medicine, Gangnam Severance Hospital, Seoul, Korea).

Research Advisory Committee: Guido Jenster (Erasmus MC, Rotterdam, the Netherlands), Anssi Auvinen (University of Tampere, Tampere, Finland), Anders Bjartell (Skåne University Hospital, Malmö, Sweden), Masoom Haider (University of Toronto, Toronto, Canada), Kees van Bochove (The Hyve B.V. Utrecht, Utrecht, the Netherlands), Ballentine Carter (Johns Hopkins University, Baltimore, USA – until 2018).

Management team: Sam Gledhill (Movember Foundation, Melbourne, Australia), Mark Buzza / Michelle Kouspou (Movember Foundation, Melbourne, Australia), Chris Bangma (Erasmus Medical Center, Rotterdam, The Netherlands), Monique Roobol (Erasmus Medical Center, Rotterdam, The Netherlands), Sophie Bruinsma / Jozien Helleman (Erasmus Medical Center, Rotterdam, The Netherlands).

Appendix

5.A Model Specification

Let T_i^* denote the true time of upgrading (increase in biopsy Gleason grade group from 1 to 2 or higher) for the i-th patient included in PRIAS. Since biopsies are conducted periodically, T_i^* is observed with interval censoring $l_i < T_i^* \leq r_i$. When upgrading is observed for the patient at his latest biopsy time r_i, then l_i denotes the time of the second latest biopsy. Otherwise, l_i denotes the time of the latest biopsy and $r_i = \infty$. Let $\boldsymbol{y}_i$ denote his observed PSA longitudinal measurements. The observed data of all n patients is denoted by $\mathcal{A}_n = \{l_i, r_i, \boldsymbol{y}_i; i = 1, \ldots, n\}$.

In our joint model, the patient-specific PSA measurements over time are modeled using a linear mixed effects sub-model. It is given by (see Panel A,

Figure 5.3):

$$\begin{aligned}\log_2\Big\{y_i(t)+1\Big\} &= m_i(t)+\varepsilon_i(t),\\ m_i(t) &= \beta_0+b_{0i}+\sum_{k=1}^{4}(\beta_k+b_{ki})B_k\Big(\frac{t-2}{2},\frac{\mathcal{K}-2}{2}\Big)\\ &\quad+\beta_5\textsf{age}_i,\end{aligned}\tag{5.1}$$

where, $m_i(t)$ denotes the measurement error free value of $\log_2(\textsf{PSA}+1)$ transformed (Pearson et al., 1994; Lin et al., 2000) measurements at time t. We model it non-linearly over time using B-splines (De Boor, 1978). To this end, our B-spline basis function $B_k\{(t-2)/2,(\mathcal{K}-2)/2\}$ has three internal knots at $\mathcal{K}=\{0.5,1.3,3\}$ years, which are the three quartiles of the observed follow-up times. The boundary knots of the spline are at 0 and 6.3 years (95-th percentile of the observed follow-up times). We mean centered (mean 2 years) and standardized (standard deviation 2 years) the follow-up time t and the knots of the B-spline $\mathcal{K}$ during parameter estimation for better convergence. The fixed effect parameters are denoted by $\{\beta_0,\ldots,\beta_5\}$, and $\{b_{0i},\ldots,b_{4i}\}$ are the patient specific random effects. The random effects follow a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{W}$. The error $\varepsilon_i(t)$ is assumed to be t-distributed with three degrees of freedom and scale σ, and is independent of the random effects.

To model the impact of PSA measurements on the risk of upgrading, our joint model uses a relative risk sub-model. More specifically, the hazard of upgrading denoted as $h_i(t)$, and the cumulative-risk of upgrading denoted as $R_i(t)$, at a time t are (see Panel C, Figure 5.3):

$$\begin{aligned}h_i(t) &= h_0(t)\exp\Big(\gamma\textsf{age}_i+\alpha_1 m_i(t)+\alpha_2\frac{\mathrm{d}m_i(t)}{\mathrm{d}t}\Big),\\ R_i(t) &= \exp\Big\{-\int_0^t h_i(s)\mathrm{d}s\Big\},\end{aligned}\tag{5.2}$$

where, γ is the parameter for the effect of age. The impact of PSA on the hazard of upgrading is modeled in two ways, namely the impact of the error free underlying PSA value $m_i(t)$ (see Panel A, Figure 5.3), and the impact of the underlying PSA velocity $\mathrm{d}m_i(t)/\mathrm{d}t$ (see Panel B, Figure 5.3). The corresponding parameters are α_1 and α_2, respectively. Lastly, $h_0(t)$ is the baseline hazard at time t, and is modeled flexibly using P-splines (Eilers and Marx, 1996). More specifically:

$$\log h_0(t) = \gamma_{h_0,0} + \sum_{q=1}^{Q} \gamma_{h_0,q} B_q(t, \boldsymbol{v}),$$

where $B_q(t, \boldsymbol{v})$ denotes the q-th basis function of a B-spline with knots $\boldsymbol{v} = v_1, \ldots, v_Q$ and vector of spline coefficients γ_{h_0}. To avoid choosing the number and position of knots in the spline, a relatively high number of knots (e.g., 15 to 20) are chosen and the corresponding B-spline regression coefficients γ_{h_0} are penalized using a differences penalty (Eilers and Marx, 1996).

5.B Full Results

Characteristics of the six validation cohorts from the GAP3 database (Bruinsma et al., 2018) are shown in Table 5.2, Table 5.3, and Table 5.4. The cause-specific cumulative upgrading-risk in these cohorts is shown in Figure 5.6.

For the relative risk sub-model, the parameter estimates in Table 5.5 show that $\log_2(\text{PSA}+1)$ velocity and age of the patient were significantly associated with the hazard of upgrading.

It is important to note that since age, and $\log_2(\text{PSA}+1)$ value and velocity are all measured on different scales, a comparison between the corresponding parameter estimates is not easy. To this end, in Table 5.6, we present the hazard ratio of upgrading, for an increase in the aforementioned variables from their 25-th to the 75-th percentile. For example, an increase in fitted $\log_2(\text{PSA}+1)$ velocity from -0.085 to 0.308 (fitted 25-th and 75-th percentiles) corresponds to a hazard ratio of 2.433. The interpretation of the rest is similar.

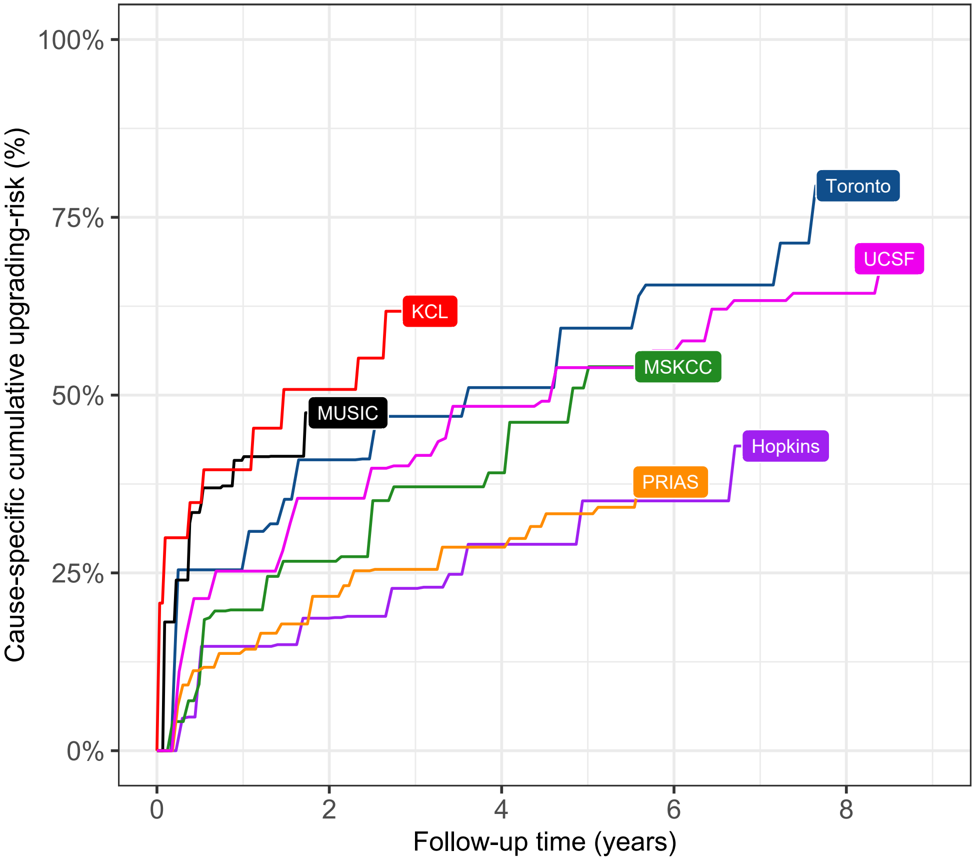

Figure 5.6: **Nonparametric estimate (Turnbull, 1976) of the cause-specific cumulative upgrading-risk** in the world's largest AS cohort PRIAS, and largest six AS cohorts from the GAP3 database (Bruinsma et al., 2018). Abbreviations are *Hopkins*: Johns Hopkins Active Surveillance, *PRIAS*: Prostate Cancer International Active Surveillance, *Toronto*: University of Toronto Active Surveillance, *MSKCC*: Memorial Sloan Kettering Cancer Center Active Surveillance, *KCL*: King's College London Active Surveillance, *MUSIC*: Michigan Urological Surgery Improvement Collaborative AS, *UCSF*: University of California San Francisco Active Surveillance.

Table 5.2: **Summary of the Hopkins and Toronto validation cohorts from the GAP3 database (Bruinsma et al., 2018)**. The primary event of interest is upgrading, that is, increase in Gleason grade group from group 1 to 2 or higher. #PSA: number of PSA, #biopsies: number of biopsies, IQR: interquartile range, PSA: prostate-specific antigen. Full names of cohorts are *Hopkins*: Johns Hopkins Active Surveillance, *Toronto*: University of Toronto Active Surveillance

Characteristic	**Hopkins**	**Toronto**
Total patients	1392	1046
Upgrading (primary event)	260	359
Median age (years)	62 (IQR: 66–69)	67 (IQR: 60–72)
Median maximum follow-up per patient (years)	3 (IQR: 1.3–5.8)	4.5 (IQR: 1.9–8.4)
Total PSA measurements	11126	13984
Median #PSA per patient	6 (IQR: 4–11)	12 (IQR: 7–19)
Median PSA (ng/mL)	4.7 (IQR: 2.9–6.7)	6 (IQR: 3.7–9.0)
Total biopsies	1926	909
Median #biopsies per patient	1 (IQR: 1–2)	1 (IQR: 1–2)

5.C Risk Predictions for Upgrading

Let us assume a new patient j, for whom we need to estimate the upgrading-risk. Let his current follow-up visit time be v, latest time of biopsy be t, observed vector PSA measurements be $\mathcal{Y}_j(v)$. The combined information from the observed data about the time of upgrading, is given by the following posterior predictive distribution $g(T_j^*)$ of his time T_j^* of upgrading:

$$\begin{aligned} g(T_j^*) &= p\Big\{T_j^* \mid T_j^* > t, \mathcal{Y}_j(v), \mathcal{A}_n\Big\} \\ &= \int\int p\Big(T_j^* \mid T_j^* > t, \boldsymbol{b}_j, \boldsymbol{\theta}\Big) p\Big\{\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(v), \boldsymbol{\theta}\Big\} p\Big(\boldsymbol{\theta} \mid \mathcal{A}_n\Big) \mathrm{d}\boldsymbol{b}_j \mathrm{d}\boldsymbol{\theta}. \end{aligned}$$

The distribution $g(T_j^*)$ depends not only depends on the observed data of the patient $T_j^* > t, \mathcal{Y}_j(v)$, but also depends on the information from the PRIAS dataset $\mathcal{A}_n$. To this the the posterior distribution of random effects $\boldsymbol{b}_j$ and

Table 5.3: **Summary of the MSKCC and UCSF validation cohorts from the GAP3 database (Bruinsma et al., 2018)**. The primary event of interest is upgrading, that is, increase in Gleason grade group from group 1 to 2 or higher. #PSA: number of PSA, #biopsies: number of biopsies, IQR: interquartile range, PSA: prostate-specific antigen. Full names of cohorts are *MSKCC*: Memorial Sloan Kettering Cancer Center Active Surveillance, *UCSF*: University of California San Francisco Active Surveillance.

Characteristic	**MSKCC**	**UCSF**
Total patients	894	1397
Upgrading (primary event)	242	547
Median age (years)	63 (IQR: 57–68)	63 (IQR: 57–68)
Median maximum follow-up per patient (years)	5.3 (IQR: 1.8–8.3)	3.6 (IQR: 1.5–7.2)
Total PSA measurements	10704	16093
Median #PSA per patient	11 (IQR: 5–17)	8 (IQR: 4–16)
Median PSA (ng/mL)	4.7 (IQR: 2.8–7.1)	5.0 (IQR: 3.4–7.2)
Total biopsies	1102	3512
Median #biopsies per patient	1 (IQR: 1–2)	2 (IQR: 2–3)

posterior distribution of the vector of all parameters $\boldsymbol{\theta}$ are utilized, respectively. The distribution $g(T_j^*)$ can be estimated as detailed in Rizopoulos et al. (2017). Since, many prostate cancer patients may not obtain upgrading in the current follow-up period of PRIAS, $g(T_j^*)$ can only be estimated for a currently limited follow-up period.

The cause-specific cumulative upgrading-risk can be derived from $g(T_j^*)$ as given in (Rizopoulos et al., 2017). It is given by:

$$R_j(u \mid t, v) = \text{Pr}\Big\{T_j^* > u \mid T_j^* > t, \mathcal{Y}_j(v), \mathcal{A}_n\Big\}, \quad u \geq t. \tag{5.3}$$

The personalized risk profile of the patient updates as more data is gathered over follow-up visits.

Table 5.4: **Summary of the MUSIC and KCL validation cohorts from the GAP3 database (Bruinsma et al., 2018)**. The primary event of interest is upgrading, that is, increase in Gleason grade group from group 1 to 2 or higher. #PSA: number of PSA, #biopsies: number of biopsies, IQR: interquartile range, PSA: prostate-specific antigen. Full names of cohorts are *KCL*: King's College London Active Surveillance, *MUSIC*: Michigan Urological Surgery Improvement Collaborative AS.

Characteristic	**MUSIC**	**KCL**
Total patients	2743	616
Upgrading (primary event)	385	198
Median age (years)	65 (IQR: 60–71)	63 (IQR: 58–68)
Median maximum follow-up per patient (years)	1.2 (IQR: 0.6–2.2)	2.4 (IQR: 1.3–3.8)
Total PSA measurements	12087	2987
Median #PSA per patient	4 (IQR: 2–6)	4 (IQR: 2–6)
Median PSA (ng/mL)	5.1 (IQR: 3.4–7.1)	6 (IQR: 4–9)
Total biopsies	1032	484
Median #biopsies per patient	1 (IQR: 1–1)	1 (IQR: 1–1)

Table 5.5: **Parameters of the relative risk sub-model**: Estimated mean and 95% credible interval for the parameters of the relative-risk sub-model.

Variable	Mean	Std. Dev	2.5%	97.5%	P
Age	0.037	0.006	0.025	0.049	<0.001
Fitted $\log_2(\text{PSA} + 1)$ value	-0.012	0.076	-0.164	0.135	0.856
Fitted $\log_2(\text{PSA} + 1)$ velocity	2.266	0.299	1.613	2.767	<0.001

Table 5.6: **Hazard ratio and 95% credible interval (CI) for upgrading**: Variables are on different scale and hence we compare an increase in the variables of relative risk sub-model from their 25-th percentile (P_{25}) to their 75-th percentile (P_{75}). Except for age, quartiles for all other variables are based on their fitted values obtained from the joint model fitted to the PRIAS dataset.

Variable	P_{25}	P_{75}	Hazard ratio [95% CI]
Age	61	71	1.455 [1.285, 1.631]
Fitted $\log_2(\text{PSA}+1)$ value	2.360	3.078	0.991 [0.889, 1.102]
Fitted $\log_2(\text{PSA}+1)$ velocity	-0.085	0.308	2.433 [1.883, 2.962]

Table 5.7: **Parameters of the relative risk sub-model in validation cohorts**. We fitted separate joint models for each of the six GAP3 validation cohorts as well. The specification of these joint models was same as that of the model for PRIAS. Two important predictors in the relative-risk sub-model, namely, the $\log_2(\text{PSA}+1)$ value and velocity have different impact on upgrading-risk across the cohorts. Table shows the mean estimate of these parameters with 95% credible interval in brackets. Strongest average effect of $\log_2(\text{PSA}+1)$ velocity is in PRIAS cohort, whereas the weakest is in MUSIC cohort. The strongest average effect of $\log_2(\text{PSA}+1)$ value is in the Toronto cohort whereas the weakest is in PRIAS cohort. Full names of cohorts are *Hopkins*: Johns Hopkins Active Surveillance, *PRIAS*: Prostate Cancer International Active Surveillance, *Toronto*: University of Toronto Active Surveillance, *MSKCC*: Memorial Sloan Kettering Cancer Center Active Surveillance, *KCL*: King's College London Active Surveillance, *MUSIC*: Michigan Urological Surgery Improvement Collaborative AS, *UCSF*: University of California San Francisco Active Surveillance.

Cohort	Fitted $\log_2(\text{PSA}+1)$ value	Fitted $\log_2(\text{PSA}+1)$ velocity
PRIAS	-0.012 [-0.164, 0.135]	2.266 [1.613, 2.767]
Hopkins	0.061 [-0.323, 0.329]	1.839 [0.761, 4.378]
MSKCC	0.336 [0.081, 0.583]	1.122 [0.421, 1.980]
Toronto	0.572 [0.347, 0.794]	0.943 [0.464, 1.554]
UCSF	0.498 [0.326, 0.673]	0.812 [0.280, 1.383]
MUSIC	0.441 [0.092, 0.767]	0.029 [-0.552, 0.512]
KCL	0.194 [-0.104, 0.540]	0.840 [-0.087, 1.665]

5.C.1 Validation of Risk Predictions

We wanted to check the usefulness of our model for not only the PRIAS patients but also for patients from other cohorts. To this end, we validated our model in the PRIAS dataset (internal validation) and the largest six cohorts from the GAP3 database (Bruinsma et al., 2018). These are the University of Toronto AS (Toronto), Johns Hopkins AS (Hopkins), Memorial Sloan Kettering Cancer Center AS (MSKCC), University of California San Francisco Active Surveillance (UCSF), King's College London AS (KCL), Michigan Urological Surgery Improvement Collaborative AS (MUSIC).

Calibration-in-the-large We first assessed calibration-in-the-large (Steyerberg et al., 2010) of our model in the aforementioned cohorts. To this end, we used our model to predict the cause-specific cumulative upgrading-risk for each patient, given their PSA measurements and biopsy results. We then averaged the resulting profiles of cause-specific cumulative upgrading-risk. Subsequently, we compared the averaged cumulative-risk profile with a non-parametric estimate (Turnbull, 1976) of the cause-specific cumulative upgrading-risk in each of the cohorts. The results are shown in Panel A of Figure 5.7. We can see that our model is miscalibrated in external cohorts, although it is fine in the Hopkins cohort. To improve our model's calibration in all cohorts, we recalibrated the baseline hazard of the joint model fitted to the PRIAS dataset, individually for each of the cohorts except the Hopkins cohort. More specifically, given the data of an external cohort $\mathcal{A}^c$, where c denotes the cohort, the recalibrated parameters $\gamma^c_{h_0}$ (Section 5.A) of the log baseline hazard are given by:

$$p(\boldsymbol{\gamma}^c_{h_0} \mid \mathcal{A}^c, \boldsymbol{b^c}, \boldsymbol{\theta}) \propto \prod_{i=1}^{n^c} p(l^c_i, r^c_i \mid \boldsymbol{b^c_i}, \boldsymbol{\theta}) p(\boldsymbol{\gamma}^c_{h_0}) \tag{5.4}$$

where n^c are the number of patients in the c-th cohort, and $\boldsymbol{\theta}$ is the vector of all parameters of the joint model fitted to the PRIAS dataset. The interval in which upgrading is observed for the i-th patient is given by l^c_i, r^c_i, with $r^c_i = \infty$ for right-censored patients. The symbol $\boldsymbol{b^c_i}$ denotes patient-specific random effects (Section 5.A) in the c-th cohort. The random effects are obtained using the joint model fitted to the PRIAS dataset before recalibration. We re-evaluated the calibration-in-the-large of our model after the recalibration of the baseline hazard individually for each cohort. The improved calibration-in-the-large is shown in Panel B of Figure 5.7.

Recalibrated PRIAS Model Versus Individual Joint Models For Each Cohort We wanted to check if our recalibrated PRIAS model performed as good as a new joint model that could be fitted to the external cohorts. To this end, we predicted cause-specific cumulative upgrading-risk for each patient from each cohort using two sets of models, namely the recalibrated PRIAS model for each cohort, and a new joint model fitted to each cohort. The difference in predicted cause-specific cumulative upgrading-risk from these models is shown in Figure 5.8. We can see that the difference is smaller in those cohorts in which the effects of $\log_2(\text{PSA} + 1)$ value and velocity were similar to that of PRIAS (Table 5.7). For example, the Hopkins cohort had parameter estimates similar to that of PRIAS, and consequently, the difference in predicted risks for this cohort is smallest. The opposite of this phenomenon holds for the MUSIC and KCL cohorts.

Validation of Dynamic Cumulative-Risk Predictions The cumulative-risk predictions from the joint model are dynamic in nature. That is, they update as more data becomes available over time. Consequently, the discrimination and prediction error of the joint model also depend on the available data. We assessed these two measures dynamically in the PRIAS cohort (interval validation) and in the largest six external cohorts that are part of the GAP3 database. For discrimination, we utilized the time-varying area under the receiver operating characteristic curve or time-varying AUC (Rizopoulos et al., 2017). For time-varying prediction error, we assessed the mean absolute prediction error or MAPE (Rizopoulos et al., 2017). The AUC indicates how well the model discriminates between patients who experience upgrading, and those do not. The MAPE indicates how accurately the model predicts upgrading. Both AUC and MAPE are restricted to $[0, 1]$. However, it is preferred that AUC $>$ 0.5 because an AUC $\leq$ 0.5 indicates that the model performs worse than random discrimination. Ideally, MAPE should be 0.

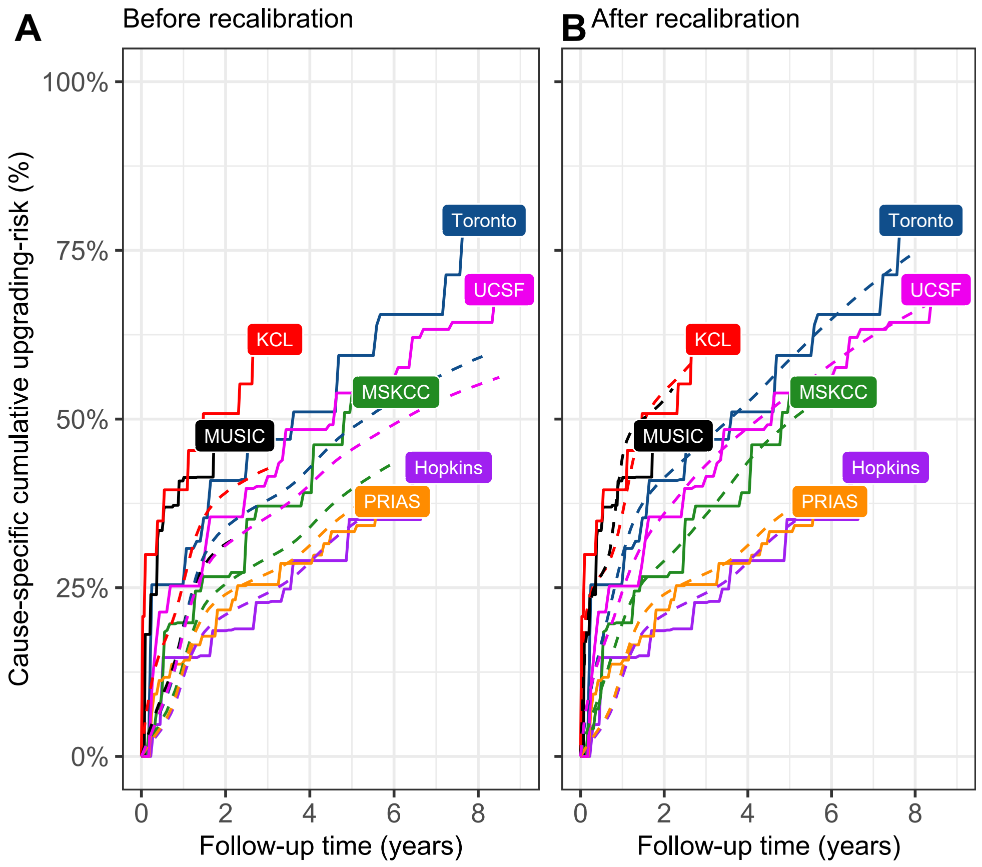

Figure 5.7: **Calibration-in-the-large of our model:**. In **Panel A** we can see that our model is not well calibrated for use in KCL, MUSIC, Toronto and MSKCC. In **Panel B** we can see that calibration of model predictions improved in KCL, MUSIC, Toronto and MSKCC cohorts after recalibrating our model. Recalibration was not necessary for Hopkins cohort. Full names of Cohorts are *PRIAS*: Prostate Cancer International Active Surveillance, *Toronto*: University of Toronto Active Surveillance, *Hopkins*: Johns Hopkins Active Surveillance, *MSKCC*: Memorial Sloan Kettering Cancer Center Active Surveillance, *KCL*: King's College London Active Surveillance, *MUSIC*: Michigan Urological Surgery Improvement Collaborative Active Surveillance, *UCSF*: University of California San Francisco Active Surveillance.

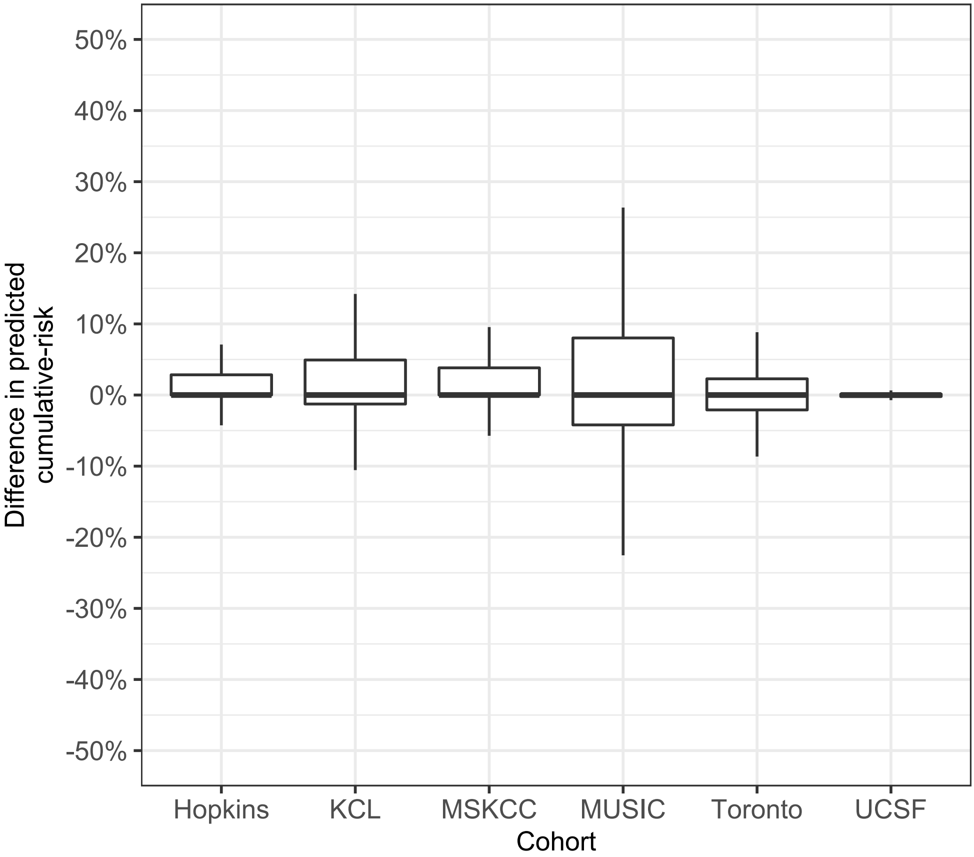

Figure 5.8: **Comparison of predictions from recalibrated PRIAS model with individual joint models fitted to external cohorts:** On Y-axis we show the difference between predicted cause-specific cumulative upgrading-risk for individual patients using two models, namely the recalibrated PRIAS model for each cohort, and individual joint model fitted to each cohort. The figure shows that the difference is smaller in those cohorts in which the effects of $\log_2(\text{PSA} + 1)$ value and velocity were similar to that of PRIAS (Table 5.7). Full names of Cohorts are *PRIAS*: Prostate Cancer International Active Surveillance, *Toronto*: University of Toronto Active Surveillance, *Hopkins*: Johns Hopkins Active Surveillance, *MSKCC*: Memorial Sloan Kettering Cancer Center Active Surveillance, *KCL*: King's College London Active Surveillance, *MUSIC*: Michigan Urological Surgery Improvement Collaborative Active Surveillance, *UCSF*: University of California San Francisco Active Surveillance.

Table 5.8: **Maximum follow-up period up to which we can reliably predict upgrading-risk and create personalized schedules**. In each cohort, this time point is chosen such that there are at least 10 patients who experience upgrading after this time point. Full names of Cohorts are *PRIAS*: Prostate Cancer International Active Surveillance, *Toronto*: University of Toronto Active Surveillance, *Hopkins*: Johns Hopkins Active Surveillance, *MSKCC*: Memorial Sloan Kettering Cancer Center Active Surveillance, *KCL*: King's College London Active Surveillance, *MUSIC*: Michigan Urological Surgery Improvement Collaborative Active Surveillance, *UCSF*: University of California San Francisco Active Surveillance.

Cohort	Maximum Prediction Time (years)
PRIAS	6
KCL	3
MUSIC	2
Toronto	8
MSKCC	6
Hopkins	7
UCSF	8.5

We calculate AUC and MAPE in a time-dependent manner. More specifically, given the time of latest biopsy t, and history of PSA measurements up to time v, we calculate AUC and MAPE for a medically relevant time frame $(t, v]$, within which the occurrence of upgrading is of interest. In the case of prostate cancer, at any point in time v, it is of interest to identify patients who may have experienced upgrading in the last one year $(v-1, v]$. That is, we set $t = v - 1$. We then calculate AUC and MAPE at a gap of every six months (follow-up schedule of PRIAS). That is, $v \epsilon \{1, 1.5, \ldots\}$ years. To obtain reliable estimates of AUC and MAPE, in each cohort, we restrict v to a maximum time point v_{max}, such that there are at least ten patients who experience upgrading after v_{max}. This maximum time point v_{max} differs between cohorts, and is given in Table 5.8.

The results for estimates of AUC and MAPE are summarized in Fig-

Table 5.9: **Internal validation of predictions of upgrading in PRIAS cohort**. The area under the receiver operating characteristic curve or AUC (measure of discrimination) and mean absolute prediction error or MAPE are calculated over the follow-up period at a gap of 6 months. In addition bootstrapped 95% confidence intervals (CI) are also presented.

Follow-up period (years)	AUC (95% CI)	MAPE (95%CI)
1.0 to 2.0	0.661 [0.647, 0.678]	0.187 [0.183, 0.191]
1.5 to 2.5	0.647 [0.596, 0.688]	0.129 [0.122, 0.140]
2.0 to 3.0	0.683 [0.642, 0.723]	0.135 [0.125, 0.146]
2.5 to 3.5	0.692 [0.632, 0.748]	0.118 [0.111, 0.128]
3.0 to 4.0	0.657 [0.603, 0.709]	0.086 [0.080, 0.092]
3.5 to 4.5	0.623 [0.582, 0.660]	0.111 [0.105, 0.116]
4.0 to 5.0	0.619 [0.582, 0.654]	0.126 [0.118, 0.131]
4.5 to 5.5	0.624 [0.537, 0.711]	0.119 [0.103, 0.135]
5.0 to 6.0	0.639 [0.582, 0.696]	0.121 [0.103, 0.138]

ure 5.9, and in Table 5.9 to Table 5.15. Results are based on the recalibrated PRIAS model for the GAP3 cohorts. The results show that AUC remains more or less constant in all cohorts as more data becomes available for patients. The AUC obtains a moderate value, roughly between 0.5 and 0.7 for all cohorts. On the other hand, MAPE reduces by a big margin after year one of follow-up. This could be because of two reasons. Firstly, MAPE at year one is based only on four PSA measurements gathered in the first year of follow-up, whereas after year one number of PSA measurements increases. Secondly, patients in year one consist of two sub-populations, namely patients with a correct Gleason grade group 1 at the time of inclusion in AS, and patients who probably had Gleason grade group 2 at inclusion but were misclassified by the urologist as Gleason grade group 1 patients. To remedy this problem, a biopsy for all patients at year one is commonly recommended in all AS programs (Bokhorst et al., 2015).

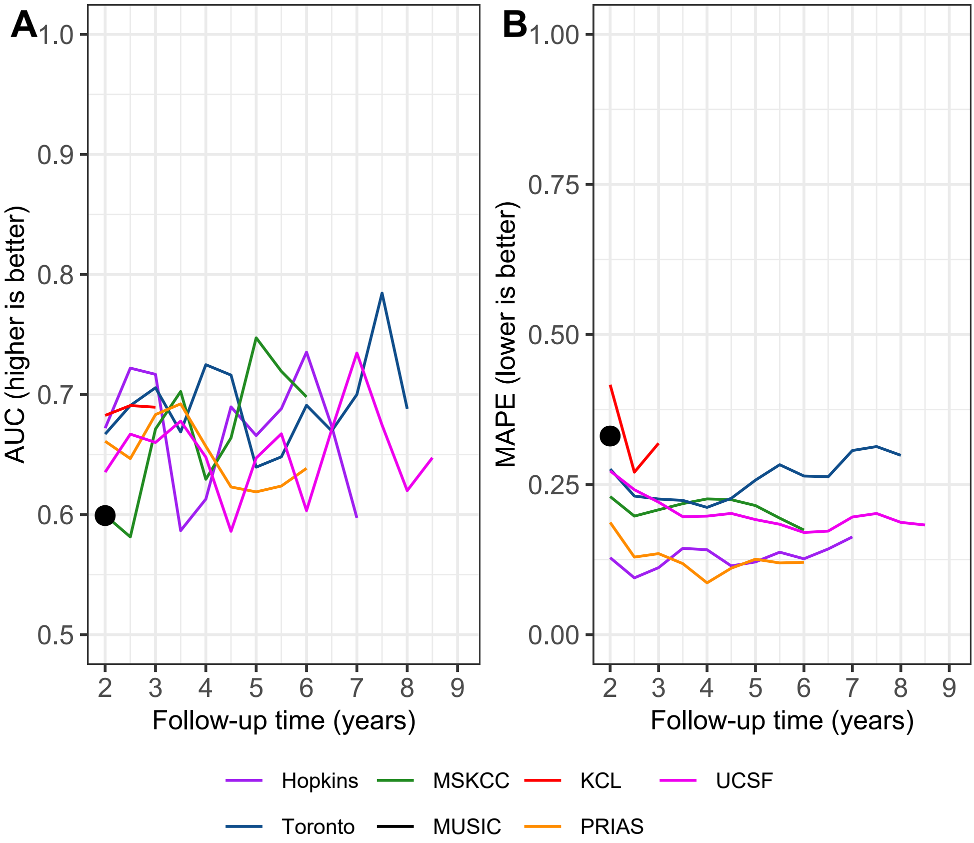

Figure 5.9: **Validation of dynamic predictions of cause-specific cumulative upgrading-risk**. In **Panel A** area under the receiver operating characteristic curve or AUC (measure of discrimination) is between 0.6 and 0.7. **Panel B** we can see that the time dependent root mean squared prediction error or MAPE is similar for PRIAS and Hopkins cohorts. The bootstrapped 95% confidence interval for these estimates are presented in Table 5.9 to Table 5.14. Full names of Cohorts are *PRIAS*: Prostate Cancer International Active Surveillance, *Toronto*: University of Toronto Active Surveillance, *Hopkins*: Johns Hopkins Active Surveillance, *MSKCC*: Memorial Sloan Kettering Cancer Center Active Surveillance, *KCL*: King's College London Active Surveillance, *MUSIC*: Michigan Urological Surgery Improvement Collaborative Active Surveillance, *UCSF*: University of California San Francisco Active Surveillance.

Table 5.10: **External validation of predictions of upgrading in University of Toronto Active Surveillance cohort**. The area under the receiver operating characteristic curve or AUC (measure of discrimination) and mean absolute prediction error or MAPE are calculated over the follow-up period at a gap of 6 months. In addition bootstrapped 95% confidence intervals (CI) are also presented.

Follow-up period (years)	AUC (95% CI)	MAPE (95%CI)
1.0 to 2.0	0.667 [0.634, 0.712]	0.276 [0.259, 0.296]
1.5 to 2.5	0.691 [0.651, 0.730]	0.231 [0.205, 0.254]
2.0 to 3.0	0.706 [0.637, 0.762]	0.226 [0.196, 0.260]
2.5 to 3.5	0.669 [0.586, 0.741]	0.224 [0.195, 0.258]
3.0 to 4.0	0.725 [0.649, 0.806]	0.212 [0.184, 0.238]
3.5 to 4.5	0.716 [0.642, 0.793]	0.227 [0.206, 0.258]
4.0 to 5.0	0.640 [0.579, 0.717]	0.257 [0.222, 0.312]
4.5 to 5.5	0.648 [0.579, 0.740]	0.283 [0.247, 0.326]
5.0 to 6.0	0.691 [0.608, 0.793]	0.264 [0.232, 0.302]
5.5 to 6.5	0.670 [0.543, 0.776]	0.263 [0.227, 0.307]
6.0 to 7.0	0.700 [0.544, 0.851]	0.307 [0.258, 0.363]
6.5 to 7.5	0.785 [0.640, 0.866]	0.313 [0.272, 0.360]
7.0 to 8.0	0.688 [0.532, 0.786]	0.299 [0.249, 0.361]

5.D Source Code

The R code for fitting the joint model to the PRIAS dataset, is at `https://github.com/anirudhtomer/prias/tree/master/src/clinical_gap3`. We refer to this location as 'R_HOME' in the rest of this document. The PRIAS dataset is not openly accessible. However, access to the database can be requested via the contact links at `https://www.prias-project.org`.

The PRIAS dataset is in the so-called wide format and also requires the removal of incorrect entries. This can be done via the R script `R_HOME/dataset_cleaning.R`. This will lead to two R objects, namely 'prias_final.id' and 'prias_long_final'. The 'prias_final.id' object contains information about the time of upgrading for PRIAS patients. The 'prias_long_final' object contains longitudinal PSA measurements, the time of biopsies and results of

Table 5.11: **External validation of predictions of upgrading in University of California San Francisco Active Surveillance cohort**. The area under the receiver operating characteristic curve or AUC (measure of discrimination) and mean absolute prediction error or MAPE are calculated over the follow-up period at a gap of 6 months. In addition bootstrapped 95% confidence intervals (CI) are also presented.

Follow-up period (years)	AUC (95% CI)	MAPE (95%CI)
1.0 to 2.0	0.635 [0.595, 0.677]	0.273 [0.266, 0.281]
1.5 to 2.5	0.667 [0.628, 0.715]	0.241 [0.224, 0.259]
2.0 to 3.0	0.660 [0.600, 0.713]	0.221 [0.205, 0.238]
2.5 to 3.5	0.678 [0.614, 0.757]	0.197 [0.175, 0.214]
3.0 to 4.0	0.648 [0.574, 0.707]	0.197 [0.179, 0.221]
3.5 to 4.5	0.586 [0.525, 0.638]	0.202 [0.180, 0.229]
4.0 to 5.0	0.647 [0.590, 0.754]	0.192 [0.168, 0.217]
4.5 to 5.5	0.667 [0.582, 0.773]	0.184 [0.159, 0.220]
5.0 to 6.0	0.603 [0.496, 0.696]	0.170 [0.144, 0.207]
5.5 to 6.5	0.671 [0.576, 0.786]	0.173 [0.145, 0.202]
6.0 to 7.0	0.735 [0.663, 0.794]	0.196 [0.166, 0.219]
6.5 to 7.5	0.675 [0.565, 0.769]	0.202 [0.168, 0.231]
7.0 to 8.0	0.620 [0.518, 0.740]	0.187 [0.144, 0.217]
7.5 to 8.5	0.647 [0.538, 0.787]	0.183 [0.146, 0.222]

biopsies.

We use a joint model for time-to-event and longitudinal data to model the evolution of PSA measurements over time, and to simultaneously model their association with the risk of upgrading. The R package we use for this purpose is called **JMbayes** (https://cran.r-project.org/web/packages/JMbayes/JMbayes.pdf). The API we use, however, is currently not hosted on CRAN, and can be found here: `https://github.com/anirudhtomer/JMbayes`. The joint model can be fitted via the script `R_HOME/analysis.R`. It takes roughly 6 hours to run on an Intel Core-i5 machine with four cores and 8GB of RAM.

The graphs presented in the main manuscript, and the appendix can be

Table 5.12: **External validation of predictions of upgrading in Johns Hopkins Active Surveillance cohort**. The area under the receiver operating characteristic curve or AUC (measure of discrimination) and mean absolute prediction error or MAPE are calculated over the follow-up period at a gap of 6 months. In addition bootstrapped 95% confidence intervals (CI) are also presented.

Follow-up period (years)	AUC (95% CI)	MAPE (95%CI)
1.0 to 2.0	0.672 [0.604, 0.744]	0.128 [0.115, 0.141]
1.5 to 2.5	0.722 [0.652, 0.792]	0.095 [0.081, 0.111]
2.0 to 3.0	0.717 [0.638, 0.777]	0.112 [0.100, 0.123]
2.5 to 3.5	0.587 [0.493, 0.704]	0.144 [0.129, 0.154]
3.0 to 4.0	0.613 [0.486, 0.742]	0.141 [0.126, 0.156]
3.5 to 4.5	0.690 [0.594, 0.783]	0.115 [0.100, 0.133]
4.0 to 5.0	0.666 [0.572, 0.754]	0.121 [0.104, 0.147]
4.5 to 5.5	0.688 [0.519, 0.779]	0.137 [0.119, 0.161]
5.0 to 6.0	0.735 [0.676, 0.820]	0.126 [0.102, 0.152]
5.5 to 6.5	0.674 [0.581, 0.765]	0.143 [0.121, 0.172]
6.0 to 7.0	0.597 [0.472, 0.712]	0.163 [0.126, 0.195]

generated by the scripts in `R_HOME/plots/`.

Validations can be done using the scripts `R_HOME/validation/auc_brier/auc_calculator.R`, and `R_HOME/validation/auc_brier/gof_calculator.R`. For external validation access to GAP3 database is required.

Once a joint model is fitted to the PRIAS dataset, personalized schedules of biopsies based on the risk of upgrading for new patients can be developed as shown in the script `R_HOME/plots/demo_schedule_supplementary.R` or directly using the script `https://raw.githubusercontent.com/anirudhtomer/prias/master/src/lastpaper/pers_schedule_api.R`.

Source code for the shiny web application which provides biopsy schedules for patients can be found at `R_HOME/shinyapp`

Table 5.13: **External validation of predictions of upgrading in Memorial Sloan Kettering Cancer Center Active Surveillance cohort**. The area under the receiver operating characteristic curve or AUC (measure of discrimination) and mean absolute prediction error or MAPE are calculated over the follow-up period at a gap of 6 months. In addition bootstrapped 95% confidence intervals (CI) are also presented.

Follow-up period (years)	AUC (95% CI)	MAPE (95%CI)
1.0 to 2.0	0.599 [0.518, 0.671]	0.230 [0.207, 0.256]
1.5 to 2.5	0.581 [0.504, 0.663]	0.198 [0.168, 0.235]
2.0 to 3.0	0.671 [0.599, 0.741]	0.208 [0.182, 0.232]
2.5 to 3.5	0.703 [0.610, 0.777]	0.218 [0.197, 0.246]
3.0 to 4.0	0.629 [0.499, 0.706]	0.226 [0.194, 0.259]
3.5 to 4.5	0.664 [0.589, 0.756]	0.225 [0.199, 0.262]
4.0 to 5.0	0.747 [0.642, 0.841]	0.215 [0.188, 0.247]
4.5 to 5.5	0.719 [0.597, 0.852]	0.194 [0.165, 0.232]
5.0 to 6.0	0.698 [0.565, 0.792]	0.174 [0.136, 0.227]

Table 5.14: **External validation of predictions of upgrading in King's College London Active Surveillance cohort**. The area under the receiver operating characteristic curve or AUC (measure of discrimination) and mean absolute prediction error or MAPE are calculated over the follow-up period at a gap of 6 months. In addition bootstrapped 95% confidence intervals (CI) are also presented.

Follow-up period (years)	AUC (95% CI)	MAPE (95%CI)
1.0 to 2.0	0.683 [0.604, 0.753]	0.416 [0.396, 0.445]
1.5 to 2.5	0.691 [0.621, 0.766]	0.271 [0.246, 0.297]
2.0 to 3.0	0.689 [0.616, 0.785]	0.319 [0.282, 0.344]

Table 5.15: **External validation of predictions of upgrading in Michigan Urological Surgery Improvement Collaborative Active Surveillance cohort**. The area under the receiver operating characteristic curve or AUC (measure of discrimination) and mean absolute prediction error or MAPE are calculated over the follow-up period at a gap of 6 months. In addition bootstrapped 95% confidence intervals (CI) are also presented.

Follow-up period (years)	AUC (95% CI)	MAPE (95%CI)
1.0 to 2.0	0.599 [0.553, 0.632]	0.331 [0.317, 0.348]

5.5 References

Ankerst, D. P., Xia, J., Thompson Jr, I. M., Hoefler, J., Newcomb, L. F., Brooks, J. D., Carroll, P. R., Ellis, W. J., Gleave, M. E., Lance, R. S., et al. (2015). Precision medicine in active surveillance for prostate cancer: development of the canary–early detection research network active surveillance biopsy risk calculator. *European Urology*, 68(6):1083–1088.

Balasubramanian, R. and Lagakos, S. W. (2003). Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika*, 90(1):171–182.

Bokhorst, L. P., Alberts, A. R., Rannikko, A., Valdagni, R., Pickles, T., Kakehi, Y., Bangma, C. H., Roobol, M. J., and PRIAS study group (2015). Compliance rates with the Prostate Cancer Research International Active Surveillance (PRIAS) protocol and disease reclassification in noncompliers. *European Urology*, 68(5):814–821.

Bratt, O., Carlsson, S., Holmberg, E., Holmberg, L., Johansson, E., Josefsson, A., Nilsson, A., Nyberg, M., Robinsson, D., Sandberg, J., et al. (2013). The study of active monitoring in Sweden (SAMS): a randomized study comparing two different follow-up schedules for active surveillance of low-risk prostate cancer. *Scandinavian Journal of Urology*, 47(5):347–355.

Briganti, A., Fossati, N., Catto, J. W., Cornford, P., Montorsi, F., Mottet, N., Wirth, M., and Van Poppel, H. (2018). Active surveillance for low-risk prostate cancer: the European Association of Urology position in 2018. *European Urology*, 74(3):357–368.

Bruinsma, S. M., Roobol, M. J., Carroll, P. R., Klotz, L., Pickles, T., Moore, C. M., Gnanapragasam, V. J., Villers, A., Rannikko, A., Valdagni, R., et al. (2017). Expert consensus document: semantics in active surveillance for men with localized prostate cancer—results of a modified Delphi consensus procedure. *Nature Reviews Urology*, 14(5):312.

Bruinsma, S. M., Zhang, L., Roobol, M. J., Bangma, C. H., Steyerberg, E. W., Nieboer, D., Van Hemelrijck, M., consortium, M. F. G. A. P. P. C. A. S. G., Trock, B., Ehdaie, B., et al. (2018). The Movember foundation's GAP3 cohort: a profile of the largest global prostate cancer active surveillance database to date. *BJU International*, 121(5):737–744.

Bul, M., Zhu, X., Valdagni, R., Pickles, T., Kakehi, Y., Rannikko, A., Bjartell, A., Van Der Schoot, D. K., Cornel, E. B., Conti, G. N., et al. (2013). Active surveillance for low-risk prostate cancer worldwide: the PRIAS study. *European Urology*, 63(4):597–603.

Chesnut, G. T., Vertosick, E. A., Benfante, N., Sjoberg, D. D., Fainberg, J., Lee, T., Eastham, J., Laudone, V., Scardino, P., Touijer, K., et al. (2020). Role of changes in magnetic resonance imaging or clinical stage in evaluation of disease progression for men with prostate cancer on active surveillance. *European Urology*, 77(4):501–507.

Coley, R. Y., Zeger, S. L., Mamawala, M., Pienta, K. J., and Carter, H. B. (2017). Prediction of the pathologic Gleason score to inform a personalized management program for prostate cancer. *European Urology*, 72(1):135–141.

Cooperberg, M. R., Brooks, J. D., Faino, A. V., Newcomb, L. F., Kearns, J. T., Carroll, P. R., Dash, A., Etzioni, R., Fabrizio, M. D., Gleave, M. E., et al. (2018). Refined analysis of prostate-specific antigen kinetics to predict prostate cancer active surveillance outcomes. *European Urology*, 74(2):211–217.

De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

de Carvalho, T. M., Heijnsdijk, E. A., and de Koning, H. J. (2017a). Estimating the risks and benefits of active surveillance protocols for prostate cancer: a microsimulation study. *BJU International*, 119(4):560–566.

de Carvalho, T. M., Heijnsdijk, E. A., and de Koning, H. J. (2017b). When should active surveillance for prostate cancer stop if no progression is detected? *The Prostate*, 77(9):962–969.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., and Humphrey, P. A. (2016). The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 40(2):244–252.

Inoue, L. Y., Lin, D. W., Newcomb, L. F., Leonardson, A. S., Ankerst, D., Gulati, R., Carter, H. B., Trock, B. J., Carroll, P. R., Cooperberg, M. R., et al. (2018). Comparative analysis of biopsy upgrading in four prostate cancer active surveillance cohorts. *Annals of Internal Medicine*, 168(1):1–9.

Kasivisvanathan, V., Giganti, F., Emberton, M., and Moore, C. M. (2020). Magnetic resonance imaging should be used in the active surveillance of patients with localised prostate cancer. *European Urology*, 77(3):318–319.

Laird, N. M., Ware, J. H., et al. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Lin, H., McCulloch, C. E., Turnbull, B. W., Slate, E. H., and Clark, L. C. (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine*, 19(10):1303–1318.

Loeb, S., Carter, H. B., Schwartz, M., Fagerlin, A., Braithwaite, R. S., and Lepor, H. (2014). Heterogeneity in active surveillance protocols worldwide. *Reviews in Urology*, 16(4):202–203.

Loeb, S., Vellekoop, A., Ahmed, H. U., Catto, J., Emberton, M., Nam, R., Rosario, D. J., Scattoni, V., and Lotan, Y. (2013). Systematic review of complications of prostate biopsy. *European Urology*, 64(6):876–892.

Makarov, D. V., Trock, B. J., Humphreys, E. B., Mangold, L. A., Walsh, P. C., Epstein, J. I., and Partin, A. W. (2007). Updated nomogram to predict pathologic stage of prostate cancer given prostate-specific antigen level, clinical stage, and biopsy Gleason score (Partin tables) based on cases from 2000 to 2005. *Urology*, 69(6):1095–1101.

Mottet, N., Bellmunt, J., Bolla, M., Briers, E., Cumberbatch, M. G., De Santis, M., Fossati, N., Gross, T., Henry, A. M., Joniau, S., et al. (2017). Eau-estro-siog guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent. *European Urology*, 71(4):618–629.

Nieboer, D., Tomer, A., Rizopoulos, D., Roobol, M. J., and Steyerberg, E. W. (2018). Active surveillance: a review of risk-based, dynamic monitoring. *Translational Andrology and Urology*, 7(1):106–115.

Nixon, R. G., Wener, M. H., Smith, K. M., Parson, R. E., Strobel, S. A., and Brawer, M. K. (1997). Biological variation of prostate specific antigen levels in serum: an evaluation of day-to-day physiological fluctuations in a well-defined cohort of 24 patients. *The Journal of Urology*, 157(6):2183–2190.

Partin, A. W., Yoo, J., Carter, H. B., Pearson, J. D., Chan, D. W., Epstein, J. I., and Walsh, P. C. (1993). The use of prostate specific antigen, clinical stage and Gleason score to predict pathological stage in men with localized prostate cancer. *The Journal of Urology*, 150(1):110–114.

Pearson, J. D., Morrell, C. H., Landis, P. K., Carter, H. B., and Brant, L. J. (1994). Mixed-effects regression models for studying the natural history of prostate disease. *Statistics in Medicine*, 13(5-7):587–601.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software*, 72(7):1–46.

Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276.

Royston, P. and Altman, D. G. (2013). External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, 13(1):33.

Schoots, I. G., Petrides, N., Giganti, F., Bokhorst, L. P., Rannikko, A., Klotz, L., Villers, A., Hugosson, J., and Moore, C. M. (2015). Magnetic resonance imaging in active surveillance of prostate cancer: a systematic review. *European Urology*, 67(4):627–636.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128.

Tomer, A., Nieboer, D., Roobol, M. J., Steyerberg, E. W., and Rizopoulos, D. (2019a). Personalized schedules for surveillance of low-risk prostate cancer patients. *Biometrics*, 75(1):153–162.

Tomer, A., Rizopoulos, D., Nieboer, D., Drost, F.-J., Roobol, M. J., and Steyerberg, E. W. (2019b). Personalized decision making for biopsies in prostate cancer active surveillance programs. *Medical Decision Making*, 39(5):499–508.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

Chapter 6

Personalized Screening Intervals for Measurement of N-terminal pro-B-type Natriuretic Peptide Improve Efficiency of Prognostication in Patients with Chronic Heart Failure

This chapter is an extended version of the paper
Schuurman, A.S., Tomer, A., Akkerhuis, K.M., Brugts, J.J., Constantinescu, A.A., van Ramshorst, J., Umans, V.A., Boersma, E., Rizopoulos, D., and Kardys, I. (2020). Personalized screening intervals for measurement of N-terminal pro-B-type natriuretic peptide improve efficiency of prognostication in patients with chronic heart failure. *European Journal of Preventive Cardiology*. Advance online publication. doi:10.1177/2047487320922639

Abstract

Aims. Personalized screening intervals for N-terminal pro-B-type natriuretic peptide (NT-proBNP) measurement in patients with chronic heart failure (CHF) could maximize information gain on individual patients' disease progression, while minimizing the number of necessary measurements. To improve prevention of clinical adverse events, we compared personalized scheduling of NT-proBNP measurements to fixed scheduling.

Methods. In 263 CHF patients from the Bio-SHiFT study, NT-proBNP was measured trimonthly according to a predefined fixed schedule. The primary endpoint (PE) comprised cardiac death, cardiac transplantation, left ventricular assist device implantation or heart failure hospitalization. We jointly modeled the repeated NT-proBNP measurements and PE. Using this fitted joint model, for each patient at each follow-up visit, we decided the optimal time point of the next NT-proBNP measurement based on the patient's individual NT-proBNP evolution. Personalized scheduling was compared to fixed scheduling by means of a simulation study, based on a replica of the Bio-SHiFT study population. Specifically, we compared the schedules' capability of identification of a high-risk interval (time-window with high risk preceding the PE; identification of its start enables appropriate timely intervention and prevention of PE occurrence), and number of measurements needed.

Results. Compared to fixed scheduling, personalized scheduling saved on average 2 measurements, while the start of the high-risk interval was similar by both approaches [personalized, Median: 6.6, IQR: 4.5-11.3; fixed, Median: 6.3, IQR: 4.2-10.3; months before occurrence of PE].

Conclusion. Personalized scheduling of NT-proBNP measurements in CHF patients, as compared to fixed scheduling, shows similar performance with regard to identification of impending adverse events, but requires fewer NT-proBNP measurements.

6.1 Introduction

Circulating biochemical markers (biomarkers) may reflect the deterioration of patients with chronic heart failure (CHF) in an earlier stage than clinical assessment does. Hence, these biomarkers carry the potential to improve the risk stratification of patients with CHF and prevention of adverse clinical events (Masson et al., 2008; Gaggin and Januzzi Jr, 2013). In the past decade, several trials on natriuretic peptide-guided therapy have been performed in which serial natriuretic peptide measurements were used to titrate medication (Khan et al., 2018; Felker et al., 2017). However, these trials have demonstrated inconclusive results. This may, in part, be explained by the fact that they mostly used predefined screening intervals (i.e., predefined time points) to assess biomarkers, as well as predefined target levels. Such predefined screening intervals and target levels do not account for individual temporal patterns of biomarkers, which may hamper their potential use for therapy guidance.

Conversely, a personalized screening approach that individualizes screening intervals and target levels based on individual temporal biomarker patterns may further improve risk assessment and therapy guidance. Such personalized screening intervals aim to maximize information gain on the individual patients' disease progression, while minimizing the necessary number of measurements, and therewith costs and patient burden (Rizopoulos et al., 2016). In order to establish such intervals and targets, a model should be applied that incorporates detailed data on individual temporal patterns. Joint modeling is a statistical approach that takes into account full individual temporal patterns of biomarkers and links these patterns to the occurrence of adverse clinical events (Rizopoulos, 2016; Rizopoulos and Takkenberg, 2014). In the Role of Biomarkers and Echocardiography in Prediction of Prognosis of Chronic Heart Failure Patients (Bio-SHiFT) study, we collected a median of 9 [interquartile range (IQR): 5–10] blood samples per patient. We demonstrated, by applying joint modeling, that individual temporal patterns of serially measured CHF-related biomarkers are associated with the prognosis of CHF patients (van Boven et al., 2018). Furthermore, we demonstrated that such a joint model when fitted on patients in Bio-SHiFT, could be used to estimate the patient-specific risk of the adverse outcome at each visit at the outpatient clinic. This risk is updated at each visit because it incorporates information on the patients' prognosis as derived from the newly available biomarker measurement (van Boven et al., 2018).

Subsequently, such a patient-specific risk profile, as derived from a joint model, can be applied to establish personalized screening intervals for future patients presenting at the outpatient clinic. This approach could contribute to improved prevention of further adverse clinical events. However, the benefits of this approach, over predefined screening intervals and targets, have not yet been investigated in CHF. Thus, in the current investigation, we aim to compare personalized scheduling to predefined fixed scheduling of N-terminal pro-B-type natriuretic peptide (NT-proBNP) measurements in individual CHF patients, in terms of the number of measurements performed according to each schedule, as well as the amount of time that remains for intervention before adverse outcome occurs. For this purpose, we use the data of the Bio-SHiFT study.

6.2 Methods

6.2.1 Study design and procedures

The design of the Bio-SHiFT study has been described in detail elsewhere (van Boven et al., 2018). Briefly, CHF patients in clinically stable conditions were recruited during their regular outpatient visits in the Erasmus MC, Rotterdam, The Netherlands, and Northwest Clinics, Alkmaar, The Netherlands. Patients were eligible if CHF (with reduced or preserved ejection fraction) was diagnosed $\geq$ 3 months ago according to the guidelines of the European Society of Cardiology (Members et al., 2012; Paulus et al., 2007; Dickstein et al., 2008). Blood samples were taken on the day of inclusion and at predefined trimonthly follow-up visits, which were scheduled to a maximum follow-up duration of 30 months. Blood sampling and study procedures are further described in the Supplemental Materials. For the current investigation, we used 263 patients who were enrolled during the first inclusion period between October 2011 and June 2013.

During follow-up, the occurrence of clinical events was recorded in the electronic case report forms, and associated hospital records and discharge letters were collected. Subsequently, a clinical event committee, blinded to the biomarker-candidate results, reviewed hospital records and discharge letters, and adjudicated the study endpoints. The primary study endpoint (PE) was defined as the composite of cardiac death, cardiac transplantation, left ventricular assist device implantation, or hospitalization for heart failure, whichever occurred first.

The Bio-SHiFT study was approved by the medical ethics committee of the Erasmus MC and was performed in accordance with the Declaration of Helsinki. Written informed consent was obtained from all patients. The Bio-SHiFT study is registered in ClinicalTrials.gov, number NCT01851538.

6.2.2 Statistical analysis

We utilized a joint model to estimate the association between longitudinally measured NT-proBNP and clinical outcome (Rizopoulos, 2012; Tsiatis and Davidian, 2004). A joint model combines a linear mixed-effect (LME) model for longitudinally measured data with a Cox regression model for time-to-event data. The association between these two types of data is modeled using patient-specific random effects. The LME model uses these random-effects to model the longitudinal temporal pattern of NT-proBNP measurements. The Cox model uses these random-effects to model the impact of the underlying trajectory of NT-proBNP measurements on the risk of PE (Rizopoulos et al., 2016; van Boven et al., 2018). The use of joint modeling is further motivated in Supplemental Materials. We used logarithmically (base 2) transformed NT-proBNP measurements in our joint model. Consequently, we were able to obtain a hazard ratio (HR) along with a 95% confidence interval (CI) that estimated the risk of the PE associated with doubling of NT-proBNP level at a given follow-up time (van Boven et al., 2018).

The potential confounders that we used in our joint model were chosen based on their independent association with the PE in multivariable Cox regression models (NYHA class and diabetes mellitus) and existing literature (age, gender, renal function, body mass index). Covariates were missing in less than 3% of the patients. Multiple imputations (5 times) of these covariates were performed in the multivariable analyses.

6.2.3 Scheduling personalized screening visits

The scheduling of personalized screening visits is based on the individual patients' longitudinal biomarker profile. A patient visiting the outpatient clinic has longitudinal NT-proBNP measurements available until a certain time point. From the aforementioned joint model, we can derive for each individual patient the cumulative-risk of the PE at a particular follow-up time point, using all of the previously measured NT-proBNP up until this time point.

For determining the optimal time point for drawing the next blood sample in a particular patient, we first need to establish the cumulative-risk of PE occurring in a certain time window. The time point for drawing the next blood sample should not be beyond the time point at which the PE occurs. For this reason, we set a maximum limit on the time window based on the cumulative-risk of the PE. Then, the time window is defined as the time between the current measurement and the maximum possible time point of drawing the next measurement. We aim to find the optimal time point to draw the next blood sample within this time window. We also need to define a risk threshold, which, if crossed within the time window, leads us to stop the further scheduling of measurements since the patient apparently needs appropriate action and/or increased surveillance, and therefore a different protocol from that point onwards. For this investigation, we have selected a risk threshold of 7.5% for the three months that follow, based on clinical considerations. Thus, if the patients' cumulative-risk of the PE exceeds 7.5% within the following three months, we stop scheduling further measurements in order to, for example, adjust therapy to avoid the occurrence of the PE. For the current investigation, we focus on the personalized screening schedules themselves, as our primary aim is to enable timely intervention before the occurrence of the PE. Hence, for now, we do not propose a specific therapy to be used at the time point that the patients' cumulative-risk of the PE exceeds the risk threshold.

On the other hand, if the cumulative-risk of the PE remains less than 7.5% within the following three months, we would like to determine the optimal time point at which to obtain the next NT-proBNP measurement. The selection of this optimal time point is based on two aspects.5 First, as stated, the cumulative-risk of PE in the time window should not exceed 7.5%. Second, obtaining an NT-proBNP measurement at this optimal time point should provide us the maximum amount of information about the future cumulative-risk of PE for this particular patient. Accordingly, we perform personalized scheduling using the stepwise approach depicted in Figure 6.1. Altogether, when applying this approach, patients with relatively stable biomarker profiles will likely not exceed the predefined risk threshold within a specified time window, and the calculations may suggest waiting for a longer time period to perform the next biomarker measurement in these patients. On the other hand, patients with worsening biomarker profiles are more likely to exceed the predefined risk threshold within a specified time window, and the calculations may suggest performing the next biomarker measurement in the short term.

6.2.4 Simulation study

After constructing the joint model and defining the thresholds needed for scheduling personalized screening visits, we proceeded to compare the personalized screening schedule to a fixed screening schedule. For the fixed schedule, we chose trimonthly intervals, in accordance with the design of the Bio-SHiFT study and daily clinical practice. Since our existing data were collected using this fixed screening schedule and hence no 'real' data on personalized screening intervals was available, the advantages of a personalized screening design were assessed by means of a simulation study. We first sim-

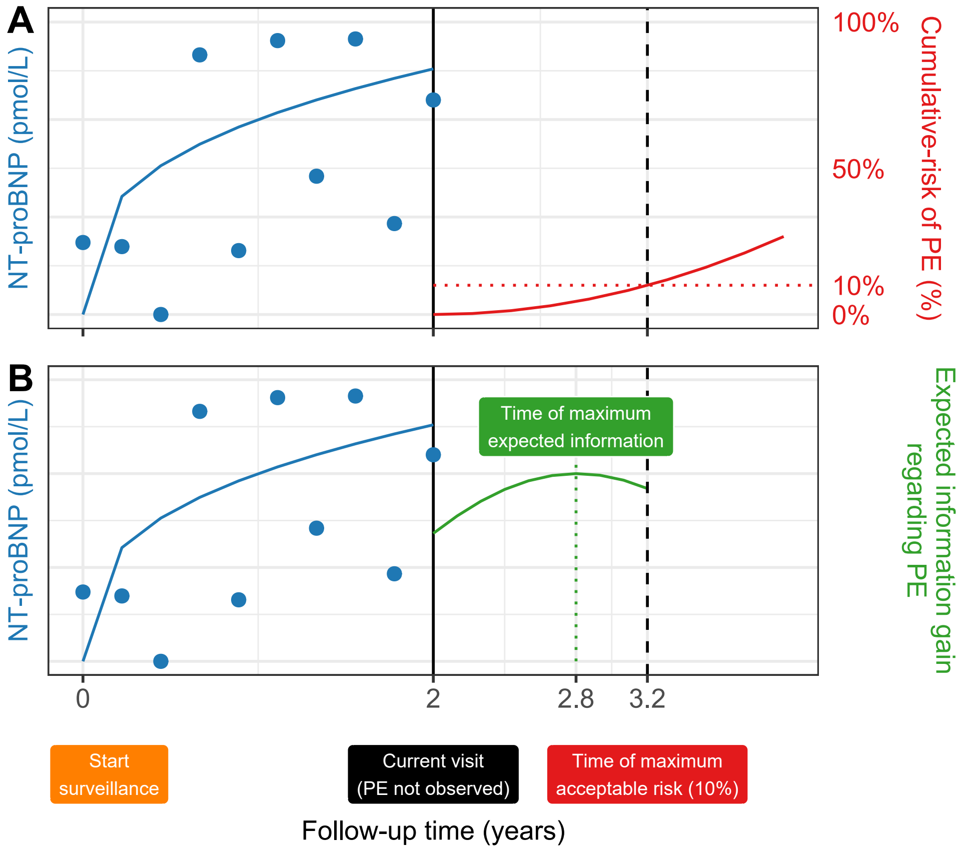

Figure 6.1: **Illustration of personalized scheduling of biomarker measurements.** We plan NT-proBNP measurements until the cumulative-risk of PE (primary endpoint) at three months from the current visit is more than 10%. **Panel A**: Example patient with longitudinal NT-proBNP measurements and fitted profile (in blue). The time of the current visit, on which PE was not observed, is year 2. Using NT-proBNP and time of current visit data, we derive a personalized cumulative-risk profile for the patient (in red). This risk profile reaches the 10% level at year 3.2, and hence, we are allowed to schedule new measurements until year 3.2. **Panel B**: We calculate the expected information gain in the patient's prognosis if a new NT-proBNP measurement is done at a future time point between the current visit at year 2 and the time of the maximum acceptable risk of 10% at year 3.2. The time of maximum expected information gain, is year 2.8, and hence, we schedule new NT-proBNP measurement at year 2.8.

ulated a dataset containing 750 patients. These 750 simulated patients had baseline characteristics and NT-proBNP profiles similar to the 263 patients included in the Bio-SHiFT since we simulated using the joint model fitted to the Bio-SHiFT data. We divided this data into training (700 patients), and testing (50 patients) set. For the training patients, using the joint model fitted to the Bio-SHiFT data, we generated NT-proBNP measurements at fixed follow-up time points. This schedule is similar to the schedule of the Bio-SHiFT study. We also generated a true PE time for these patients, as well as a random non-informative censoring time. Subsequently, we fitted a new joint model for these patients and, then, used this model to develop NT-proBNP measurement schedules for the test patients. To this end, in the test patients, we only generated the true PE time. Using such a design ensured that the 'new' patients (n=50) are comparable to the 'existing' patients (n=700) on which the model is based; if we had used the 'real' patients (n=263 from the Bio-SHiFT study), this might not have been the case.

Thus, for each of the 50 patients in the test set, we aimed to compare the efficacy of scheduling NT-proBNP measurements according to a fixed screening design and a personalized screening design. For the personalized screening design, the first three simulated NT-proBNP measurements were considered a given, in order to have a 'run-in period' for the patients' longitudinal profile of NT-proBNP, since if we have a longitudinal profile available we can apply the aforementioned stepwise approach of personalized scheduling. Apart from using the risk-threshold of 7.5% over a 3-month period, we repeated the analysis using 5% and 10% risk thresholds. We did so because 5% is a lower cumulative-risk, and consequently, scheduling will stop earlier than in case of a 7.5% risk threshold, which will give us more time to intervene with respect to the true PE time. Conversely, 10% is a higher risk percentage than 7.5%, and hence schedules based on the former will give us less time.

The performance of the personalized and fixed screening schedules was compared using two outcome measures, namely, the start of the high-risk interval and the number of scheduled measurements. The high-risk interval was defined as the estimated intervention time minus the true event time (in months) (Fig. 1D). Thus, the schedule that showed a high-risk interval that was larger in absolute terms (i.e., more negative) was preferred, because such a high-risk interval enables timely intervention. In addition, assuming that the costs of NT-proBNP measurements and outpatient visits remained the same during follow-up, we prefer a procedure that requires the fewest possible repeated measurements (Supplemental Materials). All analyses were performed with R statistical software using package JMBayes (Rizopoulos, 2016).

6.3 Results

6.3.1 Baseline characteristics

The mean age of the patients was 66.7 years, and 71.9% were men (Table 6.1). Most patients were in NYHA class I or II (73.8%). The median baseline NT-proBNP value was 137.3 pmol/L (IQR: 51.7-272.6). A total of 2022 NT-proBNP measurements were performed during follow-up before the PE occurred. The PE occurred in 70 patients (26.6%). The median maximum follow-up time was 2.1 (IQR: 1.2–2.4) years.

6.3.2 Association between temporal patterns of NT-proBNP and the PE

Serially measured NT-proBNP was associated with the PE (univariable HR per doubling of NT-proBNP: 2.13, 95%CI: 1.81–2.53, $p<0.001$). After adjustment for age, gender, diabetes mellitus, NYHA class, body mass index, and renal function, serially measured NT-proBNP remained independently associated with the PE (adjusted HR per doubling of NT-proBNP: 2.20, 95%CI :1.84–2.68, $p<0.001$). Two examples of dynamic risk assessment of individual patients from the Bio-SHiFT dataset based on the joint model are demonstrated in Supplemental Materials, Figure 1A-B.

Table 6.1: **Summary of the Bio-SHiFT dataset**. The primary study endpoint (PE) was defined as the composite of cardiac death, cardiac transplantation, left ventricular assist device implantation, or hospitalization for heart failure, whichever occurred first. Abbreviations: NYHA is New York Heart Association Classification (Bredy et al., 2018); IQR is interquartile range.

Characteristic	Value
Total patients	263
PE (primary endpoint)	70
Total NT-proBNP measurements	2022
Median NT-proBNP (pg/mL)	110.3 (IQR: 38.5–240.9)
Median age at inclusion (years)	67.9 (IQR: 58.9–75.8)
Median BMI at inclusion	26.5 (IQR: 24.4–30.1)
Median NYHA (assumed continuous)	2 (IQR: 1–3)
Gender = Female (%)	74/263 (28.1%)
Renal failure history = Yes (%)	136/263 (51.7%)
Type-II diabetes mellitus = Yes (%)	81/263 (30.8%)
Median maximum follow-up per patient (years)	2.1 (IQR: 1.2–2.4)
Median #NT-proBNP per patient	9 (IQR: 5–10)

6.3.3 Fixed versus personalized screening schedule: high-risk interval and number of measurements

The median follow-up time of the 750 patients in the simulated dataset was 1.76 years (IQR: 1.42–2.24); mean (standard deviation) was 1.85 (0.63) years and the maximum was 3.5 years. The personalized schedule used fewer measurements as compared to the fixed (Panel A, Figure 6.2). The personalized schedule used a median of 7 (IQR: 7–8) and the fixed schedule, a median of 9 (IQR: 8–10) measurements. Corresponding cost estimates are demonstrated in the Supplemental Materials. The start of the high-risk intervals for the fixed and personalized screening schedules are depicted in Figure 2B. The personalized and fixed schedules showed similar results, i.e., the difference between the estimated intervention time compared to the 'true' event time was a median of 6.6 (IQR: 4.5–11.3) months for the personalized and a median of 6.3 (IQR: 4.2–10.3) months for the fixed schedule (Panel B, Figure 6.2). In both schedules, scheduling of new sampling moments was stopped in order to undertake appropriate action, well in time before the event occurred.

Results of the analyses using risk thresholds of consecutively 5% and 10% over three months are depicted in the Supplemental Materials, Figure 6.3, and Figure 6.4, respectively. Based on a risk threshold of 5% over three months, the fixed and personalized screening schedules demonstrated similar results for the high-risk interval. However, again, the personalized screening schedule used fewer measurements as compared to the fixed screening schedule. The same was true for the risk threshold of 10%. In case of a risk threshold of 5% over three months, we found that the start of the high-risk interval was further away from the true event time as compared to a risk threshold of 7.5% over three months. Conversely, in case of a risk threshold of 10% over three months, the start of the high-risk interval was closer to the true event time as compared to a risk threshold of 7.5% over three months. These results comply with the increase in the risk threshold.

6.4 Discussion

This study aimed to optimally schedule NT-proBNP measurements for individual patients with CHF while maximizing the gain in prognostic information and reducing costs. Furthermore, to compare the efficacy of such personalized scheduling with fixed scheduling. We found that over a median follow-up time of 1.8 years, personalized scheduling required fewer NT-proBNP measurements per patient as compared to fixed scheduling while demonstrating similar performance regarding the prevention of adverse cardiac events. Since personalized scheduling required fewer measurements, this approach is expected to reduce related health care costs as well as patient burden compared to fixed scheduling.

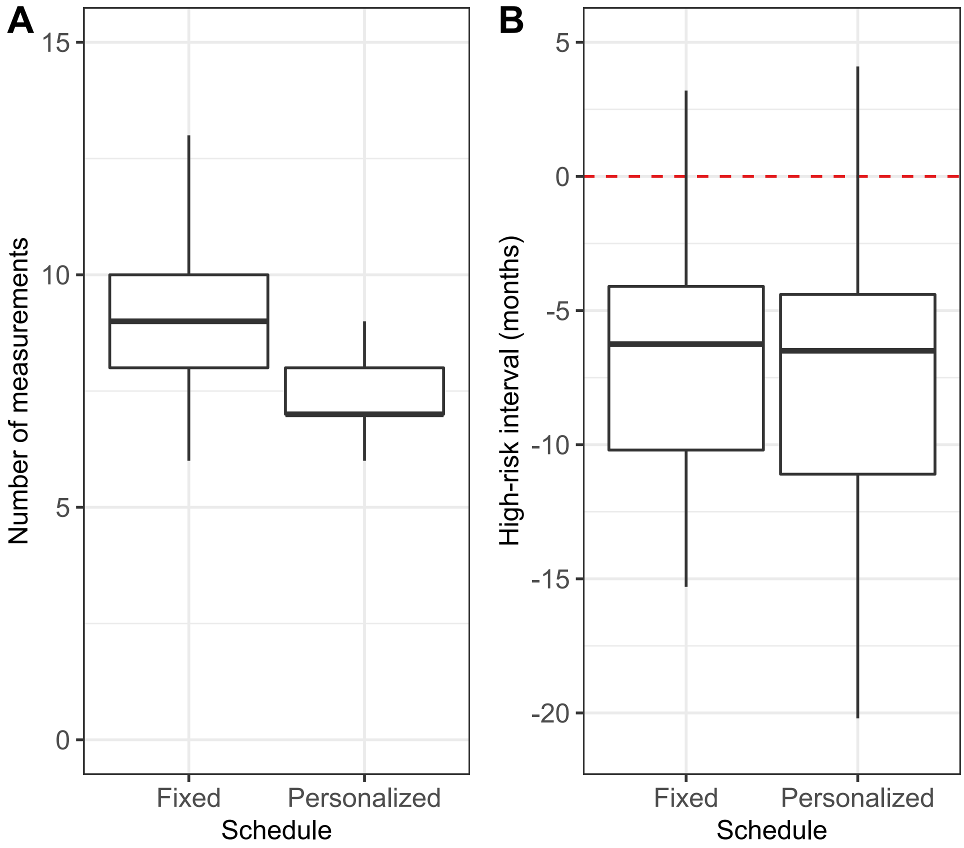

Figure 6.2: **Total NT-proBNP measurements and high-risk interval (months) preceding the primary endpoint (PE)** for fixed (quarterly measurements) and personalized schedules. Results are based on a realistic simulation study of 263 test patients. NT-proBNP is measured as per personalized and fixed schedules, until a patient's cumulative-risk of obtaining PE in the subsequent three months is above 7.5%. The boxplot for the number of measurements in Panel A is made using data of all simulated patients. The boxplot for the high-risk interval (the difference between the time at which NT-proBNP measurements are stopped and the true simulated PE time) in Panel B, is based on only those patients who observe PE. In Panel B a zero high-risk interval (dashed red line) indicates that no time is available for intervention before occurrence of PE.

The findings from our study carry important implications for future trials on biomarker-guided therapy. Previous biomarker-guided trials have generally used predefined sampling intervals and target levels (Khan et al., 2018; Felker et al., 2017). We show that, by using a personalized approach for scheduling NT-proBNP, timely intervention is enabled while using fewer NT-proBNP measurements as compared to a fixed schedule. Even though our fixed schedule consisted of rather frequent (trimonthly) NT-pro-BNP measurements, the high-risk interval identified by the personalized schedule was similar. On top of this, the fixed schedule was outperformed by the personalized schedule in terms of the number of measurements needed per patient to obtain this result. Maximizing information gain by estimating prognosis in an individual and optimal manner, while minimizing healthcare burden, may provide novel opportunities for timely adaptation of treatment. Future trials on natriuretic peptide-guided therapy for chronic heart failure may benefit from incorporating personalized screening intervals and personalized biomarker value targets; tailoring therapeutic interventions using this approach may reveal benefits that could not be demonstrated by previous trials, by nature of their design.

Previous studies on personalized scheduling of blood sampling moments for measurements of biomarkers of disease are scarce, but this topic seems to be gaining attention recently. Personalized scheduling has been applied to patients undergoing aortic allograft root implantation (Rizopoulos et al., 2016). Similarly to our study, this study used joint modeling. Aortic gradient levels were measured according to a fixed screening schedule. The authors demonstrated that personalized scheduling of aortic gradient assessments required fewer measurements and also performed better regarding the prevention of recurrent events as compared to fixed scheduling. Recently, personalized schedules for reducing the number of biopsies in low-risk prostate cancer patients have also been developed (Tomer et al., 2019). Altogether, these promising results in other disease areas concur with our conclusion that personalized screening intervals carry the potential to improve patient monitoring and to ultimately individualize and herewith improve treatment.

6.4.1 Limitations

Several aspects of this study warrant consideration. First, we made several assumptions when developing the model, defining the thresholds, and setting up of the simulation study. However, in a sensitivity analysis, we performed the simulation study for three different risk thresholds, and the results remained essentially unchanged. Second, in our investigation, we performed a so-called demonstration, meaning that the analysis was performed on one ‘test’ set of 50 patients, and was not repeated. We performed a demonstration because we aimed to provide a proof-of-concept here. A study with multiple test sets should be performed to validate our findings further. Although, it should be noted that such repeated estimations pose a heavy computational burden. Third, we did not account for the costs of implementation. Finally, while the concept of personalized screening intervals we present here seems promising, whether it would actually lead to the prevention of adverse events remains to be investigated in a clinical trial.

6.4.2 Conclusions

In conclusion, this study demonstrates for the first time that personalized scheduling of NT-proBNP measurements in patients with CHF, as compared to fixed scheduling, shows similar performance with regard to prevention of recurrent events but requires fewer NT-proBNP measurements. If such personalized scheduling were to be applied in natriuretic peptide-guided therapy, these benefits might translate into improved outcomes. Therefore, a clinical trial incorporating personalized scheduling should be considered.

Author Contributions

Study Design: Akkerhuis, Umans, Boersma, Kardys
Data analysis: Schuurman, Tomer
Drafting Manuscript: Schuurman, Tomer
Analytics Strategy: Rizopoulos, Kardys

Directing Implementation: Umans, Kardys
Quality Control: Kardys
Data Interpretation: All authors
Critical revision of manuscript for important intellectual content: All authors

Appendix

6.A Details: Materials and Methods

6.A.1 Study procedures and outcome measures

During their baseline and follow-up visits, all patients were evaluated by research physicians, who collected information on CHF-related symptoms, New York Heart Association (NYHA) class, and performed a physical examination. Information on CHF etiology, left ventricular ejection fraction, cardiovascular risk factors, medical history, and treatment was retrieved primarily from hospital records and was checked in case of ambiguities. History of cardiovascular and other comorbidities was defined as clinical diagnosis thereof reported in the hospital records.

6.A.2 Blood sampling and NT-proBNP measurement

Blood samples were processed and stored at a temperature of -80 degrees C within 2 hours after blood collection. When applicable, samples were transported to the central laboratory (Erasmus MC, Rotterdam, the Netherlands) under controlled conditions (at a temperature of -80 degrees C) until batch analysis was performed. Accordingly, results of the biomarker assays were not available to treating physicians at the time of the outpatient visits and hence did not alter patient care. Plasma NT-proBNP was analyzed using an Electrochemiluminescence immunoassay (Elecsys 2010; Roche Diagnostics, Indianapolis, IN), which measures concentrations ranging from 5 to 35000 ng/L.

6.A.3 Joint Modeling

We utilized a joint model to estimate the association between longitudinally measured NT-proBNP and clinical outcomes. A joint model combines a linear mixed-effect (LME) model for longitudinally measured data with a Cox regression model for time-to-event data. The use of joint modeling was motivated by the following considerations. In the Bio-SHiFT study, NT-proBNP levels were measured trimonthly until the PE occurred, or until the patient was censored. Thus, naturally, patients who experienced a PE had NT-proBNP measurements available over a shorter time-course than those who did not experience a PE; or in other words, further NT-proBNP measurements could be considered as missing due to occurrence of the PE. However, commonly used methods to model such longitudinal data, for example, linear mixed-effect (LME) models, assume that missing data is non-informative with regard to a patient's health status. In other words, they do not account for the fact that patients with missing NT-proBNP values are more likely to have higher NT-proBNP levels (if hypothetically, we would have been able to observe them). This may lead to bias in the parameter estimates. In addition, a classical time-dependent Cox regression may be used in order to measure the impact of NT-proBNP on the PE. However, due to the aforementioned issue, the time-dependent Cox model may also be biased. To correctly estimate the effects, the parameters of these two types of models are required to be estimated jointly. We did so by applying the joint model.

6.A.4 Costs

Based on the number of scheduled measurements by the personalized and the fixed scheduling approach, we compared the cost estimates from the perspective of Erasmus MC as well as the perspective of society at large for both scheduling approaches. Cost estimates from the perspective of the Erasmus MC included costs of NT-proBNP sampling and measurement, as well as visiting the Cardiology outpatient clinic, at this particular institution (Kanters et al., 2017; Hakkaart-van Roijen et al., 2015). Cost estimates from the perspective of society at large included average Dutch cost estimates of NT-proBNP sampling and measurement and average Dutch cost estimates of visiting a Cardiology outpatient clinic. Moreover, patients' travel costs and patients' production losses associated with visiting the outpatient clinic were included (Kanters et al., 2017; Hakkaart-van Roijen et al., 2015).

6.B Supplemental Results

6.B.1 Fixed versus personalized screening schedule: costs

Costs are depicted in Table 6.2. From the perspective of the Erasmus MC, the costs associated with a visit to the Cardiology outpatient clinic, including blood sampling and NT-proBNP measurement, were €182.1 Thus, since the personalized screening schedule required two visits less than the fixed schedule (Figure 6.1), from the perspective of the Erasmus MC the costs saved by personalized scheduling were on average €364 per patient, over a mean follow-up of 1.76 years (€207 saved per patient per year).

From the perspective of society at large, costs for visiting the outpatient clinic, blood sampling, and NT-proBNP measurement were €106.1 Travel costs and production losses amounted to €6 and €33, respectively. Altogether, costs per visit amounted to €145, with, on average, €290 saved per patient by personalized scheduling versus fixed scheduling, again over a mean follow-up of 1.76 years (€165 saved per patient per year).

In The Netherlands, the prevalence of CHF is estimated at 227,000 patients (`www.nivel.nl/node/4309`). In this context, personalized screening could reduce the involved annual costs by approximately €37 million from the perspective of society at large.

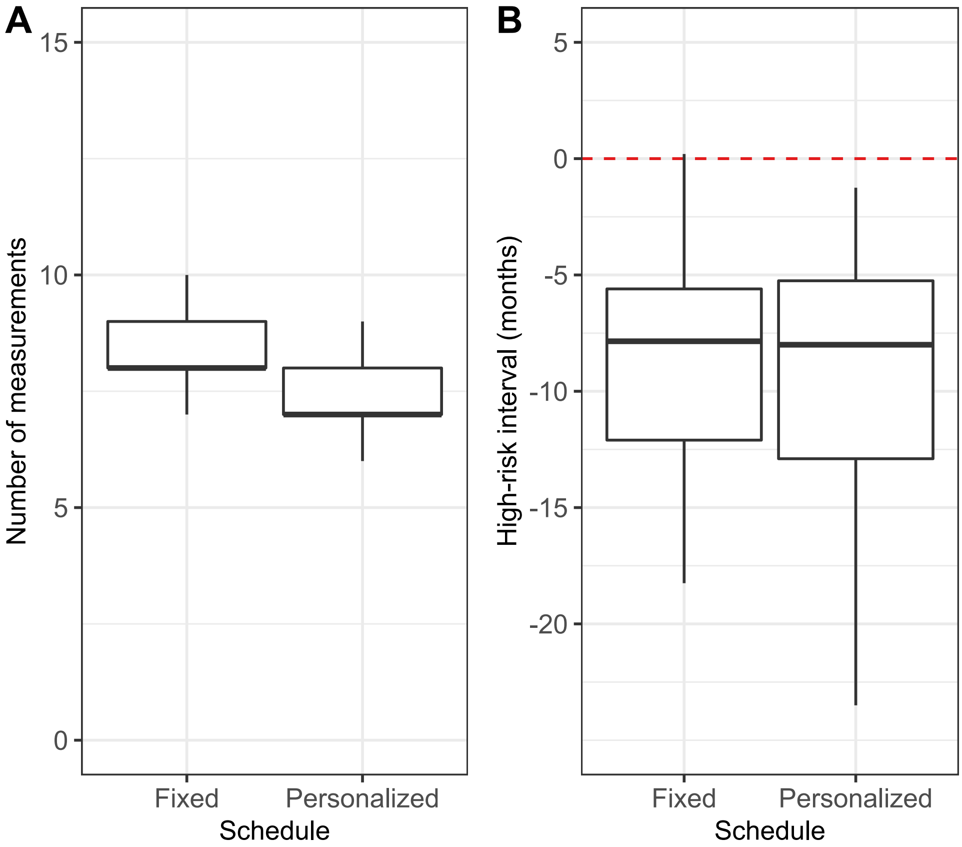

Figure 6.3: **Total NT-proBNP measurements and high-risk interval (months) preceding the primary endpoint (PE)** for fixed (quarterly measurements) and personalized schedules. Results are based on a realistic simulation study of 263 test patients. NT-proBNP is measured as per personalized and fixed schedules, until a patient's cumulative-risk of obtaining PE in the subsequent three months is above 5%. The boxplot for the number of measurements in Panel A is made using data of all simulated patients. The boxplot for the high-risk interval (the difference between the time at which NT-proBNP measurements are stopped and the true simulated PE time) in Panel B, is based on only those patients who observe PE. In Panel B a zero high-risk interval (dashed red line) indicates that no time is available for intervention before occurrence of PE.

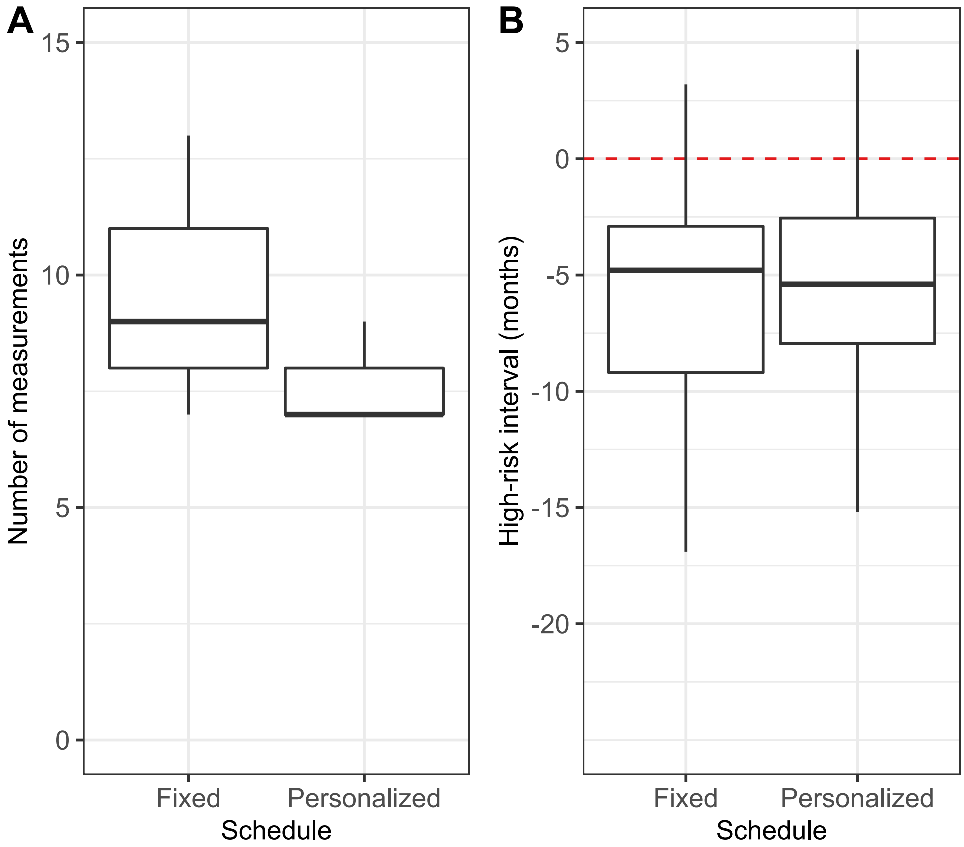

Figure 6.4: **Total NT-proBNP measurements and high-risk interval (months) preceding the primary endpoint (PE)** for fixed (quarterly measurements) and personalized schedules. Results are based on a realistic simulation study of 263 test patients. NT-proBNP is measured as per personalized and fixed schedules, until a patient's cumulative-risk of obtaining PE in the subsequent three months is above 10%. The boxplot for the number of measurements in Panel A is made using data of all simulated patients. The boxplot for the high-risk interval (the difference between the time at which NT-proBNP measurements are stopped and the true simulated PE time) in Panel B, is based on only those patients who observe PE. In Panel B a zero high-risk interval (dashed red line) indicates that no time is available for intervention before occurrence of PE.

Table 6.2: **Cost estimates from the perspective of the Erasmus MC and society at large.** Abbreviations are, CHF: chronic heart failure; NL: The Netherlands; NT-proBNP: N-terminal pro-B-type natriuretic peptide.

Costs	Erasmus MC (€)	Society at large (€)
NT-proBNP measurement, per measurement	19	15
Visit to outpatient clinic, per visit	163	91
Travel costs, per visit	-	6
Production loss costs, per visit	-	33
Total costs, per visit	182	145
Total costs for fixed schedule (median of 9 measurements)	1638	1305
Total costs for personalized schedule (median of 7 measurements)	1274	1015
Costs saved by personalized scheduling, per patient	364	290
Annual costs saved by personalized scheduling (prevalence CHF in NL estimated at 227,000 patients)	-	37,455,000

6.3 References

Bredy, C., Ministeri, M., Kempny, A., Alonso-Gonzalez, R., Swan, L., Uebing, A., Diller, G.-P., Gatzoulis, M. A., and Dimopoulos, K. (2018). New York Heart Association (NYHA) classification in adults with congenital heart disease: relation to objective measures of exercise and outcome. *European Heart Journal-Quality of Care and Clinical Outcomes*, 4(1):51–58.

Dickstein, K., Members, A. F., Cohen-Solal, A., Filippatos, G., McMurray, J. J., Ponikowski, P., Poole-Wilson, P. A., Strömberg, A., van Veldhuisen, D. J., Atar, D., Hoes, A. W., Keren, A., Mebazaa, A., Nieminen, M., Priori, S. G., Swedberg, K., Vahanian, A., for Practice Guidelines (CPG), E. C., Camm, J., De Caterina, R., Dean, V., Dickstein, K., Filippatos, G., Funck-Brentano, C., Hellemans, I., Kristensen, S. D., McGregor, K., Sechtem, U., Silber, S., Tendera, M., Widimsky, P., Zamorano, J. L., Tendera, M., Reviewers, D., Auricchio, A., Bax, J., Böhm, M., Corrà, U., della Bella, P., Elliott, P. M., Follath, F., Gheorghiade, M., Hasin, Y., Hernborg, A., Jaarsma, T., Komajda, M., Kornowski, R., Piepoli, M., Prendergast, B., Tavazzi, L., Vachiery, J.-L., Verheugt, F. W. A., Zamorano, J. L., and Zannad, F. (2008). ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008‡. *European Journal of Heart Failure*, 10(10):933–989.

Felker, G. M., Anstrom, K. J., Adams, K. F., Ezekowitz, J. A., Fiuzat, M., Houston-Miller, N., Januzzi, J. L., Mark, D. B., Piña, I. L., Passmore, G., et al. (2017). Effect of natriuretic peptide–guided therapy on hospitalization or cardiovascular mortality in high-risk patients with heart failure and reduced ejection fraction: a randomized clinical trial. *JAMA*, 318(8):713–720.

Gaggin, H. K. and Januzzi Jr, J. L. (2013). Biomarkers and diagnostics in heart failure. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1832(12):2442–2450.

Hakkaart-van Roijen, L., Van der Linden, N., Bouwmans, C., Kanters, T., and Tan, S. (2015). Costing manual: Methodology of costing research and reference prices for economic evaluations in healthcare. 2015. *Institute for Medical Technology Assessment, Erasmus University Rotterdam*.

Kanters, T. A., Bouwmans, C. A., van der Linden, N., Tan, S. S., and Hakkaart-van Roijen, L. (2017). Update of the Dutch manual for costing studies in health care. *PLoS One*, 12(11).

Khan, M. S., Siddiqi, T. J., Usman, M. S., Sreenivasan, J., Fugar, S., Riaz, H., Murad, M. H., Mookadam, F., and Figueredo, V. M. (2018). Does natriuretic peptide monitoring improve outcomes in heart failure patients? a systematic review and meta-analysis. *International Journal of Cardiology*, 263:80–87.

Masson, S., Latini, R., Anand, I. S., Barlera, S., Angelici, L., Vago, T., Tognoni, G., Cohn, J. N., Investigators, V.-H., et al. (2008). Prognostic value of changes in n-terminal pro-brain natriuretic peptide in val-heft (valsartan heart failure trial). *Journal of the American College of Cardiology*, 52(12):997–1003.

Members, A. F., McMurray, J. J., Adamopoulos, S., Anker, S. D., Auricchio, A., Böhm, M., Dickstein, K., Falk, V., Filippatos, G., Fonseca, C., Gomez-Sanchez, M. A., Jaarsma, T., Køber, L., Lip, G. Y., Maggioni, A. P., Parkhomenko, A., Pieske, B. M., Popescu, B. A., Rønnevik, P. K., Rutten, F. H., Schwitter, J., Seferovic, P., Stepinska, J., Trindade, P. T., Voors, A. A., Zannad, F., Zeiher, A., for Practice Guidelines (CPG), E. C., Bax, J. J., Baumgartner, H., Ceconi, C., Dean, V., Deaton, C., Fagard, R., Funck-Brentano, C., Hasdai, D., Hoes, A., Kirchhof, P., Knuuti, J., Kolh, P., McDonagh, T., Moulin, C., Popescu, B. A., Reiner, e., Sechtem, U., Sirnes, P. A., Tendera, M., Torbicki, A., Vahanian, A., Windecker, S., Reviewers, D., McDonagh, T., Sechtem, U., Bonet, L. A., Avraamides, P., Ben Lamin, H. A., Brignole, M., Coca, A., Cowburn, P., Dargie, H., Elliott, P., Flachskampf, F. A., Guida, G. F., Hardman, S., Iung, B., Merkely, B., Mueller, C., Nanas, J. N., Nielsen, O. W., Ørn, S., Parissis, J. T., and Ponikowski, P. (2012). ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. *European Heart Journal*, 33(14):1787–1847.

Paulus, W. J., Tschöpe, C., Sanderson, J. E., Rusconi, C., Flachskampf, F. A., Rademakers, F. E., Marino, P., Smiseth, O. A., De Keulenaer, G., Leite-Moreira, A. F., et al. (2007). How to diagnose diastolic heart failure: a consensus statement on the diagnosis of heart failure with normal left ventricular ejection fraction by the Heart Failure and Echocardiography Associations of the European Society of Cardiology. *European Heart Journal*, 28(20):2539–2550.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software*, 72(7):1–46.

Rizopoulos, D. and Takkenberg, J. (2014). Tools & techniques–statistics: Dealing with time-varying covariates in survival analysis–joint models versus Cox models. *EuroIntervention: Journal of EuroPCR in collaboration with the Working Group on Interventional Cardiology of the European Society of Cardiology*, 10(2):285–288.

Rizopoulos, D., Taylor, J. M., Van Rosmalen, J., Steyerberg, E. W., and Takkenberg, J. J. (2016). Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics*, 17(1):149–164.

Tomer, A., Nieboer, D., Roobol, M. J., Steyerberg, E. W., and Rizopoulos, D. (2019). Personalized schedules for surveillance of low-risk prostate cancer patients. *Biometrics*, 75(1):153–162.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

van Boven, N., Battes, L. C., Akkerhuis, K. M., Rizopoulos, D., Caliskan, K., Anroedh, S. S., Yassi, W., Manintveld, O. C., Cornel, J.-H., Constantinescu, A. A., et al. (2018). Toward personalized risk assessment in patients with chronic heart failure: detailed temporal patterns of NT-proBNP, troponin T, and CRP in the Bio-SHiFT study. *American Heart Journal*, 196:36–48.

Part III

Summary

Chapter 7

General Discussion

7.1 Background

Low-risk chronic non-communicable disease (e.g., localized prostate cancer, low-risk dysplasia) patients often undergo repeated invasive *tests* (biopsies, endoscopies, etc.) for confirming disease *progression*. A progression is a non-terminal event upon which patients usually undergo serious treatments, e.g., surgery, radiotherapy. Typically, invasive tests are conducted routinely according to a one-size-fits-all (e.g., yearly) fixed schedule (Bokhorst et al., 2015; Choi and Hur, 2012; Krist et al., 2007; McWilliams et al., 2008; Henderson et al., 2011). Invasive tests are burdensome (Loeb et al., 2013; Krist et al., 2007) but also indispensable for timely detection of disease progression. Specifically, frequent one-size-fits-all test schedules promise shorter time delays in observing progression at the cost of imposing an extra burden on patients who progress slowly. The vice versa holds for infrequent tests. Our aim in this thesis was to balance better the number of tests (burden) and time delay in detecting progression (shorter is beneficial) than fixed schedules. To this end, we developed and applied statistical methods for scheduling invasive diagnostic tests (e.g., biopsies, endoscopies) in a personalized manner.

To create personalized test schedules we first utilized a statistical model to predict a patient's cumulative-risk of progression over the whole follow-up period based on his accumulated clinical data. This risk profile manifested the transition of a patient's disease state over time from low-risk to progressed. Hence, subsequently, we used it to guide the timing of future invasive tests. Specifically, we derived personalized test schedules by optimizing utility functions of clinical parameters of interest (Chapters 2, 3, and 4) under the estimated patient-specific cumulative-risk of progression. We also employed the cumulative-risk of progression to assess the widely utilized approach of scheduling invasive tests using partially observable Markov decision processes (Chapter 4.E). We then implemented personalized biopsy schedules for real patients of the seven largest prostate cancer active surveillance programs (Chapter 5) in a web-application (`https://emcbiostatistics.shinyapps.io/prias_biopsy_recommender/`). The use of personalized schedules is not limited to invasive tests only. In this regard, we also applied them for planning the NT-proBNP biomarker (a simple blood test) measurements in chronic heart failure patients (Chapter 6).

7.2 Subgoals and Research Questions

7.2.1 Statistical Modeling Framework to Process Observed Patient Data

We used the framework of joint models for time-to-event and longitudinal data (Rizopoulos, 2012; Tsiatis and Davidian, 2004) to process a patient's data and predict his patient's cumulative-risk of progression. We chose this for three main reasons. First, joint models utilize patient-specific random-effects and are hence inherently personalized. Second, they accommodate outcomes of various types, including longitudinally measured, baseline patient data, and results from previous invasive tests. Third, they update risk predictions automatically over follow-up as more patient data become available.

There are some limitations of the joint models that we utilized. First, in our model, we selected predictors based on existing hypotheses regarding the clinical relevance of the predictors. However, such hypotheses can change with time, e.g., PSA velocity (Vickers et al., 2014). An alternative is choosing model predictors based on their predictive ability. Second, we ignored competing-risk scenarios and estimated only cause-specific cumulative-risk of progression. Third, our model did not account for the sampling error and inter-observer variation in invasive test results. However, these limitations apply only to the model, i.e., the scheduling methodology remains the same, even if a better model is employed.

7.2.2 Pros of Cons of Different Utility Functions

Utility functions: squared and absolute loss, Parameter of interest: time of progression In Chapter 2, we optimized two commonly used loss functions, namely, expected squared and absolute loss (Robert, 2007) for the time of progression to decide the time of the next invasive test. These loss functions plan an invasive test at the estimated mean (square loss) and median (absolute loss) time of progression of a patient. There are two limitations to this method. First, due to the limited follow-up period of real-world studies, only a restricted version of the mean time of progression can be calculated, which has no straightforward interpretation. Second, depending upon the standard deviation of the posterior predictive distribution of time of progression, the difference between the actual and mean time of progression can be substantial. Both mean and median time of progression may sound suitable when interpreted as the central tendency of the distribution for time of progression. However, they also represent the time at which a patient has a 50% risk of progression, which may seem large clinically.

Utility function: multi-linear loss (risk-based tests), Parameter of interest: time of progression Squared and absolute loss penalize equally a planned invasive test that exceeds or falls short of the time of progression. However, patients may weigh these two scenarios unequally, especially because multiple tests before the actual time of progression can be burdensome. In this regard, a multi-linear loss function (Chapter 2) can be used. This loss function plans a test at a time point at which the patient's risk of progression is equal to a particular risk threshold. Smaller risk thresholds plan a test earlier than higher risk thresholds. In other words, smaller risk thresholds penalize exceeding the time progression more than falling short of the time of progression.

The main caveat in such risk-based test decisions (Chapter 3) is the choice of risk threshold. This threshold may be chosen by patients and/or doctors according to how they weigh the relative harms of doing an unnecessary test versus a missed disease progression (e.g., 10% threshold means a

1:9 ratio) if the test is not conducted (Vickers and Elkin, 2006). Threshold choice can also be partly data-driven, e.g., based on the threshold's estimated sensitivity and specificity of diagnosing progression. In this regard, it is also possible to choose risk thresholds using more sophisticated measures than just sensitivity or specificity, e.g., Youden's index, F1 score (López-Ratón et al., 2014). However, a critical limitation of such measures is that they may automatically select a risk threshold with clinically unsuitable sensitivity and specificity. Also, typically such measures are difficult to interpret.

Utility function: Partially observable Markov decision process (POMDP) value function In Chapter 4.E we explored the framework of POMDP to schedule invasive tests. The choice of POMDP was motivated by its wide usage in numerous optimal screening and surveillance test schedules for chronic diseases (Steimle and Denton, 2017; Denton, 2018), and especially for nearly all types of cancers (Alagoz et al., 2010).

In general, POMDPs utilize estimates from previously conducted studies or surveys to build a disease state transition model. In this work, we did not have to rely on existing studies as we made a joint model for modeling the patient's disease state and the associated clinical data. We integrated joint models with POMDPs by replacing the Bayes-rule based belief (risk of disease progression, see Equation 4.11) update of POMDPs, with the dynamic predictions (Rizopoulos et al., 2017) of cumulative-risk of progression from joint models. This had the advantage that it also personalized the POMDP. Concerning the use of clinical data, POMDPs assume that the probability distribution of future longitudinal data adds extra information over observed data. However, we estimated the posterior probability distribution of future longitudinal data based on observed data using the joint model. Hence, sampling a new observation from the future distribution and using it to update belief adds no information. This was not our sole reason for ignoring the posterior probability distribution of future longitudinal data. Another was that POMDP algorithms suffer from the curse of dimensionality with continuous longitudinal outcomes (Sunberg and Kochenderfer, 2018).

We also observed a more substantial drawback of POMDPs lying in their very flexible specification. Specifically, in a simple POMDP with binary test/no test decisions, and binary disease state (low-risk, progressed), it can be shown that there exist infinite possible rewards result in the same optimal schedule (Chapter 4.E). Typically POMDP rewards are chosen based on survey results (Denton, 2018) and translated as quality-adjusted life-years saved. However, with infinite optimal reward sets, any reward set can be cherry-picked, including those that correspond to (improbable) thousands of quality-adjusted life-years saved. Hence, POMDPs may find the most optimal schedule, but to achieve that, the choice of suitable rewards is tough in practice.

Utility function: Euclidean distance, Parameters of interest: expected number of tests and time delay in detecting progression Motivated by the POMDP framework we improved over planning one future test at a time (Chapter 2 and 3), by replacing it with a whole schedule of tests planned until a maximum future time point (e.g., end of the study period) in Chapter 4. To this end, we made risk threshold based test decisions iteratively at each future follow-up visit (e.g., future visits for biomarker measurement) of the patient. An important question, in this case, is which risk threshold yields the optimal schedule? To assist patients and doctors in this endeavor, we proposed a utility function to find the optimal risk threshold based schedule. The utility function was the Euclidean distance between a risk-based schedule and a perfect schedule (one test planned at the exact time of progression) in a bi-dimensional space of the expected number of tests and time delay in detecting progression. While we included only risk-based schedules in the Euclidean space, given L future visits of a patient, the Euclidean distance can be used to find the optimal schedule among all 2^L possible schedules.

The main advantage of this approach is that we directly minimize quantities that manifest burden and benefit. Second, personalized schedules get updated with newly collected data over follow-up. Third, Euclidean distance is easier to understand compared to squared loss or the recursive POMDP value function. A drawback of our approach is that we are only able to schedule tests up to a maximum horizon time. Another caveat is that a fair comparison of time delays between different schedules for the same patient requires a compulsory test at a common horizon time point in all schedules.

7.2.3 Criteria for Comparison of Schedules

An important question for patients and doctors is if personalized schedules are any better than fixed schedules. Especially if personalized schedules improve upon patient deaths and/or progression to an advanced disease state (e.g., metastasis) compared to fixed schedules. However, to our knowledge, currently, there are no running studies that compare personalized versus fixed schedules. Also, reliable data on the number of patient deaths are difficult to obtain in low-grade diseases. This is because, in such diseases, the prevalence of death from disease can be quite low. For example, in the PRIAS prostate cancer dataset, only two out of 7813 patients were reported to die from prostate cancer.

In search of the criteria for comparison of schedules, in Chapter 4.3.4 we proposed a method to calculate expected number of tests (burden) per schedule and expected time delay in detecting progression (shorter is beneficial) for any schedule, fixed or personalized. These two are easily-quantifiable surrogates for important clinical aspects, such as the window of opportunity for curative treatment, risk of adverse downstream outcomes, quality-adjusted remaining lifetime, and additional complications in treating a delayed progression. Based on these two quantities, patients/doctors can compare any schedule with any other and decide which schedule suits them best. Both expected number of tests and expected time delay in detecting progression are calculated in a personalized manner. That is, two patients may be prescribed the exact same schedule of tests, but their expected time delay in detecting progression and the expected number of tests will depend on their disease progression risk profile.

7.2.4 Factors Affecting Performance of Personalized Schedules

We utilized the estimated cumulative-risk of disease progression for developing personalized schedules throughout this work. Thus, how accurately this risk profile resembles the actual disease state dictates the performance of the personalized schedules. In this regard, both the quality of the clinical data and the accuracy of the model are important. For example, in the prostate cancer surveillance scenario, we found that prostate-specific antigen (PSA) and its derivatives, such as PSA velocity, were not very strong predictors of progression (Chapter 5). Besides, the PSA measurement error was best modeled with a t-distribution with three degrees of freedom. As a result, until a rise in PSA was evident via multiple observations and the surge was sharp as well, the predicted risk profile remained almost the same. Another effect of this was that once a negative result was obtained on a biopsy, a patient's predicted risk remained low until PSA consistently showed a sharp rise. Consequently, our model discriminated poorly (Chapter 5.C) between patients who obtained progression versus patients who did not obtain progression within a short time period of their last biopsy (e.g., an year). Overall, while building a model for personalized schedules, the focus should be on predictive ability of the model.

7.2.5 Reusing a Test Scheduling Framework Across Different Cohorts and Diseases

The utility functions and methodology we developed in this work is generic for scheduling tests in both screening and surveillance scenarios. It is also not necessary to rely only on joint models. The proposed utility functions only require an estimate of cumulative-risk of progression. While we focused on balancing the number of tests and time delay in detecting progression, users can extend the proposed methodology to include other aspects such as quality-adjusted life years. Our methodology may work differently in certain diseases. For example, in Barrett's esophagus (Choi and Hur, 2012), longitudinal data (e.g., p53 and SOX2 protein expressions) is collected when endoscopy (invasive test) is conducted. Consequently, risk predictions and personalized schedules do not update unless a test is conducted. Based on our findings from validation of the PRIAS (prostate cancer data introduced in Chapter 1.3.1) based model in external datasets (Chapter 5.C), a model may require recalibration (e.g., of baseline-risk) before reusing in other cohorts. Alternatively, one may fit a new model to the new cohort before using the model for personalized schedules.

7.3 Recommendations for Practice, and Future Improvements

Based on the work done in this thesis, we have certain recommendation for practitioners as well a list of improvements that may be researched in future. We next provide these recommendations and improvements in each of the four steps of creating invasive test schedules (defined in Chapter 1.1.3). These are, namely, processing the observed data of the patient, choosing the reward/utility/loss function and the corresponding clinical parameters, comparing proposed personalized schedules with currently practiced schedules, and implementing personalized schedules in a computer application for practitioners.

7.3.1 Observed Data of the Patient

Our overall recommendation while developing a model for personalized schedules is to focus on the predictive performance of the model. For example, in the prostate cancer surveillance joint model, we used a limited set of predictors (e.g., PSA value and velocity). However, this did not necessarily lead to the model with the best prediction error, and/or capacity of discrimination (e.g., the area under the receiver operating characteristic curve) between patients who are progressing and non-progressing patients. In this regard, although a joint model is a suitable candidate for modeling both longitudinal and time-to-event data, practitioners may also explore other models, and strive to achieve the best predictions. In addition, there are four broad areas of improvement in modeling observed data of the patient than the standard joint model we used.

Sampling error In general invasive procedures such as biopsies are prone to sampling error, especially when the site for sampling the tissues is not chosen carefully. In this regard, Coley et al. (2017) have proposed a joint model that predicts the time of actual disease progression given the time of observed progression. Actual progression means progression in the absence of sampling error, e.g., progression based on complete tumor such as via surgical removal, rather than biopsy samples.

Inter-observer variation Invasive test results are also prone to inter-observer variation. Models that have been proposed (Balasubramanian and Lagakos, 2003) to handle this problem require an estimate of the size and direction of the inter-observation variation. However, if the inter-observer variation for a test is inherently large, then overall model parameter estimates for risk of progression may also inflate.

Recurrent cancer We did not consider the scenario of surveillance of recurrent occurrence of progression (e.g., recurrent breast cancer). Although, if a model predicts risk of progression while accounting for recurrent events (Rizopoulos, 2012), the scheduling methodology that we proposed may not change.

Competing-risks In this work, we assumed all competing events to be non-informative censoring. However, treatment without progression or death may occur in a substantial number of patients. This affects our methodology in three ways. First, while creating risk-based schedules; second, during the calculation of the expected number of tests given a schedule; and third, in the calculation of expected time delay in detecting progression given a schedule. Among these, schedules based on cause-specific cumulative-risk may be alternatively obtained using the cumulative-incidence function (Andrinopoulou et al., 2017; Putter et al., 2007). The expected number of tests requires a model in which all events are combined to form a composite event. This is because the occurrence of any of the competing events will stop surveillance. On the other hand, time delay in detecting progression may only be defined if progression occurs before any other competing event.

7.3.2 Choice of Reward/Utility Function

The majority of the scheduling methods we explored in this work optimized the expected value of a utility function. Although, another key aspect is the variance of the utility. This is because a schedule with sub-optimal expected utility but lower utility variance may seem more trustworthy owing to its consistency. For example, the squared loss has a high variance in delay in detecting progression despite having expected delay equal to zero (Chapter 2). Second, we relied on the number of invasive tests and time delay in detecting disease progression as measures of burden and benefit of scheduling an invasive test. However, it is also important to consider patient anxiety, risk of complications, risk of non-compliance, and quality-adjusted life-years (QALY). Among these, the reference measure in a cost-utility analysis is QALY (Sassi, 2006). Although it can be included in the utility functions we proposed (Chapter 4), the biggest challenge is calculating QALYs correctly in different disease surveillance.

Optimizing the schedule of invasive tests and longitudinal outcomes together In Chapters 2 to 5, we optimized the schedule of invasive tests, whereas in Chapter 6 we optimized the schedule of the longitudinal outcome. In diseases such as chronic heart failure (Chapter 6), being able to measure the longitudinal outcome also indicates the survival of the patient. This is because if the patient obtains the event (cardiac failure), the longitudinal outcome cannot be measured. Whereas in a disease such as Barrett's esophagus, longitudinal outcomes are also measured via endoscopy (invasive). Hence, there is no need to optimize the schedule of invasive tests and the longitudinal outcomes separately. Conversely, in prostate cancer active surveillance, longitudinal outcomes can be measured even after patient obtains progression because progression in prostate cancer active surveillance is a non-terminal event. Hence, an interesting question is if we can optimize both the schedule of invasive tests and multiple longitudinal measurements together.

Optimizing the schedule of invasive tests and longitudinal outcomes together has the potential to reduce patient burden even further. However, it has certain caveats too. For example, in Chapter 5, we observed that PSA is not a strong indicator of progression. Thus, fewer PSA measurements will mean increased variance of estimates of cumulative-risk of progression. On the other hand, in diseases where the biomarker is a strong indicator of progression, measuring it less frequently may still provide reasonably accurate estimates of progression. Hence, our overall recommendation while pursuing this research problem is that for all longitudinal biomarkers, we first calculate a quantitative estimate of their burden (both financial and medical) and their predictive ability. Perhaps, optimal results can be obtained by employing a combination of, frequently measuring a cheaper but poor indicator of progression with infrequently measuring an expensive but reliable indicator of progression.

7.3.3 Simulation of Invasive Test Randomized Clinical Trials

While simulating hypothetical patients, we ignored the correlation between a patient's baseline features, and the patient's non-compliance to invasive tests. Generating correlated predictors will make the simulation cohort more realistic. Whereas, accounting for patient non-compliance to invasive tests will correct the currently likely over-estimated benefit of personalized schedules over fixed schedules.

7.3.4 Linking Patient Databases with Web-interfaces for Personalized Schedules

Risk-calculators for disease progression, including our web-application for prostate cancer surveillance patients, are not difficult to create given the current support for web-based technologies in the R framework. The larger challenge is linking the risk calculators with patient databases. We recommend the current industry standard of RESTful Web services (Rodriguez, 2008) for this purpose.

7.4 General Conclusion

In this work, we explored personalized schedules for invasive diagnostic tests in chronic non-communicable disease surveillance. To detect disease progression timely, in surveillance, typically invasive tests are planned in a one-size-fits-all manner, or flowcharts are used for test protocols. Neither of these methods exploit patient data fully. In contrast, the proposed personalized schedules rely on joint models for time-to-event and longitudinal data, which utilize complete patient data, including baseline covariates, longitudinal outcomes, and results of previous tests. The model building process is crucial in obtaining effective personalized schedules. Specifically, a model that predicts progression with a low error and a high capacity for discrimination will lead to personalized schedules that better balance the burden and benefit of repeated tests.

Personalized schedules are not a panacea, and there is no single schedule that is suitable for all patients. In this regard, our methodology for estimating the expected number of tests and expected time delay in detecting progression in a patient-specific manner for any schedule can assist patients and doctors in shared decision making of an appropriate schedule. It is also essential to implement personalized schedules in Internet and web-applications, separately for each disease surveillance. Currently, we have done so for prostate cancer active surveillance patients. We hope this work will motivate surveillance studies to investigate personalized schedules more, e.g., via a randomized clinical trial.

7.5 References

Alagoz, O., Ayer, T., and Erenay, F. S. (2010). Operations research models for cancer screening. *Wiley encyclopedia of operations research and management science*.

Andrinopoulou, E.-R., Rizopoulos, D., Takkenberg, J. J., and Lesaffre, E. (2017). Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. *Statistical methods in medical research*, 26(4):1787–1801.

Balasubramanian, R. and Lagakos, S. W. (2003). Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika*, 90(1):171–182.

Bokhorst, L. P., Alberts, A. R., Rannikko, A., Valdagni, R., Pickles, T., Kakehi, Y., Bangma, C. H., Roobol, M. J., and PRIAS study group (2015). Compliance rates with the Prostate Cancer Research International Active Surveillance (PRIAS) protocol and disease reclassification in noncompliers. *European Urology*, 68(5):814–821.

Choi, S. E. and Hur, C. (2012). Screening and surveillance for barrett's esophagus: current issues and future directions. *Current opinion in gastroenterology*, 28(4):377.

Coley, R. Y., Zeger, S. L., Mamawala, M., Pienta, K. J., and Carter, H. B. (2017). Prediction of the pathologic Gleason score to inform a personalized management program for prostate cancer. *European Urology*, 72(1):135–141.

Denton, B. T. (2018). Optimization of sequential decision making for chronic diseases: From data to decisions. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, pages 316–348. INFORMS.

Henderson, L., Nankivell, B., and Chapman*, J. (2011). Surveillance protocol kidney transplant biopsies: their evolving role in clinical practice. *American Journal of Transplantation*, 11(8):1570–1575.

Krist, A. H., Jones, R. M., Woolf, S. H., Woessner, S. E., Merenstein, D., Kerns, J. W., Foliaco, W., and Jackson, P. (2007). Timing of repeat colonoscopy: disparity between guidelines and endoscopists' recommendation. *American journal of preventive medicine*, 33(6):471–478.

Loeb, S., Vellekoop, A., Ahmed, H. U., Catto, J., Emberton, M., Nam, R., Rosario, D. J., Scattoni, V., and Lotan, Y. (2013). Systematic review of complications of prostate biopsy. *European Urology*, 64(6):876–892.

López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C., and Gude-Sampedro, F. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, 61(8):1–36.

McWilliams, T. J., Williams, T. J., Whitford, H. M., and Snell, G. I. (2008). Surveillance bronchoscopy in lung transplant recipients: risk versus benefit. *The Journal of Heart and Lung Transplantation*, 27(11):1203–1209.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press.

Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Rodriguez, A. (2008). Restful web services: The basics. *IBM developerWorks*, 33:18.

Sassi, F. (2006). Calculating qalys, comparing qaly and daly calculations. *Health policy and planning*, 21(5):402–408.

Steimle, L. N. and Denton, B. T. (2017). Markov decision processes for screening and treatment of chronic diseases. In *Markov Decision Processes in Practice*, pages 189–222. Springer.

Sunberg, Z. N. and Kochenderfer, M. J. (2018). Online algorithms for pomdps with continuous state, action, and observation spaces. In *Twenty-Eighth International Conference on Automated Planning and Scheduling*.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.

Vickers, A. J., Thompson, I. M., Klein, E., Carroll, P. R., and Scardino, P. T. (2014). A commentary on psa velocity and doubling time for clinical decisions in prostate cancer. *Urology*, 83(3):592–598.

English Summary, Nederlandse Samenvatting, PhD Portfolio, CV, and Acknowledgements

Summary

Low-grade chronic non-communicable disease (e.g., localized prostate cancer, low-grade dysplasia) patients often undergo repeated invasive *tests* (biopsies, endoscopies, etc.) for confirming disease *progression*. A progression is a non-terminal event, and upon progression, patients usually undergo serious treatments, e.g., surgery, radiotherapy. For detecting progression, invasive tests are typically conducted as per a one-size-fits-all (e.g., yearly) fixed schedule. Such invasive tests have two main features. First, they are indispensable because they are the benchmark/reference criteria for diagnosing progression. Second, they are burdensome for patients. Since invasive tests can only be conducted with a time gap between them, there is always a time delay in observing progression. In general, this delay is shorter when the tests are planned frequently. However, this simultaneously leads to an extra burden of biopsies on slow-progressing patients. The proportion of slow-progressing patients can be moderate to large in low-grade diseases that are the focus of our work. For the overall patient population, one-size-fits-all infrequent tests are also not a solution because they may lead to a large delay in patients who need early detection and care. Hence, our aim in this thesis was to balance better the number of tests (burden) and time delay in detecting progression (shorter is beneficial) in the whole patient population than one-size-fits-all schedules. To this end, we developed and applied statistical methods for scheduling invasive diagnostic tests (e.g., biopsies, endoscopies) in a personalized manner.

For creating personalized test schedules, our first step was to develop a statistical model. The purpose of this model was to use patients' accumulated clinical data to predict their cumulative-risk of progression over the whole follow-up period. A risk profile manifests the transition of a patient's disease state over time, from low-grade to progressed. Hence, subsequently, we used the risk profile to guide the timing of future invasive tests. That is, our second step was to obtain patient-specific test schedules using their estimated risk profiles. To this end, we optimized loss functions (e.g., squared loss) of certain clinical parameters of interest (e.g., time delay in detecting

progression). The various parameters we utilized are detailed in Chapters 2, 3, and 4. In each chapter, we optimized the loss functions with respect to the estimated patient-specific cumulative-risk of progression.

The choice of loss functions and parameters lead to different types of test schedules. For example, in Chapter 2, we chose three standard loss functions from Bayesian decision theory. They were, namely, squared loss, absolute loss, and multilinear loss. The parameter optimized via these loss functions was the time difference between the time of the future test and the true time of progression. Squared and absolute losses resulted in tests planned at a patient's estimated mean and median time of progression, respectively. Whereas, multilinear loss planned the future test at a time point where the patient's predicted risk of progression was equal to a certain threshold. Squared and absolute loss functions aim to plan a test exactly at the true time of progression so that progression is observed without any delay. However, in doing so, they ignore the variance of the posterior predictive distribution of the time of progression of a patient. Consequently, when squared/absolute loss functions are applied repeatedly until progression is detected, they may lead to very few tests but also a large time delay in detecting progression. On the other hand, planning a test when the risk of progression is equal to a certain threshold (multilinear loss), allows patients and doctors to weigh the unnecessary tests versus time delay in detecting progression. Particularly, choosing smaller risk thresholds means a patient is willing to undergo more tests but does not want the time delay in detecting progression to be high.

In a risk threshold based approach, the key question is how to choose an appropriate threshold? Typically, thresholds are chosen based on receiver operating characteristic curve analysis or on how patients weigh the burden of an unnecessary test against a large delay in detecting progression. To further facilitate decision making for an appropriate threshold, we conducted a realistic simulation randomized controlled trial with different risk thresholds for the prostate cancer active surveillance scenario in Chapter 3. While the results of this simulation study are only applicable for the study cohort (PRIAS prostate cancer surveillance) to which we fitted our dataset, certain results are generalizable across all diseases. The most important of the results is that thresholds should not be chosen using measures of diagnostic accuracy, such as Youden's J or F1 score. This is because they do not allow controlling sensitivity and specificity of a threshold.

In Chapter 2 and Chapter 3 we only planned one future test at a time. Later in Chapter 4, we extended our methodology to allow planning a full test schedule at once. We also calculated new measures of efficacy of schedules to assist patients and doctors in finding an optimal schedule. These measures were the expected number of tests and the expected time delay in detecting progression. We calculated these two in a personalized manner. Our choice of these criteria is motivated by two reasons. First, we argue that time delay in detection of progression is an easily-quantifiable surrogate for important clinical aspects such as the window of opportunity for curative treatment, risk of adverse downstream outcomes, quality-adjusted remaining lifetime, and additional complications in treating a delayed progression. Similarly, the number and timing of tests manifest financial costs of tests, risk of side-effects, and reduction in quality of life, etc. Second, both the number of tests and time delay in detecting progression are easy to understand for both patients/doctors and can facilitate *shared decision making* of test schedules.

A personalized schedule is only as good as the predictive performance of the underlying statistical model. In this regard, we externally validated the model we proposed for prostate cancer active surveillance. For validation, we employed the largest six cohorts of the Movember Foundations' Global Action Plan (GAP3) database (Chapter 5). We calculated the time-dependent mean absolute prediction error and time-dependent area under the receiver operating characteristic curve (AUC) in each cohort. The results indicated that our model had a moderate prediction error and moderate AUC. It is important to note that our patient population had a very low risk of metastases and mortality. Also, a PRIAS based simulation study has concluded that after the first biopsy, future biopsies leading to a time delay in detecting progression up to three years may lead to very limited adverse outcomes. Thus, even with a moderate predictive performance of the model, personalized schedules based on our model can be useful for our patients. Besides, patients can review the expected time delay in detecting progression for different schedules to limit their risks. To assist them in comparing multiple schedules side by side, we also implemented biopsy schedules for real patients of the validated cohorts in a web-application (`https://emcbiostatistics.shinyapps.io/prias_biopsy_recommender/`). The use of personalized schedules, however, is not limited to invasive tests only. We demonstrated this by planning the NT-proBNP biomarker (a simple blood test) measurements in chronic heart failure patients using an existing personalized methodology (Chapter 6).

Nederlandse Samenvatting

Patiënten met laaggradige chronische niet-overdraagbare ziektes (zoals gelokaliseerde prostaatkanker en laaggradige dysplasie) krijgen vaak herhaalde invasieve testen (biopsieën, endoscopieën, enz.) om de progressie van ziekte te kunnen detecteren. Progressie gaat vaak gepaard met zware behandelingen met ernstige bijwerkingen, zoals chirurgie of radiotherapie. Om progressie te kunnen vaststellen worden invasieve testen vaak routinematig uitgevoerd volgens een vast (bijv. jaarlijks) schema dat voor alle patiënten gelijk is. Deze invasieve testen zijn belastend, maar tegelijkertijd noodzakelijk voor het tijdig detecteren van ziekteprogressie. Het frequent uitvoeren van invasieve testen kan leiden tot een versnelde detectie van ziekteprogressie, maar het is onnodig en belastend voor patiënten die slechts langzaam achteruitgaan. Daarentegen hebben teststrategieën waarbij niet frequent wordt getest, het nadeel dat de ziekteprogressie vertraagd wordt gedetecteerd. Het doel van dit proefschrift was om met geïndividualiseerde schema's een beter evenwicht te vinden tussen het aantal testen enerzijds en de vertraging in de detectie van progressie anderzijds, dan mogelijk zou zijn met vaste schema's. Hiervoor hebben we statistische methoden ontwikkeld en toegepast met als doel om invasieve diagnostische testen (bijvoorbeeld biopsieën, endoscopieën) op een gepersonaliseerde manier te kunnen plannen.

De eerste stap in het creëren van geïndividualiseerde testschema's was het ontwikkelen van een statistisch model. Dit model hebben we gebruikt om op basis van de verzamelde klinische data van een patiënt, het cumulatieve risico op progressie over de gehele follow-up periode te voorspellen. Zo'n risicoprofiel voorspe lt de progressie van de ziekte van een patiënt over de tijd, van laaggradig naar gevorderd. Dit risicoprofiel gebruikten we vervolgens als leidraad om de timing van toekomstige invasieve testen te bepalen. De tweede stap was dus het creëren van een geïndividualiseerd testschema voor een patiënt op basis van zijn geschatte risicoprofiel. Voor dit doeleinde hebben we 'doelfuncties' (bijvoorbeeld een kwadratische functie) van relevante klinische parameters (zoals de vertraging in detectie van ziekteprogressie) geoptimaliseerd.. De parameters die we hiervoor hebben gebruikt staan beschreven in Hoofdstukken 2, 3 en 4. In elk hoofdstuk hebben we de doelfunctie geoptimaliseerd ten opzichte van het geschatte cumulatieve risico op progressie, geïndividualiseerd voor de patiënt.

De gekozen doelfunctie en parameters heeft invloed op het testschema. In Hoofdstuk 2 bijvoorbeeld, hebben we gebruik gemaakt van drie standaard doelfuncties uit de Bayesiaanse besliskunde, namelijk 'squared loss', 'abslote loss', en 'multilinear loss'. Op basis van deze doelfuncties hebben we het tijdsverschil tussen de toekomstige test en het echte moment van progressie geoptimaliseerd. Het gebruik van 'squared' en 'absolute loss' resulteerde in testen op respectievelijk het gemiddelde en het mediane tijdstip van progressie, maar 'multilinear loss' resulteerde in het tijdstip waarop het voorspelde risico van een patiënt een bepaalde drempelwaarde heeft bereikt. Het gebruik van 'squared' en 'abslote loss' resulteert in het testen op het precieze moment van progressie, en leidt dus tot geen vertraging in het detecteren hiervan. Echter, door gebruik te maken van deze functies wordt geen rekening gehouden met de variantie van de 'posterior predictive distribution' (posterior voorspelverdeling) van het tijdstip van progressie van een patiënt. Het herhaaldelijk toepassen van deze functies tot het moment van progressie, betekent dat minder testen uitgevoerd worden maar dit kan ook leiden tot een vertraging in het detecteren van progressie. Daarentegen biedt het plannen van een test op het moment dat het risico van progressie een bepaalde drempelwaarde heeft bereikt, artsen de mogelijkheid om een afweging te maken tussen het onnodig uitvoeren van testen en een vertraging in het detecteren van progressie. Het kiezen van een lage drempelwaarde betekent bijvoorbeeld dat een patiënt ervoor kiest om veel testen te ondergaan en zo weinig mogelijk risico te lopen dat progressie te laat wordt gedetecteerd.

Als gebruik wordt gemaakt van een dremelwaarde, is het de vraag hoe een geschikte drempelwaarde gekozen dient te worden. Meestal worden drempelwaardes gekozen op basis van 'receiver operating characteristic curve' (ROC analyse) of op basis van hoe voor patiënten de last van het uitvoeren van een onnodige testen opweegt tegen het te laat detecteren van progressie. Om de keuze van een geschikte drempelwaarde verder te vergemakkelijken, hebben we in Hoofdstuk 3, een realistische gerandomiseerde, gecontroleerde simulatiestudie uitgevoerd met verschillende drempelwaardes voor het actieve surveillancescenario voor prostaatkanker. Hoewel de resultaten van dit simulatieonderzoek alleen van toepassing zijn op het onderzoekscohort van onze dataset, zijn bepaalde resultaten generaliseerbaar naar alle ziektes. De belangrijkste uitkomst was dat drempelwaardes niet gekozen moeten worden op basis van maatstaven voor diagnostische nauwkeurigheid, zoals Youden's J of de F1 score, aangezien het voor dergelijke maatstaven niet mogelijk is de gevoeligheid en specificiteit van een drempelwaarde te controleren.

In Hoofdstuk 2 en Hoofdstuk 3 hebben we maar één toekomstige test tegelijk gepland. In Hoofdstuk 4 hebben we onze methode uitgebreid zodat een volledig testschema in één keer gepland kan worden. Om patiënten en artsen te assisteren bij het vinden van een optimaal schema, hebben we het verwachte aantal testen en de verwachte vertraging in het detecteren van progressie voor bepaalde testschema's berekend, beide geïndividualiseerd voor de specifieke patiënt. Onze beweegredenen voor de keuze voor deze criteria zijn als volgt. Ten eerste stellen we dat vertraging in het detecteren van progressie een makkelijk te kwantificeren surrogaatuitkomst is voor belangrijke klinische aspecten zoals de kans op curatieve behandeling, het risico op nadelige 'downstream' uitkomsten, de voor kwaliteit gecorrigeerde resterende levensduur en aanvullende complicaties bij de behandeling van een progressie die te laat wordt gedetecteerd. Ten tweede zijn de criteria (het aantal testen en de vertraging in het detecteren van progressie) voor zowel patiënten als artsen gemakkelijk te begrijpen, en ze kunnen een tussen arts en patiënt gedeelde besluitvorming van testschema's faciliteren.

Een geïndividualiseerd schema is slechts zo goed als de voorspellende waarde van het onderliggende statistische model. Om deze voorspellende waarde te bepalen, hebben we het model dat we hebben voorgesteld voor actieve surveillance van prostaatkanker extern gevalideerd in de grootste zes cohorten van de Movember Foundations Global Action Plan (GAP3) database (Hoofdstuk 5). We hebben voor elk cohort de tijdsafhankelijke gemiddelde absolute voorspellingsfout en de tijdsafhankelijke oppervlakte onder de receiver operating characteristic (AUC) curve berekend. Ons model had een redelijk lage voorspellingsfout en AUC. We hebben ook geïndividualiseerde biopsieschema's geïmplementeerd voor echte patiënten van de gevalideerde cohorten in een webapplicatie (`https://emcbiostatistics.shinyapps.io/prias_biopsy_recommender/`). Het gebruik van geïndividualiseerde schema's is echter niet beperkt tot invasieve testen. We hebben dit aangetoond door het plannen van metingen van de biomarker NT-proBNP (een eenvoudige bloedtest) voor patiënten met chronisch hartfalen op basis van een bestaande geïndividualiseerde methode (Hoofdstuk 6).

PhD Portfolio

PhD Candidate:	Anirudh Tomer
Department:	Department of Biostatistics Erasmus MC Rotterdam the Netherlands
PhD Period:	Sep 2016 – Aug 2020
Promoters:	Prof. dr. Dimitris Rizopoulos Prof. dr. Ewout W. Steyerberg

Publications

Tomer, A., Nieboer, D., Roobol, M.J., Steyerberg, E.W., and Rizopoulos, D. (2019), Personalized schedules for surveillance of low-risk prostate cancer patients. *Biometrics*, 75: 153–162.

Tomer, A., Rizopoulos, D., Nieboer, D., Drost, F.J., Roobol, M.J., and Steyerberg, E.W. (2019). Personalized decision making for biopsies in prostate cancer active surveillance programs. *Medical Decision Making*, 39(5): 499–508.

Tomer, A., Nieboer, D., Roobol, M.J., Steyerberg, E.W., and Rizopoulos, D. (2020), Personalized schedules for burdensome surveillance tests. Submitted to *Journal of the American Statistical Association*

Tomer, A., Nieboer, D., Roobol, M.J., Bjartell, A., Steyerberg, E.W., and Rizopoulos, D. (2020), Personalized Biopsy Schedules Based on Risk of Gleason Upgrading for Low-Risk Prostate Cancer Active Surveillance Patients. *BJU International*. Advance online publication. doi:10.1111/bju.15136

Schuurman, A.S., Tomer, A., Akkerhuis, K.M., Brugts, J.J., Constantinescu, A.A., van Ramshorst, J., Umans, V.A., Boersma, E., Rizopoulos, D., and Kardys, I. (2020). Personalized screening intervals for measurement of N-terminal pro-B-type natriuretic peptide improve efficiency of prognostication in patients with chronic heart failure. *European Journal of Preventive Cardiology*. Advance online publication. doi:10.1177/2047487320922639

Papageorgiou, G., Mauff, K., Tomer, A., and Rizopoulos, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual Review of Statistics and Its Application*, 6(1): 223–240.

Nieboer, D., Tomer, A., Rizopoulos, D., Roobol, M.J., and Steyerberg, E.W. (2018). Active surveillance: a review of risk-based, dynamic monitoring. *Translational Andrology and Urology*, 7(1), 106–115.

1. PhD Training	year	ECTS
International Conferences		
29th IBS Conference, Barcelona, Spain	2018	1.5
(*runner-up best oral presentation*)		
10th EMR-IBS Conference, Jerusalem, Israel	2018	1.0
(*student award for attending conference*)		
40th ISCB Conference, Leuven, Belgium	2019	1.0
35th EAU Conference, Netherlands	2020	1.0
Seminars and Workshops (the Netherlands)		
Fortnightly Biostatistics CQM Seminar	2016 – 2020	3.0
Yearly Erasmus Statistics Day	2016 – 2020	1.0
Erasmus MC – LUMC Joint Statistics Meeting	2016	0.2
BMS-ANed PhD Day	2017	0.2
BMS-ANed Spring Meeting	2018	0.2
VVSOR R Shiny Workshop (*presenter*)	2019	1.0
BMS-ANed PhD Day	2019	0.2
Courses		
Research Integrity	2017	0.3

2. Teaching and Work Experience	year	ECTS
Teaching Assistant		
SPSS Practicals (VO1)	2016 – 2020	1.0
Biostatistical Methods II (EP03)	2016 – 2020	1.0
Repeated Measurements (CE08)	2016 – 2020	1.0
CPO Consultant		
Erasmus MC CPO Statistical Consultant (*65 clients*)	2016 – 2020	10.0

CV

The author was born in Jorhat, India in February 1990. He obtained a BEng degree in Computer Engineering from the University of Pune in 2011. Subsequently, he worked as a software developer at TIBCO Software for three years. Due to a repetitive strain injury, in 2014, he switched careers from software programming to statistics. He graduated with an MSc in Statistics (magna cum laude) at KU Leuven, Belgium, in 2016. In September 2016 he commenced his PhD project under the supervision of Professor Dimitris Rizopoulos and Professor Ewout W. Steyerberg at Erasmus University Medical Center, the Netherlands.

Acknowledgements

This work was not possible without the blessings, well wishes, and the support of my family. With limited financial resources, my parents made enormous sacrifices to send me to university. It is because of their hard work and strength that I was able to study abroad, stay focused, and write this thesis. I am indebted to my father for his ever-present support for my decisions, to my mother for motivating me to maintain good health, and to my sister for always telling me that she believes in me. I married my lovely wife, Nikhita, in the second year of my PhD. Her love, care, support, and urging me to stop working at 5 pm, saved me from a possible PhD burnout.

I am grateful to my friends and soul brothers/sisters for listening to both the eureka and frustrating moments of my PhD. Outside work, we hiked, camped, made bonfires and barbecues, and cycled together. Special thanks to Kamen and his parents (Bulgaria), and Karin (Luxembourg) for making me a part of their families. Visiting them gave insights into European culture, and brought me closer to nature and delicious food. Being away from home, my friends turned a 'sometimes' mundane PhD routine into a fun life overall.

Isaac Newton said, *'If I have seen further it is by standing on the shoulders of Giants'*. I am no Newton, but I had the fortune of receiving support from two giants of statistics; my advisors Dimitris, and Ewout. I am indebted to them for allowing me to pursue research under their guidance. Dimitris was especially gentle towards my research mistakes. He provided me ample space and time throughout my PhD to understand the subject matter and grow into the researcher role. Truth be told, ideas in this work are only my minor extensions to his novel conceptions. Yet, he was kind enough to give me credit for the work. His humility, compassion, and friendliness towards his PhD students have inspired me to follow in his footsteps in the future. Without his support, I could not have written this thesis timely. I enjoyed statistical discussions with him because he evaluated our ideas based on merit. Dimitris consistently motivated me to leave no stone unturned and emphasized presenting our work in an easy to understand manner. I will carry these lessons for a lifetime.

I am also thankful to my co-advisor Ewout Steyerberg. In our meetings, he spoke with the eloquence of a spiritual guru on a range of topics. His critique on our papers was invariably to the point; his counterexamples exposed weaknesses of our methodologies; and his suggestions were often the hard truth, which is bitter and difficult to absorb, usually frustrating, and yet the right thing to do. Ewout inspired me to be concise in my scientific writing and taught me to prioritize the usefulness of an idea over its mathematical complexity. I am grateful to him for sharing his wisdom whenever we met.

I was lucky to often collaborate with Daan, a through and through seasoned statistician. I am grateful to him for assisting me formulate response letters for reviewers, prediction modeling, and expert advice on prostate cancer surveillance. My kind regards and best wishes to my colleagues from the Department of Urology. Although mostly we know each other only by face, they made me feel welcome whenever I visited them with questions.

I am grateful to all of my colleagues for sharing their knowledge during my PhD. Greg went above and beyond to help me even before I started. Elrozy and Joost always responded to my knocks on their office doors with a smile. They never even asked me to come later. Talking with them over lunch and tea lowered my work-related stress. I am indebted to both of them. Joost, thank you for providing useful advice for administrative matters related to my stay in the Netherlands. Special thanks to Paloma and Floor for being great friends. Paloma and Dino, your Bolivian dishes are delicious. Thank you for the evening walks and smiles. I am glad Nikki and I could share our cultures and cuisines with Floor and Paloma. Floor took extra efforts to make my wife and I feel welcome in the Netherlands. She is one of the kindest people I have met. I wish everyone gets a Dutch friend like her. Lastly, I would like to thank Eline for being a true friend in need. I do not have words to express my gratitude for the countless number of times she helped me in dire situations outside work. I may not be able to return her favors, but I promise to pay them forward to others.

In the end, I am grateful to the almighty for keeping me in good health, for giving me the means to see mountains up close, eat tasty food from around the world, and for showing miracles whenever I almost gave up.