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

Journal:	*BJU International*
Manuscript ID	BJU-2020-0444.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Tomer, Anirudh; Erasmus MC, Department of Biostatistics Nieboer, Daan; Erasmus MC, Department of Public Health; Erasmus MC, Department of Urology Roobol, Monique; Erasmus MC, Department of Urology Bjartell, Anders; Skanes universitetssjukhus Malmo, Department of Urology Steyerberg, Ewout; LUMC, Department of Biomedical Data Sciences; Erasmus MC, Department of Public Health Rizopoulos, Dimitris; Erasmus MC, Department of Biostatistics
Keywords:	Active Surveillance, Biopsies, Personalized Medicine, Prostate Cancer, Shared Decision Making
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 GAP3 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.

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

Anirudh Tomer, MSc[a,*], Daan Nieboer, MSc[b,c], Monique J. Roobol, PhD[c], Anders Bjartell, MD, PhD[d], Ewout W. Steyerberg, PhD[b,e], Dimitris Rizopoulos, PhD[a], Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium[f]

[a] *Department of Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands*
[b] *Department of Public Health, Erasmus University Medical Center, Rotterdam, the Netherlands*
[c] *Department of Urology, Erasmus University Medical Center, Rotterdam, the Netherlands*
[d] *Department of Urology, Skåne University Hospital, Malmö, Sweden*
[e] *Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands*
[f] *The Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium members presented in Appendix A*

∗Corresponding author (Anirudh Tomer): Erasmus MC, kamer flex Na-2823, PO Box 2040, 3000 CA Rotterdam, the Netherlands. Tel: +31 10 70 43393.

Email addresses: a.tomer@erasmusmc.nl (Anirudh Tomer, MSc), d.nieboer@erasmusmc.nl (Daan Nieboer, MSc), m.roobol@erasmusmc.nl (Monique J. Roobol, PhD), anders.bjartell@med.lu.se (Anders Bjartell, MD, PhD), e.w.steyerberg@lumc.nl (Ewout W. Steyerberg, PhD), d.rizopoulos@erasmusmc.nl (Dimitris Rizopoulos, PhD).

Word count: Main text 3636, Abstract 429.

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

Anirudh Tomer, MSc[a,*], Daan Nieboer, MSc[b,c], Monique J. Roobol, PhD[c], Anders Bjartell, MD, PhD[d], Ewout W. Steyerberg, PhD[b,e], Dimitris Rizopoulos, PhD[a], Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium[f]

[a] *Department of Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands*
[b] *Department of Public Health, Erasmus University Medical Center, Rotterdam, the Netherlands*
[c] *Department of Urology, Erasmus University Medical Center, Rotterdam, the Netherlands*
[d] *Department of Urology, Skåne University Hospital, Malmö, Sweden*
[e] *Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands*
[f] *The Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium members presented in Appendix A*

∗Corresponding author (Anirudh Tomer): Erasmus MC, kamer flex Na-2823, PO Box 2040, 3000 CA Rotterdam, the Netherlands. Tel: +31 10 70 43393.

Email addresses: a.tomer@erasmusmc.nl (Anirudh Tomer, MSc), d.nieboer@erasmusmc.nl (Daan Nieboer, MSc), m.roobol@erasmusmc.nl (Monique J. Roobol, PhD), anders.bjartell@med.lu.se (Anders Bjartell, MD, PhD), e.w.steyerberg@lumc.nl (Ewout W. Steyerberg, PhD), d.rizopoulos@erasmusmc.nl (Dimitris Rizopoulos, PhD).

Word count: Main text 3636, Abstract 429.

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 GAP3 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.

Keywords: Active Surveillance, Biopsies, Personalized Medicine, Prostate Cancer, Shared Decision Making.

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 [1]. 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 [2]. When it increases from group 1 (Gleason 3+3) to 2 (Gleason 3+4) or higher, it is called *upgrading* [3]. Upgrading is an important endpoint in AS upon which patients are commonly advised curative treatment [4].

Biopsies in AS are always conducted with a time gap between them. Consequently, upgrading is always detected with a time delay (Figure 1) that

cannot be measured directly. In this regard, to detect upgrading timely, many patients are prescribed fixed and frequent biopsies, most often annually [5]. 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 [6]. Thus, biopsy burden and patient non-compliance to frequent biopsies [7] have raised concerns regarding the optimal biopsy schedule [8, 9] in AS.

Except for the confirmatory biopsy at year one of AS [7], opinions and practice regarding the timing of remaining biopsies lack agreement [10]. Some AS programs utilize patients' observed PSA, DRE, previous biopsy Gleason grade, and lately, MRI results to decide biopsies [11, 4, 10]. In contrast, others discourage schedules based on clinical data and MRI results [12, 5], and instead support periodical one-size-fits-all biopsy schedules. Furthermore, some suggest replacing frequent periodical schedules with infrequent ones (e.g., biennially) [8, 13]. 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 [8]. 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 [10] (Figure 2). While, upgrading-risk calculators are not new [14, 15, 16, 17], 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 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 PSA measurements of patients, previous biopsy results, and baseline characteristics. 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 Movember Foundation's Global Action Plan (GAP3) database [18]. Last, we aim to implement the validated model and methodology in a web-application.

2. Patients and Methods

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 [4], dated April 2019 (Table 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 [19]. 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 [10].

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 [18] version 3.1, covering nearly 73% of the GAP3 patients: the University of Toronto AS (Toronto), Johns Hopkins AS (Hopkins), Memorial Sloan Kettering Cancer 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 Supplementary A.2.

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 [11, 12, 20]. 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.

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 [21, 14, 22].

Our joint model consisted of two sub-models. Namely, a linear mixed-effects sub-model [23] 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 [24], 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 3).

Subsequently, we calculated the mathematical derivative of the patient's fitted PSA profile (Equation 2, Supplementary A), to obtain his follow-up time specific instantaneous PSA velocity (Panel B, Figure 3). This instantaneous velocity is a stronger predictor of upgrading than the widely used average PSA velocity [25].

Table 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 [2] 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)

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 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 Supplementary A). This approach also

accommodates watchful waiting and treatment protocols that are also based on patient data. Specifically, the parameters of our two sub-models were estimated jointly under the Bayesian paradigm (Supplementary A) using the R package **JMbayes** [26].

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 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 risk predictions for a patient are not calculated only once. Rather, as illustrated in Figure 5 of Supplementary B, risk-predictions update over the follow-up, to account for additional patient data (e.g., new biopsy results, PSA measurements) that becomes available. We validated our model internally in the PRIAS cohort, and externally in the largest six GAP3 database cohorts. We employed calibration plots [27, 28] and follow-up *time-dependent* mean absolute risk prediction error or MAPE [29] to graphically and quantitatively evaluate 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 [29].

The aforementioned *time-dependent* AUC and MAPE [29] are temporal extensions of their standard versions [28] 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 (Supplementary B.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. For example, calculations for AUC and MAPE for the time interval year two to year three do not utilize data of patients who progressed

before year two. Last, to resolve any potential model miscalibration in validation cohorts, we aimed to recalibrate our model's baseline hazard of upgrading (Supplementary B.1), individually for each cohort.

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 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 (Supplementary Table 8).

The time-dependent AUC, calibration plot, and time-dependent MAPE of our model are shown in Figure 4, and Supplementary Figure 8. 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, Supplementary Table 8), and large (0.2 to 0.3) otherwise. Our model was miscalibrated for validation cohorts (Panel B, Figure 4) because cohorts had differences in inclusion criteria (e.g., PSA density) and follow-up protocols [18] 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 (Supplementary Figure 6). We compared risk predictions from the recalibrated models, with predictions from separately fitted cohort-specific

joint models (Supplementary Figure 7). The difference in predictions was lowest in the Johns Hopkins cohort (impact of PSA on upgrading-risk similar to PRIAS). Comprehensive results are in Supplementary A.3 and Supplementary B.

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) 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 Supplementary C 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, and in Supplementary Figure 10–12.

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, 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 all available clinical data of the patient (see Supplementary C). 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.

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. Users must first choose the cohort to which the patient belongs (left panel), and then upload patient data in Microsoft Excel format. Internally, the web-application loads the appropriate validated and recalibrated model for that cohort. The maximum follow-up time up to which predictions can be obtained depends on each cohort (Supplementary Table 9). 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.

4. Discussion

We successfully developed and externally validated a statistical model for predicting upgrading-risk [3] in prostate cancer AS, and providing risk-based personalized biopsy decisions. Our work has four novel features over earlier risk calculators [14, 15]. 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 [18]. 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 following 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. A benefit of our model is that it allows the biopsy schedule, schedule of longitudinal measurements, and loss to follow-up in each cohort to depend on patient age and PSA characteristics. Consequently, in future, when MRI data is included in the model, our model will also accommodate biopsy schedules dependent on MRI data or MRI schedules dependent on previous biopsy results, PSA characteristics, and age of the patient (mathematical proof in Appendix A.2). An additional advantage of our model and resulting personalized schedules is that they update as more patient data becomes available over follow-up. 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 [28] 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 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 [7], the availability of our methodology as a web-application may encourage patients/doctors to consider upgrading-risk based personalized schedules instead. 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 personalized schedules is quite low because of the low rate of metastases and prostate cancer specific mortality in AS patients (Table 1). Second, studies [31, 8] 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 [32]. 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 (Supplementary Table 9). 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. In this regard, the current model can be extended to include both MRI and PSA density data as predictors when they become available in the future. Our model schedules biopsies in a personalized manner, but the patient burden can be decreased even more if we also personalize the schedule of PSA measurements. A caveat of doing so is that reduction in the number of PSA measurements can also lead to an increase in the variance of risk estimates, and also affect the performance of personalized schedules. Although we have done a simulation study to conclude that personalized schedules may not be any more harmful than PRIAS or annual schedule [32], with an infrequent PSA schedule, these conclusions may not hold. Hence, we do not recommend any changes in the schedule of PSA measurements from the current protocol of PSA measurements every six months. At the same time, personalizing the schedule of both biopsies and PSA measurements together is a research problem we intend to tackle in the near 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 [14, 33] will be interesting to investigate further. Even with an enhanced risk prediction model, the methodology for

personalized scheduling and calculation of expected time delay (Supplementary C) 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. 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.

Author Contributions

Anirudh Tomer had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Tomer, Nieboer, Roobol, Bjartell, and Rizopoulos

Acquisition of data: Tomer, Nieboer, and Roobol

Analysis and interpretation of data: Tomer, Nieboer, and Rizopoulos

Drafting of the manuscript: Tomer, and Rizopoulos

Critical revision of the manuscript for important intellectual content: Tomer, Nieboer, Roobol, Bjartell, Steyerberg, and Rizopoulos

Statistical analyses: Tomer, Nieboer, Steyerberg, and Rizopoulos

Obtaining funding: Roobol, Steyerberg, and Rizopoulos

Administrative, technical or material support: Nieboer

Supervision: Roobol, and Rizopoulos

Other: none

Acknowledgments

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 einfrastructure 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.

Appendix A. 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, Universita` 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, Go¨teborg, Sweden), Jose Rubio-Briones (Instituto Valenciano de Oncologia, 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 Health-Care, 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), Hel´en Ahlgren (Sahlgrenska University Hospital, Go¨teborg, Sweden), Juanma Mascaro´s (Instituto Valenciano de Oncolog´ıa, 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 (Sk˚ane University Hospital, Malmo¨, 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).

References

[1] Briganti A, Fossati N, Catto JW, Cornford P, Montorsi F, Mottet N, et al. Active surveillance for low-risk prostate cancer: the European Association of Urology position in 2018. European urology 2018;74(3):357– 68.

[2] Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. The American journal of surgical pathology 2016;40(2):244–52.

[3] Bruinsma SM, Roobol MJ, Carroll PR, Klotz L, Pickles T, Moore CM, et al. Expert consensus document: semantics in active surveillance for men with localized prostate cancer—results of a modified Delphi consensus procedure. Nature reviews urology 2017;14(5):312.

[4] Bul M, Zhu X, Valdagni R, Pickles T, Kakehi Y, Rannikko A, et al. Active surveillance for low-risk prostate cancer worldwide: the PRIAS study. European urology 2013;63(4):597–603.

[5] Loeb S, Carter HB, Schwartz M, Fagerlin A, Braithwaite RS, Lepor H. Heterogeneity in active surveillance protocols worldwide. Reviews in urology 2014;16(4):202–3.

[6] Loeb S, Vellekoop A, Ahmed HU, Catto J, Emberton M, Nam R, et al. Systematic review of complications of prostate biopsy. European urology 2013;64(6):876–92.

[7] Bokhorst LP, Alberts AR, Rannikko A, Valdagni R, Pickles T, Kakehi Y, et al. Compliance rates with the Prostate Cancer Research International Active Surveillance (PRIAS) protocol and disease reclassification in noncompliers. European Urology 2015;68(5):814–21.

[8] Inoue LY, Lin DW, Newcomb LF, Leonardson AS, Ankerst D, Gulati R, et al. Comparative analysis of biopsy upgrading in four prostate cancer active surveillance cohorts. Annals of internal medicine 2018;168(1):1–9.

[9] Bratt O, Carlsson S, Holmberg E, Holmberg L, Johansson E, Josefsson A, et al. 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
2013;47(5):347–55.

[10] Nieboer D, Tomer A, Rizopoulos D, Roobol MJ, Steyerberg EW. Active surveillance: a review of risk-based, dynamic monitoring. Translational andrology and urology 2018;7(1):106–15.

[11] Kasivisvanathan V, Giganti F, Emberton M, Moore CM. Magnetic resonance imaging should be used in the active surveillance of patients with localised prostate cancer. European urology 2020;77(3):318–9.

[12] Chesnut GT, Vertosick EA, Benfante N, Sjoberg DD, Fainberg J, Lee T, et al. 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 2020;77(4):501–7.

[13] de Carvalho TM, Heijnsdijk EA, de Koning HJ. Estimating the risks and benefits of active surveillance protocols for prostate cancer: a microsimulation study. BJU international 2017;119(4):560–6.

[14] Coley RY, Zeger SL, Mamawala M, Pienta KJ, Carter HB. Prediction of the pathologic Gleason score to inform a personalized management program for prostate cancer. European urology 2017;72(1):135–41.

[15] Ankerst DP, Xia J, Thompson Jr IM, Hoefler J, Newcomb LF, Brooks JD, et al. Precision medicine in active surveillance for prostate cancer: development of the canary–early detection research network active surveillance biopsy risk calculator. European urology 2015;68(6):1083–8.

[16] Partin AW, Yoo J, Carter HB, Pearson JD, Chan DW, Epstein JI, et al.
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 1993;150(1):110–4.

[17] Makarov DV, Trock BJ, Humphreys EB, Mangold LA, Walsh PC, Epstein JI, et al. 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 2007;69(6):1095–101.

[18] Bruinsma SM, Zhang L, Roobol MJ, Bangma CH, Steyerberg EW, Nieboer D, et al. The Movember foundation's GAP3 cohort: a profile of the largest global prostate cancer active surveillance database to date. BJU international 2018;121(5):737–44.

[19] Mottet N, Bellmunt J, Bolla M, Briers E, Cumberbatch MG, De Santis M, et al. Eau-estro-siog guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent. European Urology 2017;71(4):618–29.

[20] Schoots IG, Petrides N, Giganti F, Bokhorst LP, Rannikko A, Klotz L, et al. Magnetic resonance imaging in active surveillance of prostate cancer: a systematic review. European urology 2015;67(4):627–36.

[21] Tomer A, Nieboer D, Roobol MJ, Steyerberg EW, Rizopoulos D. Personalized schedules for surveillance of low-risk prostate cancer patients. Biometrics 2019;75(1):153–62.

[22] Rizopoulos D. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. CRC Press; 2012. ISBN 9781439872864.

[23] Laird NM, Ware JH, et al. Random-effects models for longitudinal data. Biometrics 1982;38(4):963–74.

[24] Nixon RG, Wener MH, Smith KM, Parson RE, Strobel SA, Brawer MK. 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 1997;157(6):2183–90.

[25] Cooperberg MR, Brooks JD, Faino AV, Newcomb LF, Kearns JT, Carroll PR, et al. Refined analysis of prostate-specific antigen kinetics to predict prostate cancer active surveillance outcomes. European urology 2018;74(2):211–7.

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

[27] Royston P, Altman DG. External validation of a cox prognostic model: principles and methods. BMC medical research methodology 2013;13(1):33.

[28] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass) 2010;21(1):128.

[29] Rizopoulos D, Molenberghs G, Lesaffre EM. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biometrical Journal 2017;59(6):1261–76.

[30] Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society Series B (Methodological) 1976;38(3):290–5.

[31] de Carvalho TM, Heijnsdijk EA, de Koning HJ. When should active surveillance for prostate cancer stop if no progression is detected? The Prostate 2017;77(9):962–9.

[32] Tomer A, Rizopoulos D, Nieboer D, Drost FJ, Roobol MJ, Steyerberg EW. Personalized decision making for biopsies in prostate cancer active surveillance programs. Medical Decision Making 2019;39(5):499–508.

[33] Balasubramanian R, Lagakos SW. Estimation of a failure time distribution based on imperfect diagnostic tests. Biometrika 2003;90(1):171–82.

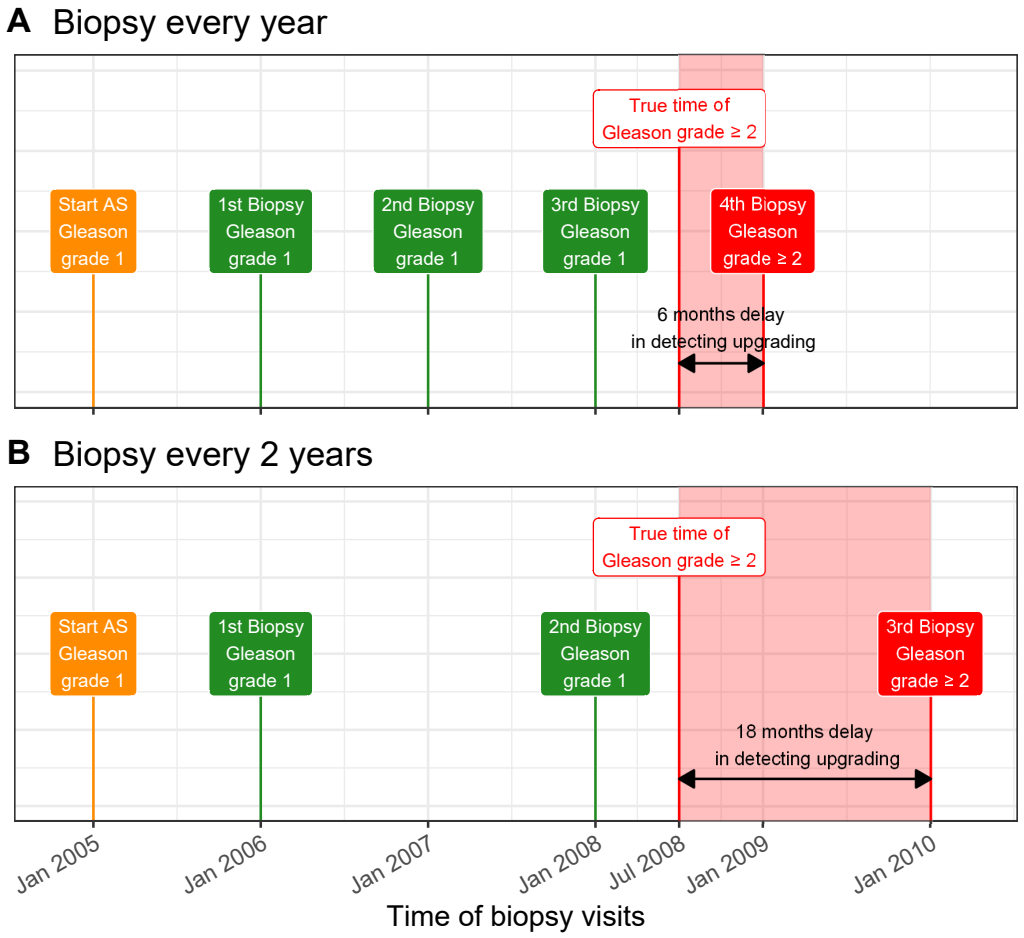

Figure 1: **Trade-off between the timing and number of biopsies (burden) and time delay in detecting Gleason upgrading (shorter is better):** The true time of Gleason upgrading (increase in Gleason grade group from group 1 to 2 or higher) for the patient in this figure is July 2008. When biopsies are 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.

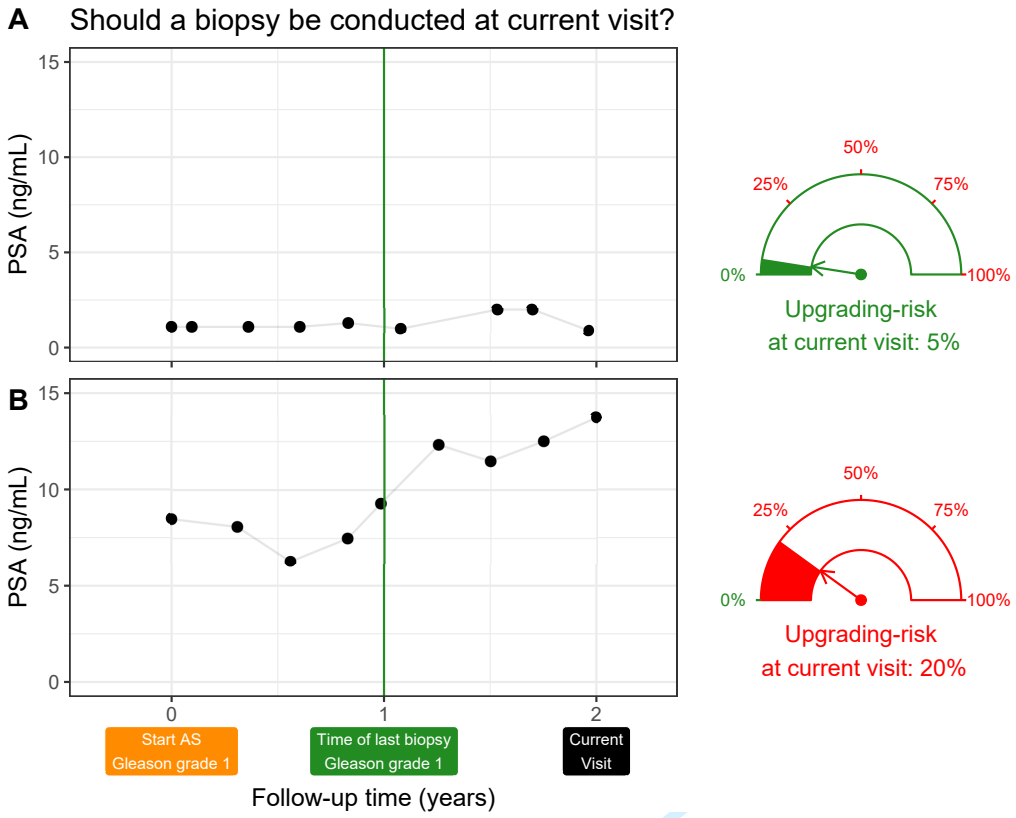

Figure 2: **Motivation for upgrading-risk based personalized biopsy decisions**: To utilize patients' complete longitudinal data and results from previous biopsies in making biopsy decisions. For this purpose, we first process data using a statistical model and then utilize the patient-specific predictions 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.

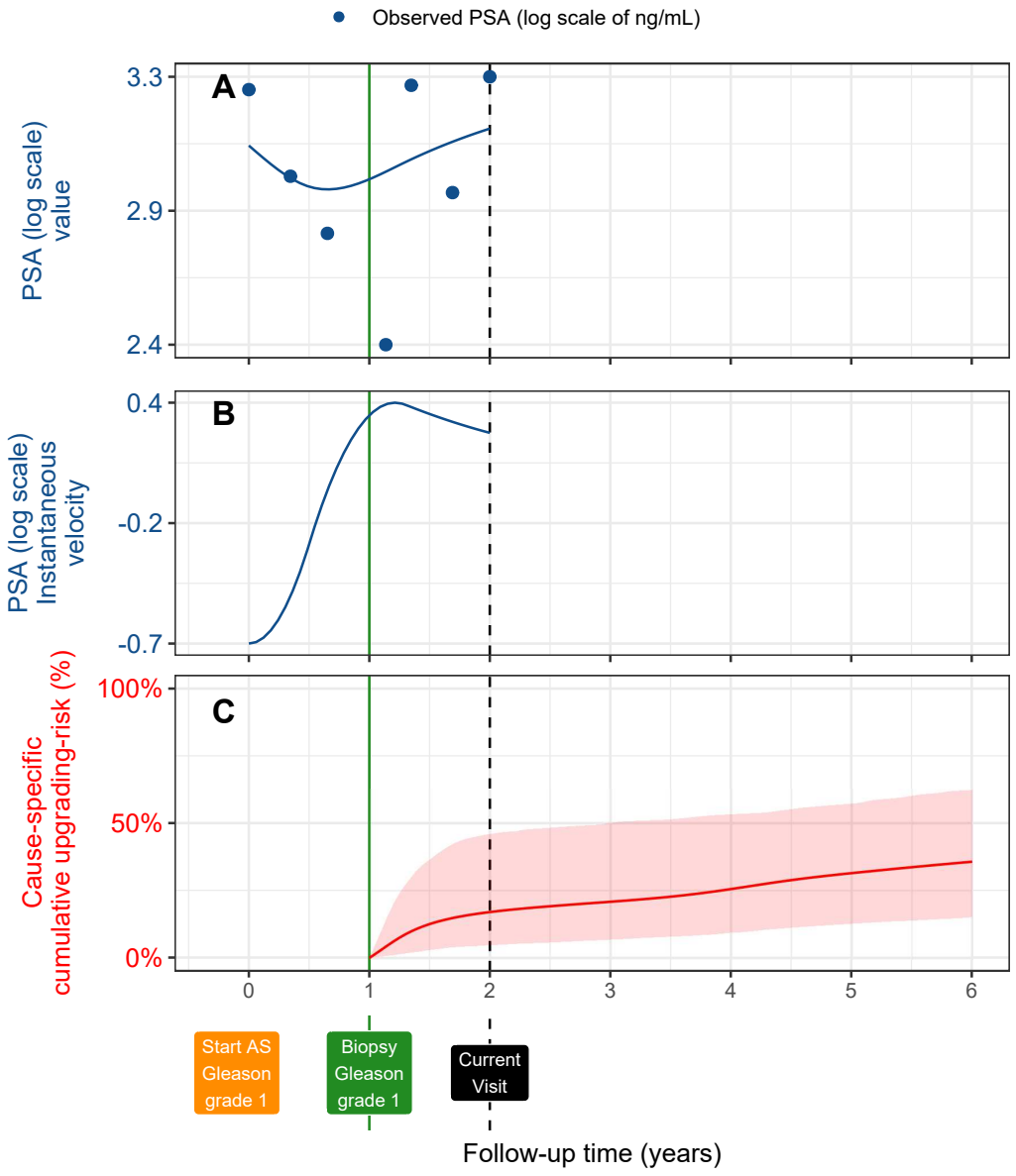

Figure 3: **Illustration of the joint model on a real PRIAS patient**. **Panel A:** Observed PSA (blue dots) and fitted PSA (solid blue line), log-transformed from ng/mL. **Panel B:** Estimated instantaneous velocity of PSA (log-transformed). **Panel C**: Predicted cause-specific cumulative upgrading-risk (95% credible interval shaded). Upgrading is defined as an increase in the Gleason grade group from group 1 [2] 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.

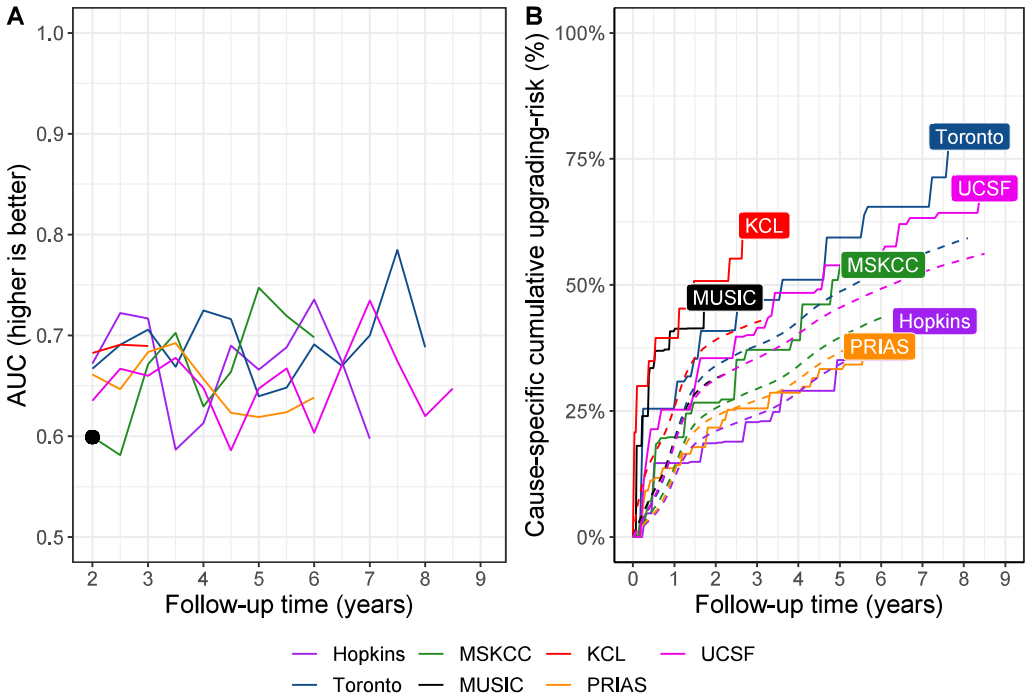

Figure 4: **Model Validation Results**. **Panel A**: time-dependent area under the receiver operating characteristic curve or AUC (measure of discrimination). AUC at year one is not shown because we do not intend to replace the confirmatory biopsy at year one. **Panel B**: calibration-at-large indicates model miscalibration. This is because solid lines depicting the non-parameteric estimate of the cause-specific cumulative upgrading-risk [30], 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 (Supplementary Figure 6). 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.

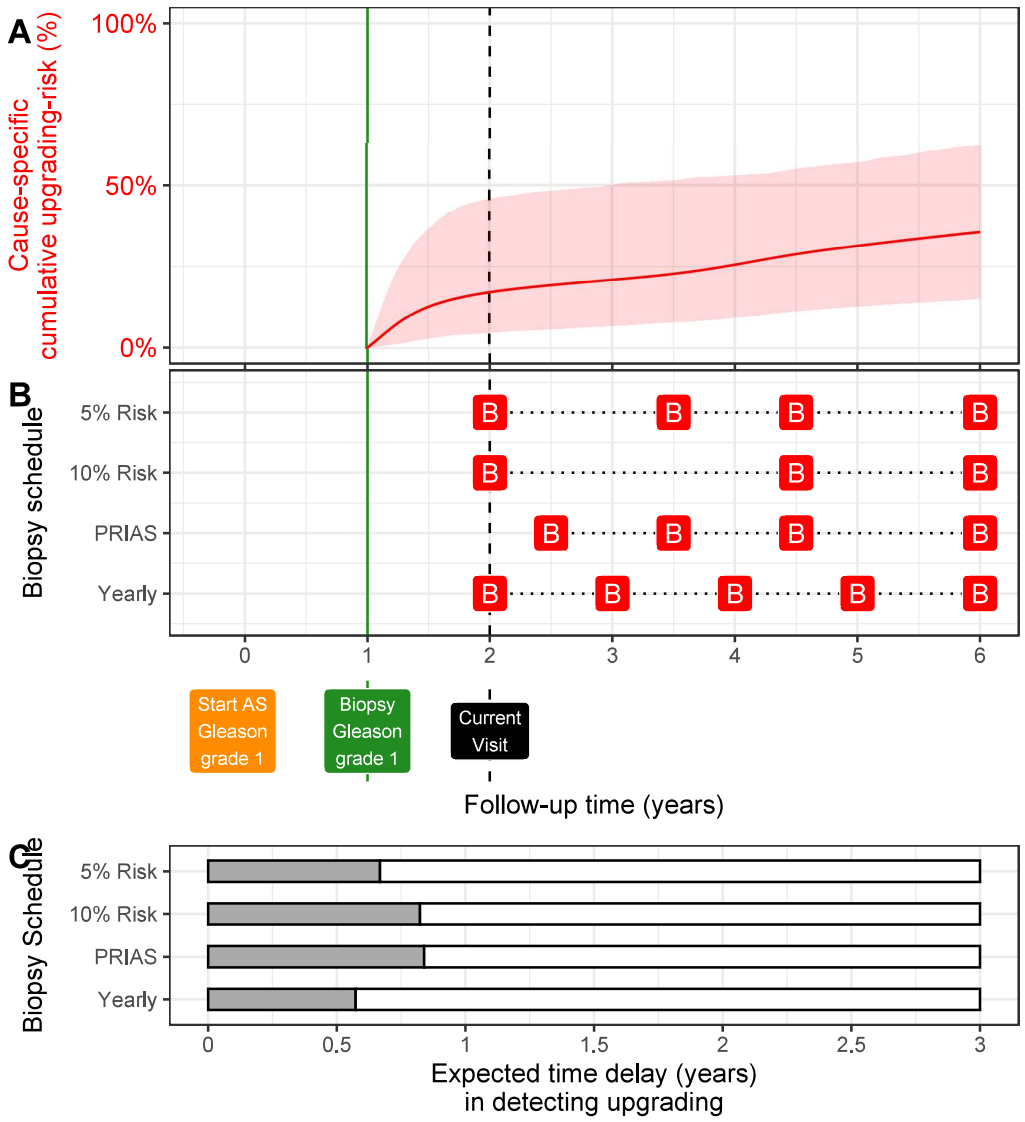

Figure 5: **Illustration of personalized and fixed schedules of biopsies for patient from Figure 3**. **Panel A:** Predicted cumulative upgrading-risk (95% credible interval shaded). **Panel B:** Different biopsy schedules with a red 'B' indicating a future biopsy. Risk: 5% and Risk: 10% are personalized schedules in which a biopsy is planned whenever the conditional cause-specific cumulative upgrading-risk is above 5% or 10% risk, respectively. Smaller risk thresholds lead to more frequently planned biopsies. 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. The maximum time delay with limited adverse consequences is three years [13]. A compulsory biopsy was scheduled at year six (maximum biopsy scheduling time in PRIAS, Supplementary C) in all schedules for a meaningful comparison between delays.

Supplementary Materials for "Risk of Upgrading Based Personalized Biopsy Schedules for Prostate Cancer Active Surveillance Patients"

Anirudh Tomer, MSc[a,*], Daan Nieboer, MSc[b], Monique J. Roobol, PhD[c], Anders Bjartell, MD, PhD[d], Ewout W. Steyerberg, PhD[b,e], Dimitris Rizopoulos, PhD[a], Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium[f]

[a] *Department of Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands*
[b] *Department of Public Health, Erasmus University Medical Center, Rotterdam, the Netherlands*
[c] *Department of Urology, Erasmus University Medical Center, Rotterdam, the Netherlands*
[d] *Department of Urology, Skåne University Hospital, Malmö, Sweden*
[e] *Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands*
[f] *The Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium members presented in Appendix F*

Appendix A. A Joint Model for the Longitudinal PSA, and Time to Gleason Upgrading

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

*Corresponding author (Anirudh Tomer): Erasmus MC, kamer flex Na-2823, PO Box 2040, 3000 CA Rotterdam, the Netherlands. Tel: +31 10 70 43393

Email addresses: a.tomer@erasmusmc.nl (Anirudh Tomer, MSc), d.nieboer@erasmusmc.nl (Daan Nieboer, MSc), m.roobol@erasmusmc.nl (Monique J. Roobol, PhD), anders.bjartell@med.lu.se (Anders Bjartell, MD, PhD), e.w.steyerberg@lumc.nl (Ewout W. Steyerberg, PhD), d.rizopoulos@erasmusmc.nl (Dimitris Rizopoulos, PhD)

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 1):

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

where, $m_i(t)$ denotes the measurement error free value of $\log_2(\text{PSA}+1)$ transformed [2, 3] measurements at time t. We model it non-linearly over time using B-splines [4]. 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 (see Appendix B.1) 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 1):

$$\begin{aligned} h_i(t) &= h_0(t)\exp\Big(\gamma \text{age}_i + \alpha_1 m_i(t) + \alpha_2 \frac{\mathrm{d}m_i(t)}{\mathrm{d}t}\Big), \\ R_i(t) &= \exp\Big\{-\int_0^t h_i(s)\mathrm{d}s\Big\}, \end{aligned} \tag{2}$$

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

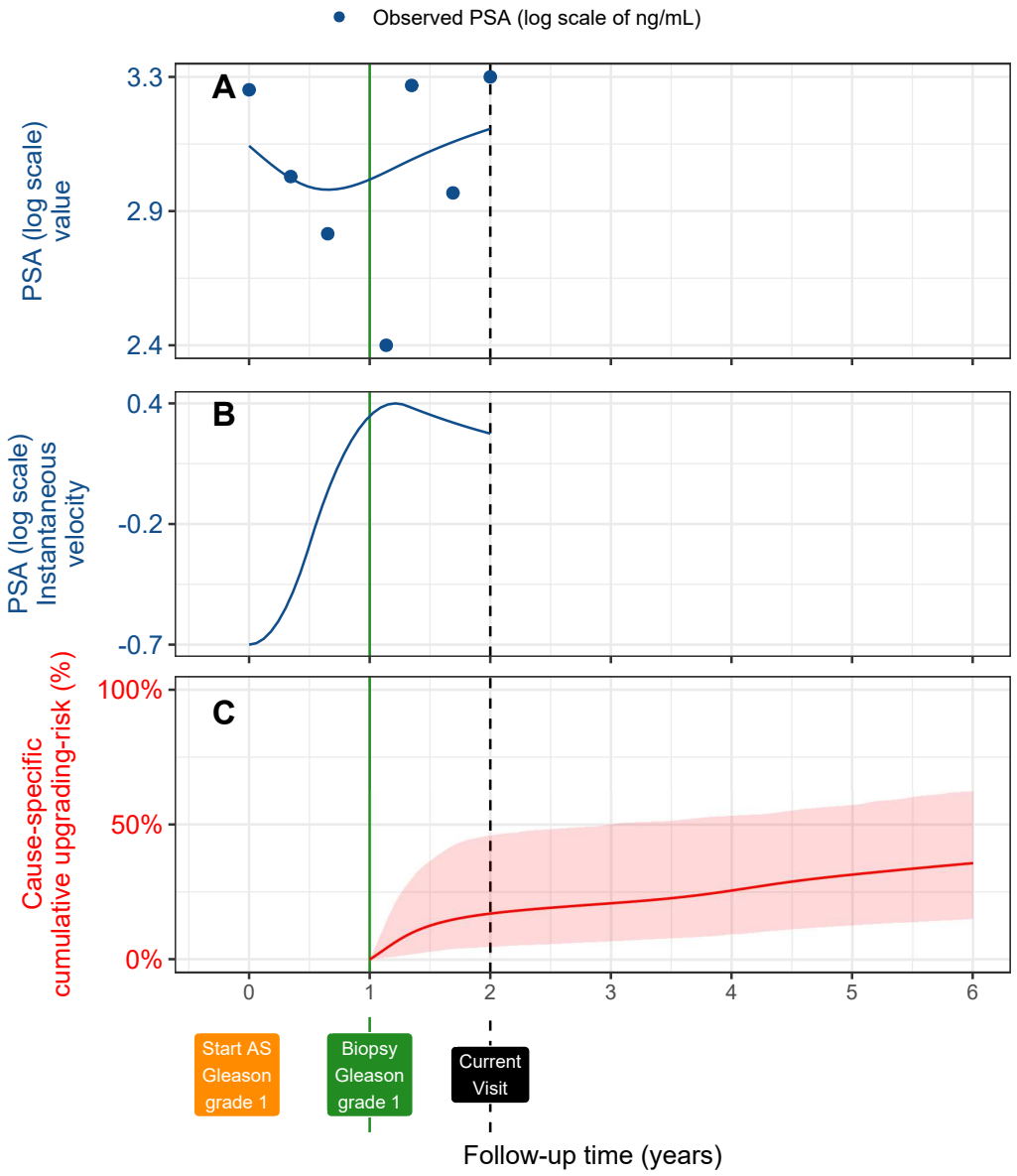

Figure 1: **Illustration of the joint model on a real PRIAS dataset patient**. **Panel A:** Observed (blue dots) and fitted PSA (solid blue line) measurements, log-transformed. **Panel B:** Estimated instantaneous velocity of PSA (log-transformed). **Panel C**: Predicted cumulative-risk of upgrading (95% credible interval shaded). Upgrading is defined as an increase in Gleason grade group [1] from grade group 1 to 2 or higher. This risk of upgrading is available starting from the time of the latest negative biopsy (vertical green line at year 1 of follow-up). The joint model estimated it by combining the fitted PSA value and velocity (both on the log scale of PSA) and time of the latest negative biopsy. Black dashed line at year 4 denotes the time of current visit.

free underlying PSA value $m_i(t)$ (see Panel A, Figure 1), and the impact of the underlying PSA velocity $\mathrm{d}m_i(t)/\mathrm{d}t$ (see Panel B, Figure 1). 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 [5]. 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 [5].

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 of upgrading, and the PSA measurements taken over time are all mutually independent. Under this assumption the posterior distribution of the parameters is given by:

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

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

$$p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^{n_i}} \exp\Big\{ -\frac{\sum_{j=1}^{n_i} (y_{ij} - m_{ij})^2}{2\sigma^2} \Big\},$$

where n_i is the number of PSA measurements of the i-th patient. The likelihood contribution of the time of upgrading outcome is given by:

$$p(l_i, r_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \exp\Big\{ -\int_0^{l_i} h_i(s)\mathrm{d}s \Big\} - \exp\Big\{ -\int_0^{r_i} h_i(s)\mathrm{d}s \Big\}. \qquad (3)$$

The integrals in (3) do not have a closed-form solution, and therefore we use a 15-point Gauss-Kronrod quadrature rule to approximate them.

We use independent normal priors with zero mean and variance 100 for the fixed effects $\{\beta_0, \ldots, \beta_5\}$, and inverse Gamma prior with shape and rate both equal to 0.01 for the parameter σ^2. For the variance-covariance matrix $\boldsymbol{W}$ of the random effects, we take inverse Wishart prior with an identity scale matrix and degrees of freedom equal to 5 (number of random effects). For the relative risk model's parameter γ and the association parameters α_1, α_2, we use independent normal priors with zero mean and variance 100.

Appendix A.1. Assumption of t-distributed (df=3) Error Terms

With regards to the choice of the distribution for the error term ε for the PSA measurements (see Equation 1), we attempted fitting multiple joint models differing in error distribution, namely t-distribution with three, and four degrees of freedom, and a normal distribution for the error term. However, the model assumption for the error term was best met by the model with t-distribution having three degrees of freedom. The quantile-quantile plot of subject-specific residuals for the corresponding model in Panel A of Figure 2, shows that the assumption of t-distributed (df=3) errors is reasonably met by the fitted model.

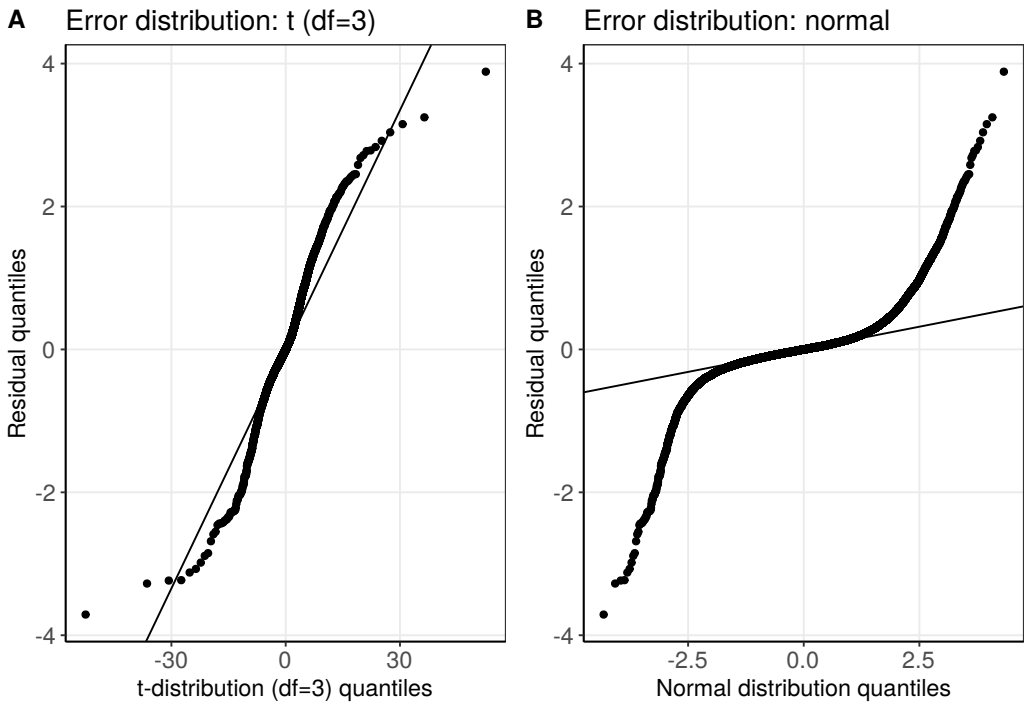

Figure 2: **Quantile-quantile plot** of subject-specific PSA residuals from two different joint models fitted to the PRIAS dataset. **Panel A**: model assuming a t-distribution (df=3) for the error term ε (see Equation 1). **Panel B**: model assuming a normal distribution for the error term ε. We selected the model with t-distributed error terms.

Appendix A.2. PSA Dependent Biopsy Schedule of PRIAS, and Competing Risks

PSA dependent interval censored time of upgrading: The true time of upgrading T_i^* is not known for any of the patients in PRIAS. In order 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 thereafter. 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 loss to follow-up as non-informative censoring, which may not always be true. This is especially the case when the reason of loss to follow-up is unknown. However, when the reason of 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 unknown reason of dropout is not specific to only our model. For the remaining patients who are removed from AS on the basis 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 [6], 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 [7], despite the interval censoring observed.

Appendix A.3. Results

Characteristics of the six validation cohorts from the GAP3 database [8] are shown in Table 1, Table 2, and Table 3. The cause-specific cumulative upgrading-risk in these cohorts is shown in Figure 3.

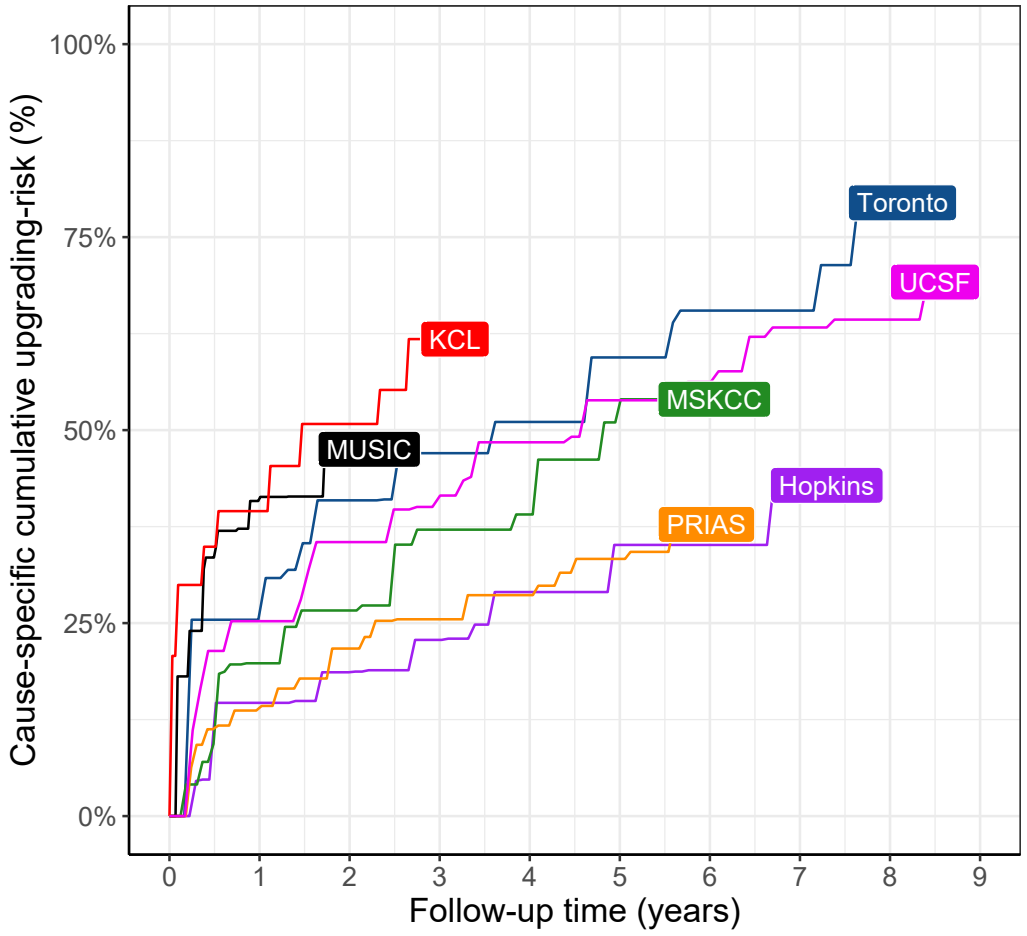

Figure 3: **Nonparametric estimate [9] of the cause-specific cumulative upgrading-risk** in the world's largest AS cohort PRIAS, and largest six AS cohorts from the GAP3 database [8]. Abbreviations are *Hopkins*: Johns Hopkins Active Surveillance, *PRIAS*: Prostate Cancer International Active Surveillance, *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 1: **Summary of the Hopkins and Toronto validation cohorts from the GAP3 database [8]**. 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)

Table 2: **Summary of the MSKCC and UCSF validation cohorts from the GAP3 database [8]**. 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)

Table 3: **Summary of the MUSIC and KCL validation cohorts from the GAP3 database [8]**. 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 4: **Estimated variance-covariance matrix $\boldsymbol{W}$** of the random effects $\boldsymbol{b} = (b_0, b_1, b_2, b_3, b_4)$ from the joint model fitted to the PRIAS dataset. The variances of the random effects are highlighted along the diagonal of the variance-covariance matrix.

Random Effects	b_0	b_1	b_2	b_3	b_4
b_0	**0.229**	0.030	0.023	0.073	0.007
b_1	0.030	**0.149**	0.098	0.171	0.085
b_2	0.023	0.098	**0.276**	0.335	0.236
b_3	0.073	0.171	0.335	**0.560**	0.359
b_4	0.007	0.085	0.236	0.359	**0.351**

The joint model was fitted using the R package **JMbayes** [10]. This package utilizes the Bayesian methodology to estimate model parameters. The corresponding posterior parameter estimates are shown in Table 5 (longitudinal sub-model for PSA outcome) and Table 6 (relative risk sub-model). The parameter estimates for the variance-covariance matrix $\boldsymbol{W}$ from the longitudinal sub-model for PSA are shown in the following Table 4:

For the PSA mixed effects sub-model parameter estimates (see Equation 1), in Table 5 we can see that the age of the patient trivially affects the baseline $\log_2(\text{PSA} + 1)$ measurement. Since the longitudinal evolution of $\log_2(\text{PSA} + 1)$ measurements is modeled with non-linear terms, the interpretation of the coefficients corresponding to time is not straightforward. In lieu of the interpretation, in Figure 4 we present plots of observed versus fitted

Table 5: **Parameters of the longitudinal sub-model**: Estimated mean and 95% credible interval for parameters in Equation (1).

Variable	Mean	Std. Dev	2.5%	97.5%	P
Intercept	2.129	0.060	2.009	2.244	<0.001
Age	0.008	0.001	0.007	0.010	<0.001
Spline: [0.0, 0.5] years	0.063	0.007	0.051	0.075	<0.001
Spline: [0.5, 1.3] years	0.196	0.010	0.177	0.217	<0.001
Spline: [1.3, 3.0] years	0.244	0.014	0.217	0.272	<0.001
Spline: [3.0, 6.3] years	0.382	0.014	0.356	0.410	<0.001
σ	0.139	0.001	0.138	0.140	

Table 6: **Parameters of the relative risk sub-model**: Estimated mean and 95% credible interval for the parameters in Equation (2).

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

PSA profiles for nine randomly selected patients.

For the relative risk sub-model (see Equation 2), the parameter estimates in Table 6 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 7, 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.

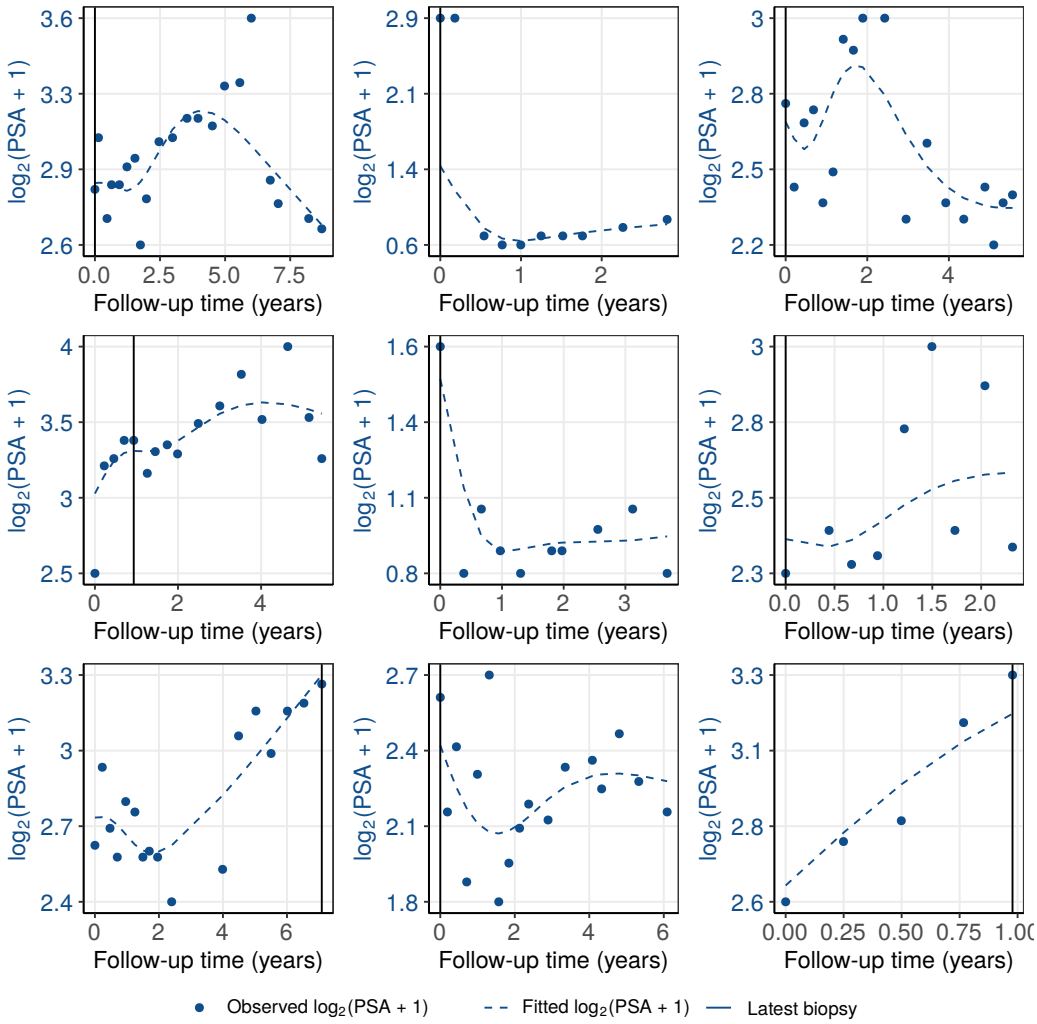

Figure 4: **Fitted versus observed** $\log_2(\mathbf{PSA}+1)$ **profiles** for nine randomly selected PRIAS patients. The fitted profiles utilize information from the observed PSA measurements, and time of the latest biopsy.

Table 7: **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 8: **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]

Appendix B. 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:

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

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 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. [11]. 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 [11]. It is given by:

$$R_j(u \mid t, v) = \Pr\{T_j^* > u \mid T_j^* > t, \mathcal{Y}_j(v), \mathcal{A}_n\}, \quad u \geq t. \tag{4}$$

The personalized risk profile of the patient (see Panel C, Figure 5) updates as more data is gathered over follow-up visits.

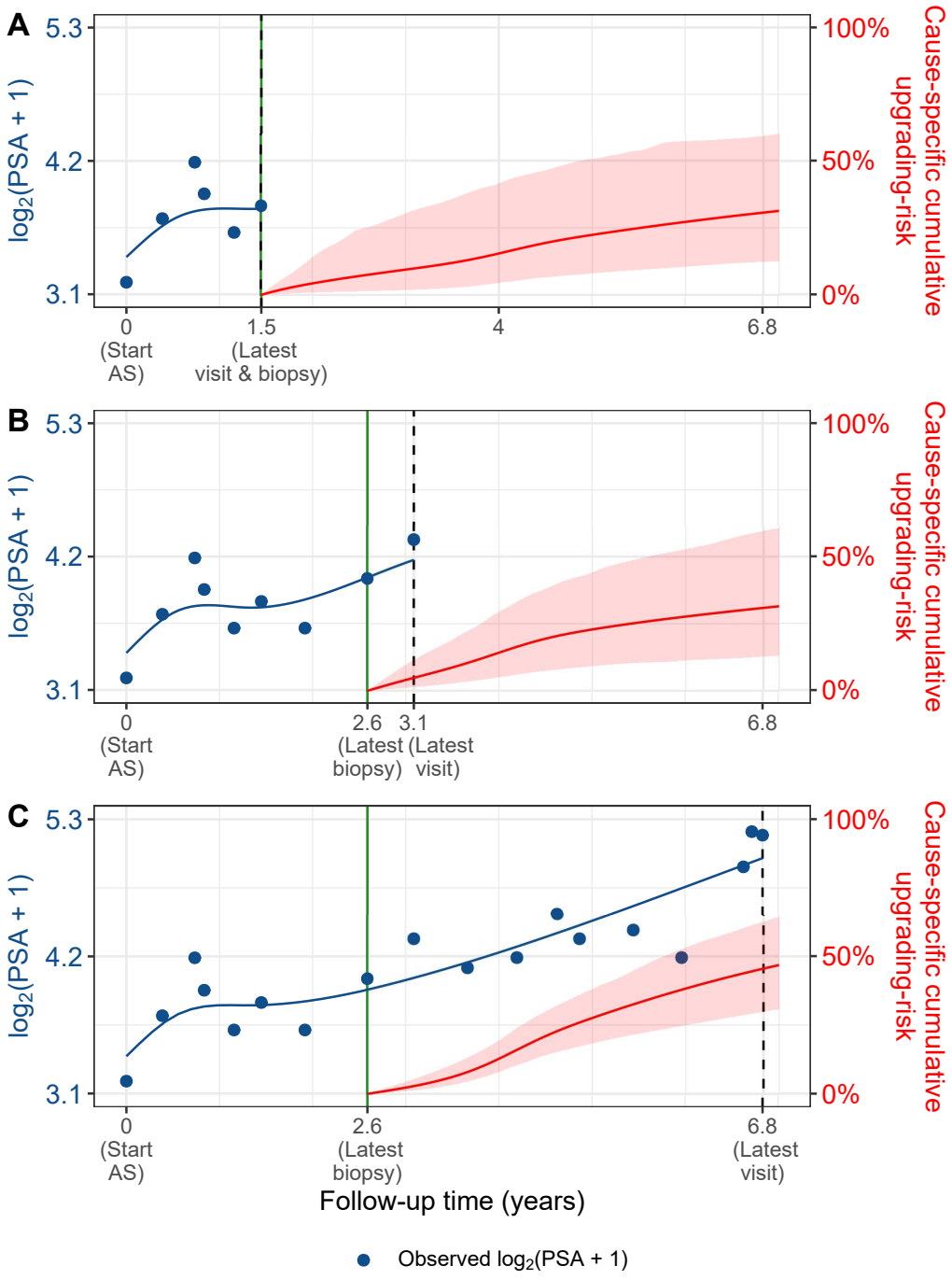

Figure 5: **Cause-specific cumulative upgrading-risk changing dynamically over follow-up** as more patient data is gathered. The three **Panels A,B and C:** are ordered by the time of the latest visit (dashed vertical black line) of a new patient. At each of the latest follow-up visits, we combine the accumulated PSA measurements (shown in blue), and latest time of negative biopsy (solid vertical green line) to obtain the updated cumulative-risk profile (shown in red) of the patient.

Appendix B.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 [8]. 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 [12] 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 [9] of the cause-specific cumulative upgrading-risk in each of the cohorts. The results are shown in Panel A of Figure 6. 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 $\boldsymbol{\gamma}^c_{h_0}$ (Appendix A) of the log baseline hazard are given by:

$$p(\boldsymbol{\gamma}^c_{h_0} \mid \mathcal{A}^c, \boldsymbol{b^c}, \boldsymbol{\theta}) \propto \prod_{i=1}^{n^c} p(l^c_i, r^c_i \mid \boldsymbol{b^c_i}, \boldsymbol{\theta}) p(\boldsymbol{\gamma}^c_{h_0}) \tag{5}$$

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 (Appendix 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 6.

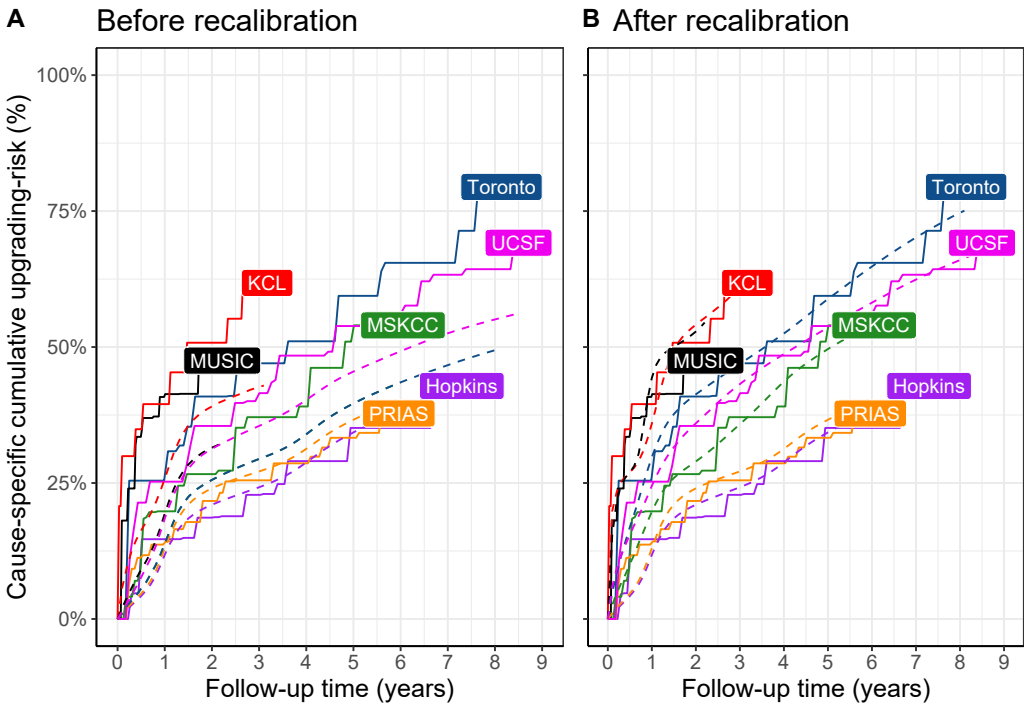

Figure 6: **Calibration-in-the-large of our model:**. In **Panel A** we can see that our model is not well calibrated for use in KCL, MUSIC, Toronto and MSKCC. In **Panel B** we can see that calibration of model predictions improved in KCL, MUSIC, Toronto and MSKCC cohorts after recalibrating our model. 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.

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 7. 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 8). 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.

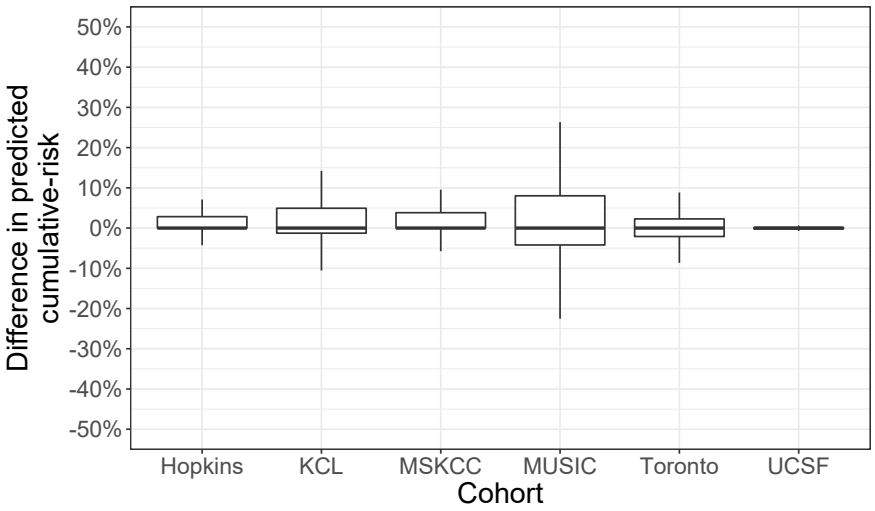

Figure 7: **Comparison of predictions from recalibrated PRIAS model with individual joint models fitted to external cohorts:** On Y-axis we show the difference between predicted cause-specific cumulative upgrading-risk for individual patients using two models, namely the recalibrated PRIAS model for 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 8). 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.

Validation of Dynamic Cumulative-Risk Predictions As shown in Figure 5, 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 [11]. For time-varying prediction error, we assessed the mean absolute prediction error or MAPE [11]. 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 ≤ 0.5 indicates that the model performs worse than random discrimination. Ideally, MAPE should be 0.

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 9.

The results for estimates of AUC and MAPE are summarized in Figure 8, and in Table 10 to Table 16. 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

Table 9: **Maximum follow-up period up to which we can reliably predict upgrading-risk**. 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

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 [13].

Table 10: **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]

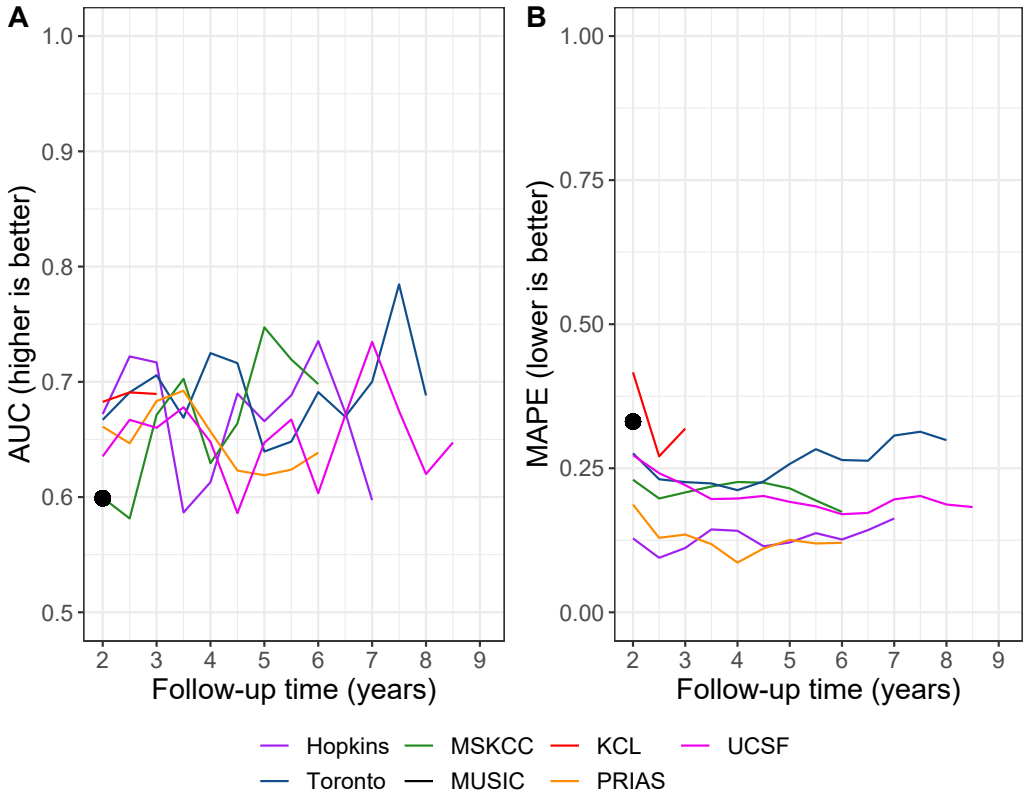

Figure 8: **Validation of dynamic predictions of cause-specific cumulative upgrading-risk**. In **Panel A** area under the receiver operating characteristic curve or AUC (measure of discrimination) is between 0.6 and 0.7. **Panel B** we can see that the time dependent root mean squared prediction error or MAPE is similar for PRIAS and Hopkins cohorts. The bootstrapped 95% confidence interval for these estimates are presented in Table 10 to Table 15. 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 11: **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]

Table 12: **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]

Table 13: **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]

Table 14: **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 15: **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 16: **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]

Appendix C. Personalized Biopsies Based on Cause-Specific Cumulative Upgrading-Risk

Consider some real patients from the PRIAS database shown in Figure 10– 12. In line with the protocols of most AS cohorts [14], we first schedule a compulsory biopsy at year one of follow-up. This promises early detection of Gleason upgrade for patients misdiagnosed as low-grade cancer patients or patients who chose AS despite having a higher grade at diagnosis. We also maintain a recommended minimum gap of one year between consecutive biopsies [13]. That is, we intend to develop a personalized schedule of biopsies for these patients starting from the second year. The added benefit of planning biopsies year two onwards is that due to the longitudinal measurements accumulated over two years, and year one biopsy results, we are able to make reasonably accurate predictions of the cause-specific cumulative upgrading-risk.

Using the joint model fitted to the PRIAS dataset, we first obtain a patient's cause-specific cumulative upgrading-risk over the entire future follow-up period (see 4), given their accumulated two year clinical data. Typically biopsies may be decided on the same visit on which PSA is measured. Let $U = u_1, \ldots, u_L$ represent a schedule of such visits (e.g., every six months in prostate cancer for PSA measurement), where $u_1 = v$ is also the time of the current visit, and u_L is the horizon up to which we intend to plan biopsies. Depending upon how much training/validation data is available, this horizon differs between cohorts (Table 17). First, we make L successive decisions for conducting biopsies on each of the L future visit times $u_l \in U$. Specifically, we decide to conduct a biopsy at time u_l if the conditional cumulative-risk of upgrading at u_l is larger than a certain risk threshold $0 \leq \kappa \leq 1$ (e.g., $\kappa = 12\%$ risk as shown in Figure 9). If a biopsy gets planned at time u_l, then the successive biopsy decision at time u_{l+1} is made using an updated cumulative-risk profile. This updated cumulative-risk profile accounts for the possibility that upgrading may occur after time $u_l < T_j^*$. The biopsy decisions on each future visit time u_l are defined as:

$$Q_j^{\kappa}(u_l \mid t_l, v) = I\big\{R_j(u_l \mid t_l, v) \geq \kappa\big\},$$

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

The cumulative-risk $R_j(u_l \mid t_l, v)$ at future visit time u_l utilizes the time t_l

as the time of the last conducted biopsy on which upgrading may not be observed. However, the contribution of the observed longitudinal data $\mathcal{Y}_j(v)$ in the risk function remains the same over all time points in U. The biopsy decision at time u_l is denoted by $Q_j^\kappa(u_l \mid t_l, v)$. Via the indicator function $I(\cdot)$ it obtains a value 1 (or 0) when a biopsy is to be conducted (or not conducted) at time u_l. The subset of future time points in U on which a biopsy is to be performed results into a personalized schedule of planned future biopsies, given by:

$$S_j^\kappa(U \mid t, v) = \left\{u_l \in U \mid Q_j^\kappa(u_l \mid t_l, v) = 1\right\}. \tag{6}$$

The personalized schedule in (6) is updated as more patient data becomes available over subsequent follow-up visits.

Appendix C.1. Expected Time Delay in Detecting Upgrading

The schedule $S_j^\kappa(U \mid t, v)$ manifests a personalized biopsy plan for the j-the patient. However, the time delay in detecting upgrading that may subsequently be observed depends on the true time of upgrading T_j^* of the patient. Since two different patients with the same timing of biopsies will expect different time delays, we estimate it in a patient-specific manner as well. Although, this calculation is not limited to personalized schedules only, but can be done for any schedule S of biopsies with N time points $S = \{s_n \mid n = 1, \ldots, N\}$.

For each of the N planned biopsies there exist N possible time intervals $s_{n-1} < T_j^* \leq s_n$ in which upgrading may be observed. Correspondingly, there are N possible time delays in detecting upgrading $s_n - T_j^*$. Given a schedule S, the true time delay in detecting upgrading D_j that the patient will experience can be defined as:

$$D_j(S \mid t) = \left\{ \begin{array}{ll} s_1 - T_j^*, & \text{if} \quad t < T_j^* \leq s_1 \\ \ldots & \\ s_N - T_j^*, & \text{if} \quad s_{N-1} < T_j^* \leq s_N \end{array} \right\}. \tag{7}$$

The time delay is cannot be defined for the scenario in which the patient obtains upgrading after the time of the last biopsy in the schedule $T_j^* > s_N$. Hence, this delay should be interpreted as the delay that will be observed if the patient will experience upgrading before time of the last planned biopsy at $T_j^* \leq s_N$. To estimate the expected value of $D_j(\cdot)$ in a patient-specific manner, we exploit the personalized cumulative-risk profile of the patient

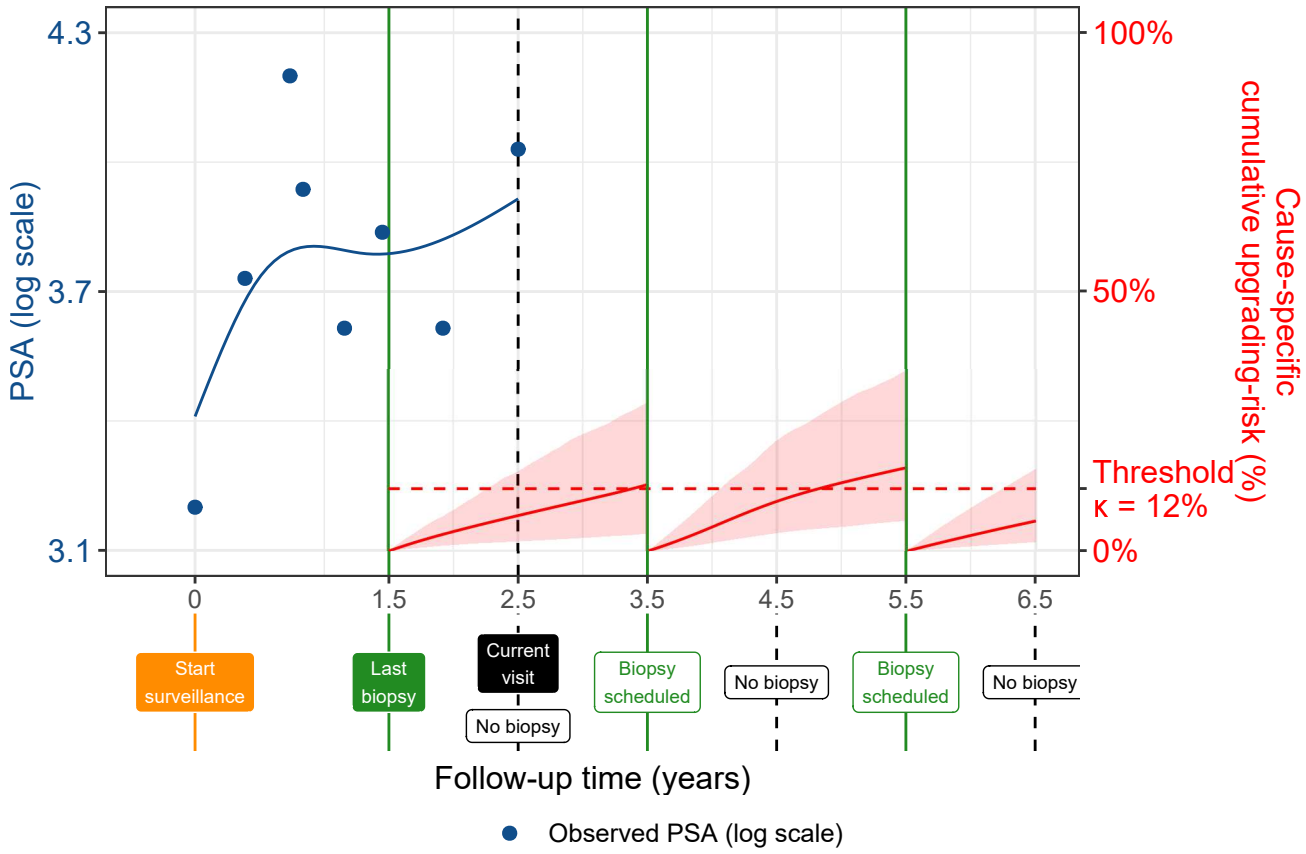

Figure 9: **Illustration of Personalized Biopsy Decisions Using Patient-specific Conditional Cumulative Upgrading-risk**. The last biopsy on which upgrading was not observed was conducted at $t = 1.5$ years. The current visit time of the patient is $v = 2.5$ years. Decisions for biopsy need to be made at a gap of every one year starting from the current visit until a horizon of 6.5 years. That is, $U = \{2.5, 3.5, 4.5, 5.5, 6.5\}$ years. Based on an example risk threshold of 12% ($\kappa = 0.12$) the future biopsy decisions at time points in U lead to a personalized schedule $S_j^{\kappa^*}(U \mid t = 1.5, v = 2.5) = \{3.5, 5.5\}$ years. The conditional cumulative-risk profiles $R_j(u_l \mid t_l, v)$ employed in (Appendix C) are shown with red line (confidence interval shaded). It is called 'conditional' because, for example, the second biopsy at future time 5.5 years, is scheduled after accounting for the possibility that upgrading (true time T_j^*) may not have occurred until the time of the previously scheduled biopsy at time $T_j^* > 3.5$ years. All values are illustrative.

defined in (4). Specifically, the expected time delay $E\{D_j(\cdot)\}$ can be calculated as the weighted sum of N possible time delays defined in (7). The n-th weight is equal to the probability of the patient obtaining upgrading in the n-th interval $s_{n-1} < T_j^* \leq s_n$.

$$E\Big\{D_j(S \mid t)\Big\} = \sum_{n=1}^{N} \Big\{s_n - E(T_j^* \mid s_{n-1}, s_n, v)\Big\} \times \Pr\Big\{s_{n-1} < T_j^* \leq s_n \mid T_j^* \leq s_N, \mathcal{Y}_j(v), \mathcal{A}_n\Big\}, \quad s_0 = t$$

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

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

The personalized expected time delay in detecting upgrading has the advantage that it is updated over follow-up as more patient data become available. Since it can be calculated for any schedule, patients and doctors can utilize it along with the plan of biopsies to compare schedules before making a decision. Although, in order to have a fair comparison of time delays between different schedules for the same patient, a compulsory biopsy at a common horizon time point should be planned in all schedules.

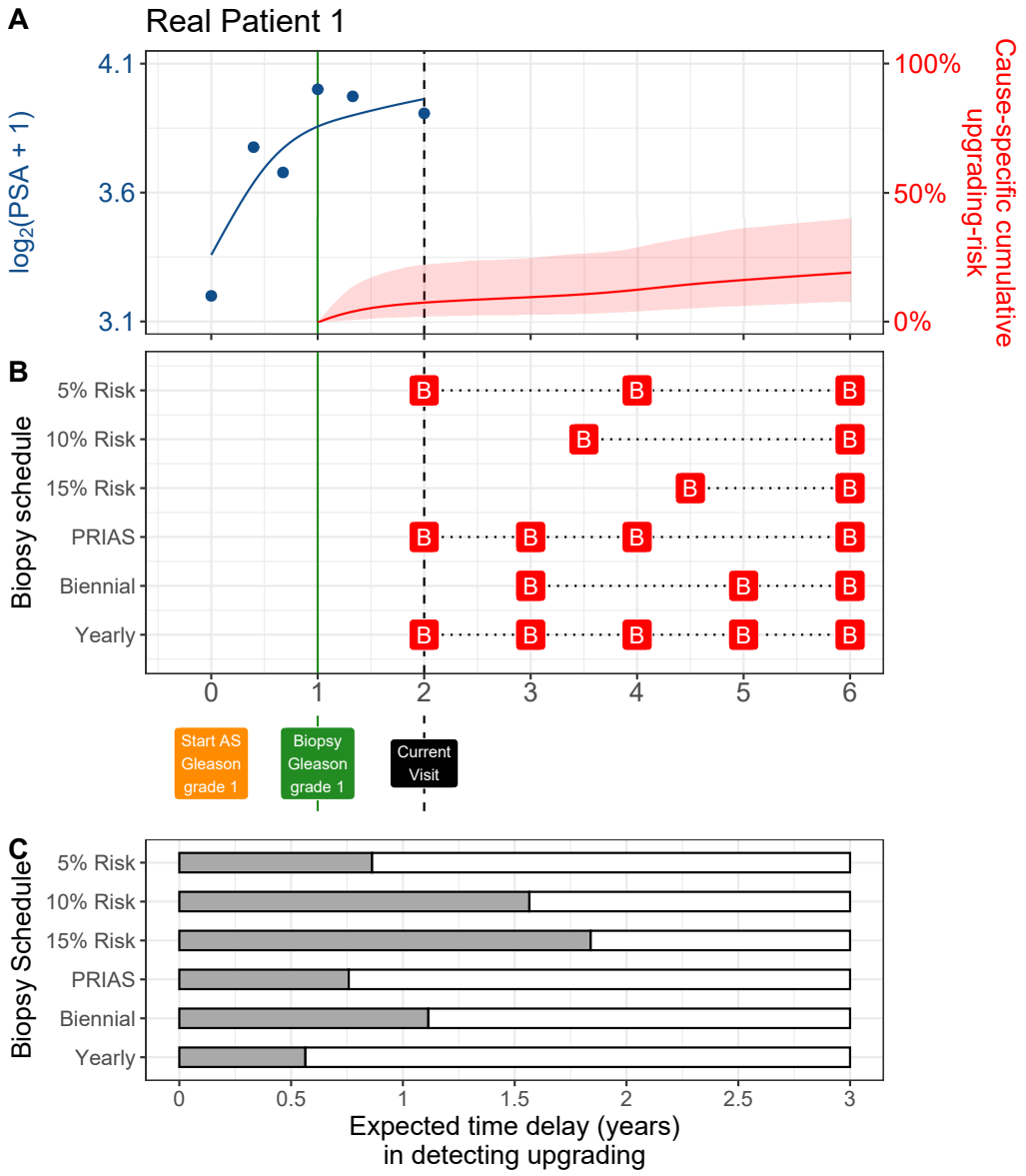

Figure 10: **Personalized and fixed schedules of biopsies for patient 1**. **Panel A:** shows the observed and fitted $\log_2(\text{PSA} + 1)$ measurements (Equation 1), and the dynamic cause-specific cumulative upgrading-risk (see Appendix B) over follow-up period. **Panel B** shows the personalized and fixed schedules of biopsies with a 'B' indicating times of biopsies. Smaller risk thresholds lead to more frequently planned biopsies. **Panel C** various schedules are compared in terms of the expected time delay in detecting upgrading (years) if patient progresses before year six. The maximum time delay with limited adverse consequences is three years [15]. A compulsory biopsy was scheduled at year six (maximum biopsy scheduling time in PRIAS, Table 17) in all schedules for a meaningful comparison between them.

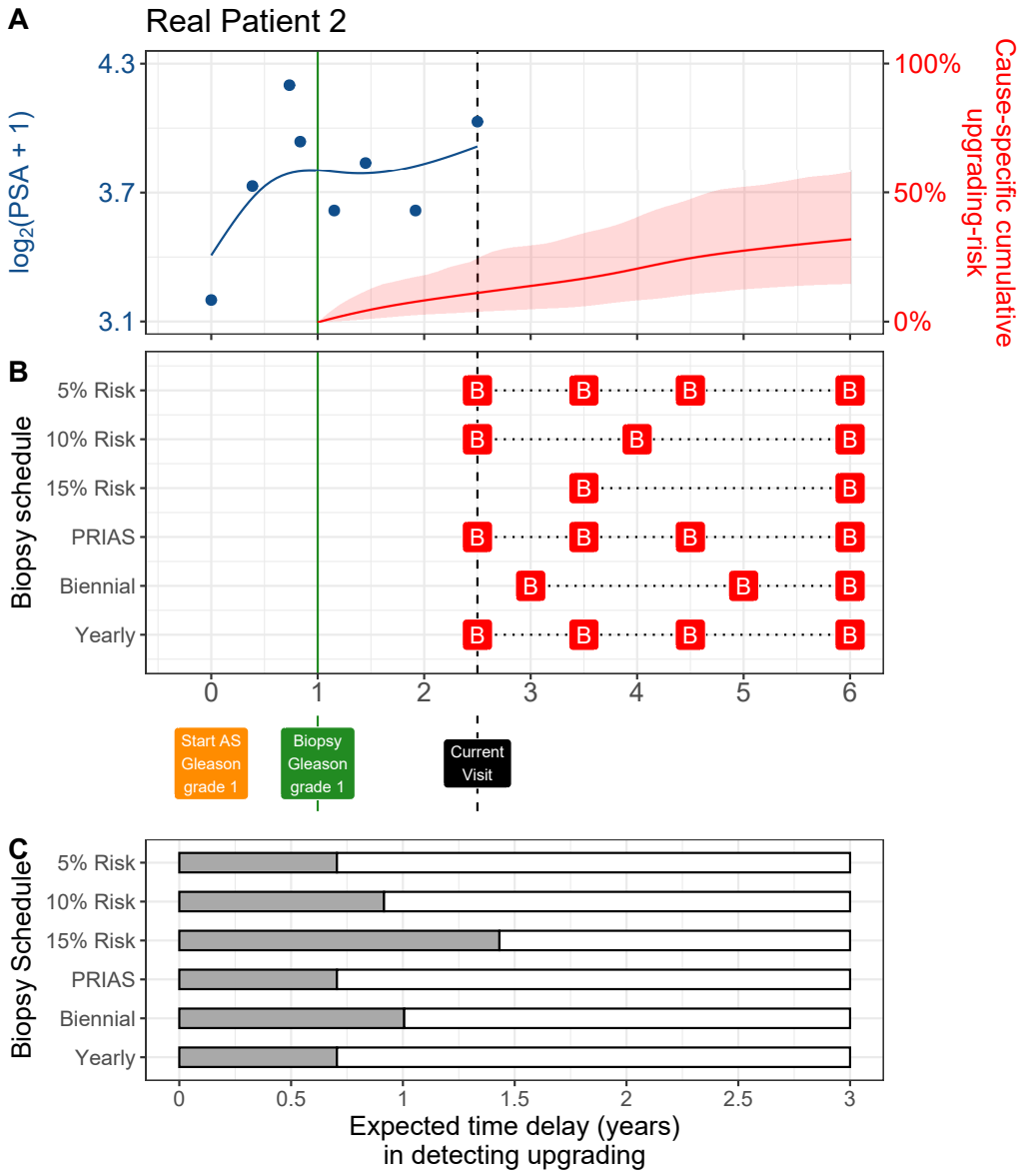

Figure 11: **Personalized and fixed schedules of biopsies for patient 2**. **Panel A:** shows the observed and fitted $\log_2(\text{PSA} + 1)$ measurements (Equation 1), and the dynamic cause-specific cumulative upgrading-risk (see Appendix B) over follow-up period. **Panel B** shows the personalized and fixed schedules of biopsies with a 'B' indicating times of biopsies. Smaller risk thresholds lead to more frequently planned biopsies. **Panel C** various schedules are compared in terms of the expected time delay in detecting upgrading (years) if patient progresses before year six. The maximum time delay with limited adverse consequences is three years [15]. A compulsory biopsy was scheduled at year six (maximum biopsy scheduling time in PRIAS, Table 17) in all schedules for a meaningful comparison between them.

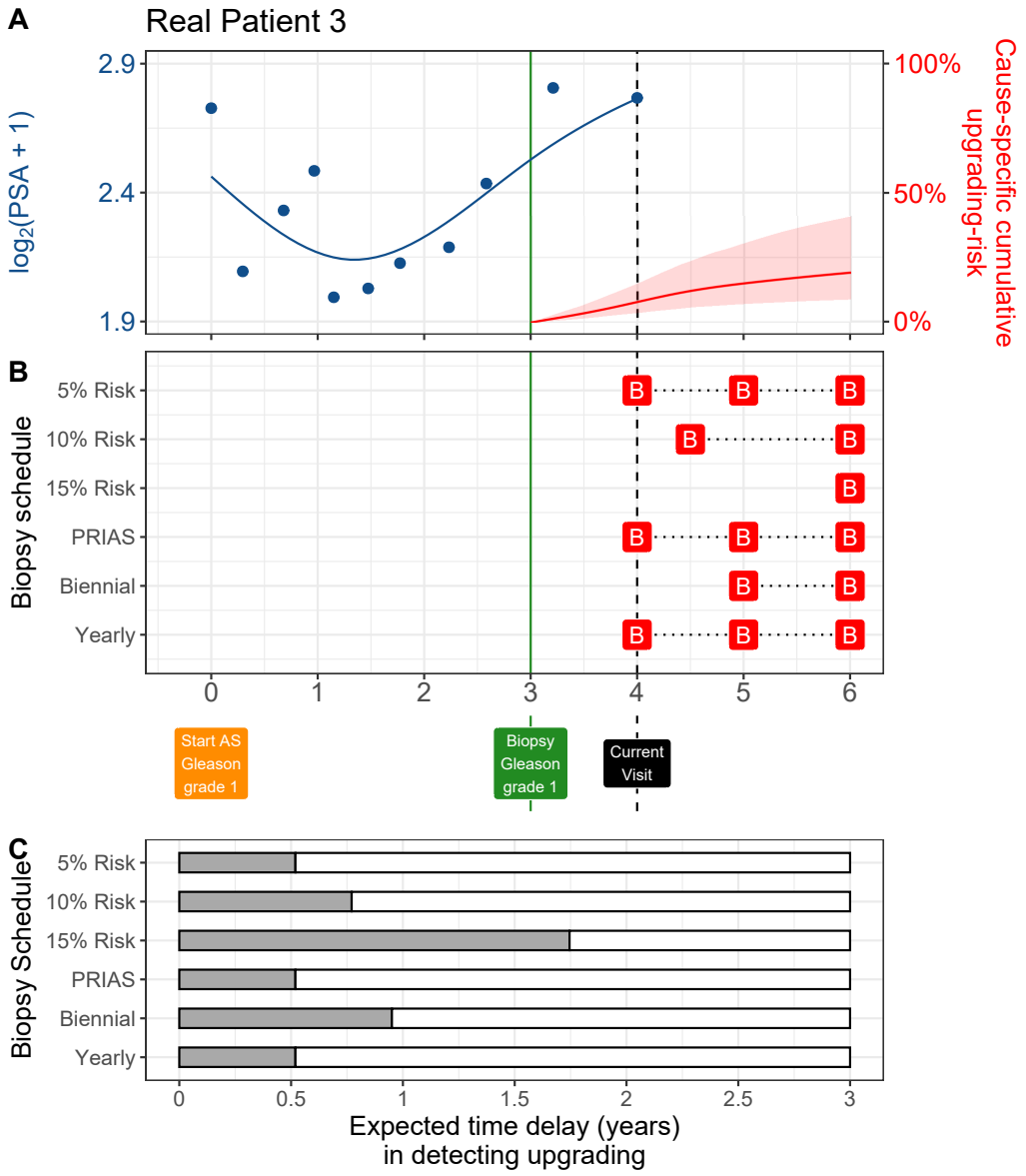

Figure 12: **Personalized and fixed schedules of biopsies for patient 3**. **Panel A:** shows the observed and fitted $\log_2(\text{PSA} + 1)$ measurements (Equation 1), and the dynamic cause-specific cumulative upgrading-risk (see Appendix B) over follow-up period. **Panel B** shows the personalized and fixed schedules of biopsies with a 'B' indicating times of biopsies. Smaller risk thresholds lead to more frequently planned biopsies. **Panel C** various schedules are compared in terms of the expected time delay in detecting upgrading (years) if patient progresses before year six. The maximum time delay with limited adverse consequences is three years [15]. A compulsory biopsy was scheduled at year six (maximum biopsy scheduling time in PRIAS, Table 17) in all schedules for a meaningful comparison between them.

Table 17: **Maximum follow-up period up to which we can reliably make 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 Personalized Schedule Time (years)
PRIAS	6
KCL	3
MUSIC	2
Toronto	8
MSKCC	6
Hopkins	7
UCSF	8.5

Appendix D. Web-Application for Practical Use of Personalized Schedule of Biopsies

We implemented our methodology in a web-application to assist patients and doctors in better decision making. It works on desktop as well as mobile devices. The cohorts that are currently supported in this web-application are PRIAS and the largest six cohorts from the GAP3 database [8]. These are the University of Toronto AS (Toronto), Johns Hopkins AS (Hopkins), Memorial Sloan Kettering Cancer Center AS (MSKCC), King's College London AS (KCL), Michigan Urological Surgery Improvement Collaborative AS (MUSIC), and University of California San Francisco Active Surveillance (UCSF). The web application is hosted at `https://emcbiostatistics.shinyapps.io/prias_biopsy_recommender/`.

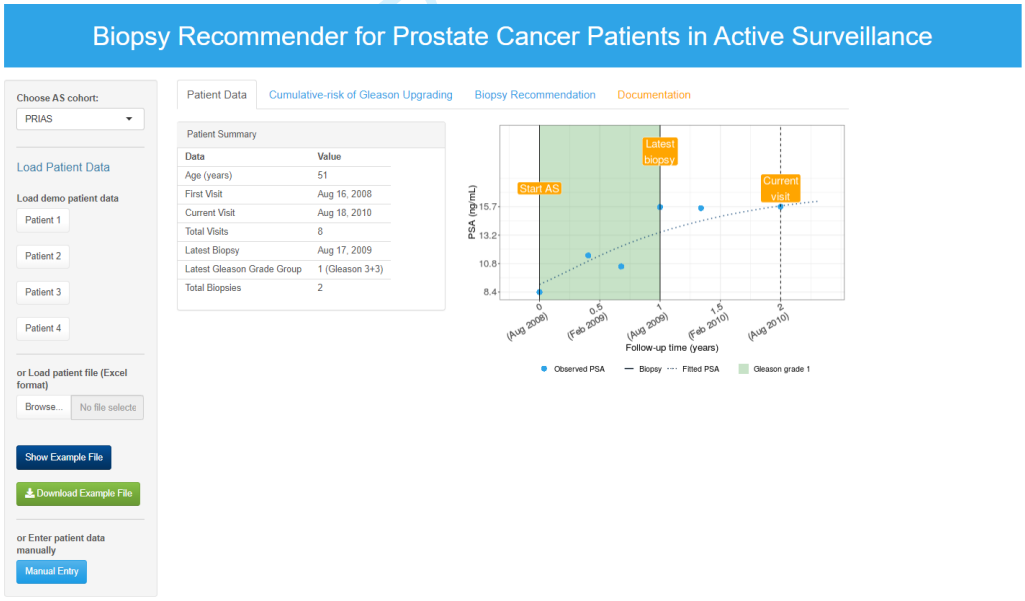

Figure 13: Landing page of the web-application. Panel on the left allows users to load patient data and panel on the right provides information. Patient data can be entered manually, or via Excel files. In addition, demo patient data is already uploaded to assist users in understanding the web-application.

Appendix E. 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.

Appendix E.1. Fitting the Joint Model to the PRIAS dataset

Accessing the dataset: 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`.

Formatting the dataset: This dataset, however, 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 biopsies.

Fitting the joint model: 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 supplementary material can be generated by the scripts in `R_HOME/plots/`.

Appendix E.2. Validation of Predictions of Upgrading

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.

Appendix E.3. Creating Personalized Schedules of Biopsies

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`.

Appendix E.4. Source Code for Web Application

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

Appendix F. Appendix A. 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).

References

1. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American journal of surgical pathology* 2016;40(2):244–52.

2. Pearson JD, Morrell CH, Landis PK, Carter HB, Brant LJ. Mixed-effects regression models for studying the natural history of prostate disease. *Statistics in Medicine* 1994;13(5-7):587–601.

3. Lin H, McCulloch CE, Turnbull BW, Slate EH, Clark LC. A latent

class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine* 2000;19(10):1303–18.

4. De Boor C. A practical guide to splines; vol. 27. Springer-Verlag New York; 1978.

5. Eilers PH, Marx BD. Flexible smoothing with B-splines and penalties. *Statistical Science* 1996;11(2):89–121.

6. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 2004;14(3):809–34.

7. Gentleman R, Geyer CJ. Maximum likelihood for interval censored data: Consistency and computation. *Biometrika* 1994;81(3):618–23.

8. Bruinsma SM, Zhang L, Roobol MJ, Bangma CH, Steyerberg EW, Nieboer D, Van Hemelrijck M, consortium MFGAPPCASG, Trock B, Ehdaie B, et al. The Movember foundation's GAP3 cohort: a profile of the largest global prostate cancer active surveillance database to date. *BJU international* 2018;121(5):737–44.

9. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B (Methodological)* 1976;38(3):290–5.

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

11. Rizopoulos D, Molenberghs G, Lesaffre EM. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal* 2017;59(6):1261–76.

12. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass)* 2010;21(1):128.

13. Bokhorst LP, Alberts AR, Rannikko A, Valdagni R, Pickles T, Kakehi Y, Bangma CH, Roobol MJ, PRIAS study group . Compliance rates with the Prostate Cancer Research International Active Surveillance (PRIAS)

protocol and disease reclassification in noncompliers. *European Urology* 2015;68(5):814–21.

14. Nieboer D, Tomer A, Rizopoulos D, Roobol MJ, Steyerberg EW. Active surveillance: a review of risk-based, dynamic monitoring. *Translational andrology and urology* 2018;7(1):106–15.

15. de Carvalho TM, Heijnsdijk EA, de Koning HJ. Estimating the risks and benefits of active surveillance protocols for prostate cancer: a microsimulation study. *BJU international* 2017;119(4):560–6.

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

Anirudh Tomer, MSc[a,*], Daan Nieboer, MSc[b,c], Monique J. Roobol, PhD[c], Anders Bjartell, MD, PhD[d], Ewout W. Steyerberg, PhD[b,e], Dimitris Rizopoulos, PhD[a], Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium[f]

[a] *Department of Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands*
[b] *Department of Public Health, Erasmus University Medical Center, Rotterdam, the Netherlands*
[c] *Department of Urology, Erasmus University Medical Center, Rotterdam, the Netherlands*
[d] *Department of Urology, Skåne University Hospital, Malmö, Sweden*
[e] *Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands*
[f] *The Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium members presented in Appendix A*

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,

*Word count: 3636

*Corresponding author (Anirudh Tomer): Erasmus MC, kamer flex Na-2823, PO Box 2040, 3000 CA Rotterdam, the Netherlands. Tel: +31 10 70 43393.

Email addresses: `a.tomer@erasmusmc.nl` (Anirudh Tomer, MSc), `d.nieboer@erasmusmc.nl` (Daan Nieboer, MSc), `m.roobol@erasmusmc.nl` (Monique J. Roobol, PhD), `anders.bjartell@med.lu.se` (Anders Bjartell, MD, PhD), `e.w.steyerberg@lumc.nl` (Ewout W. Steyerberg, PhD), `d.rizopoulos@erasmusmc.nl` (Dimitris Rizopoulos, PhD)

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.

Keywords: Active Surveillance, Biopsies, Personalized Medicine, Prostate Cancer, Shared Decision Making.

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 [1]. 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 [2]. When it increases from group 1 (Gleason 3+3) to 2 (Gleason 3+4) or higher, it is called *upgrading* [3]. Upgrading is an important endpoint in AS upon which patients are commonly advised curative treatment [4].

Biopsies in AS are always conducted with a time gap between them. Consequently, upgrading is always detected with a time delay (Figure 1) that cannot be measured directly. In this regard, to detect upgrading timely, many patients are prescribed fixed and frequent biopsies, most often annually [5]. 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[6]. Thus, biopsy burden and patient non-compliance to frequent biopsies [7] have raised concerns regarding the optimal biopsy schedule [8, 9] in AS.

Except for the confirmatory biopsy at year one of AS [7], opinions and practice regarding the timing of remaining biopsies lack agreement [10]. Some AS programs utilize patients' observed PSA, DRE, previous biopsy Gleason grade, and lately, MRI results to decide biopsies [11, 4, 10]. In contrast, others discourage schedules based on clinical data and MRI results [12, 5], and instead support periodical one-size-fits-all biopsy schedules. Furthermore, some suggest replacing frequent periodical schedules with infrequent ones (e.g., biennially) [8, 13]. 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 [8]. Whereas, since observed clinical data has measurement error (e.g., PSA fluctuations),

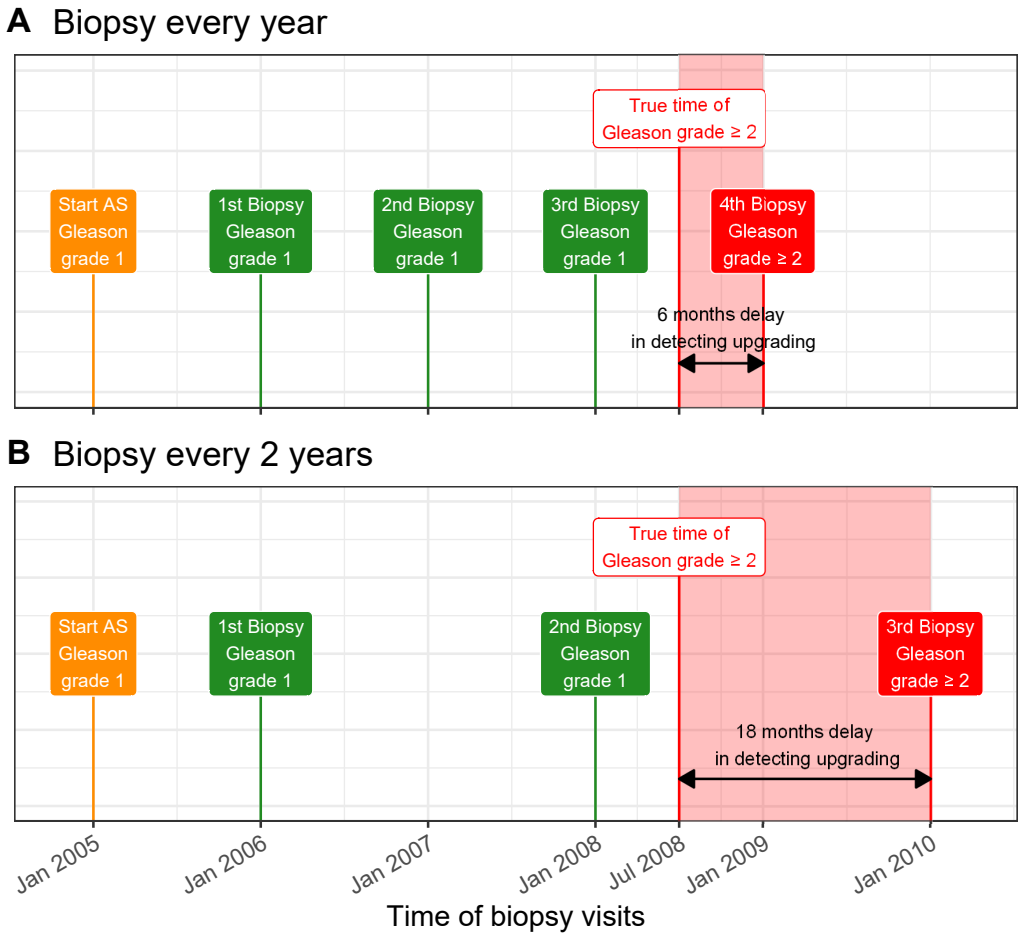

Figure 1: **Trade-off between the timing and number of biopsies (burden) and time delay in detecting Gleason upgrading (shorter is better):** The true time of Gleason upgrading (increase in Gleason grade group from group 1 to 2 or higher) for the patient in this figure is July 2008. When biopsies are 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.

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 [10] (Figure 2). While, upgrading-risk calculators are not new [14, 15, 16, 17], 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 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 PSA measurements of patients, previous biopsy results, and baseline characteristics. 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,

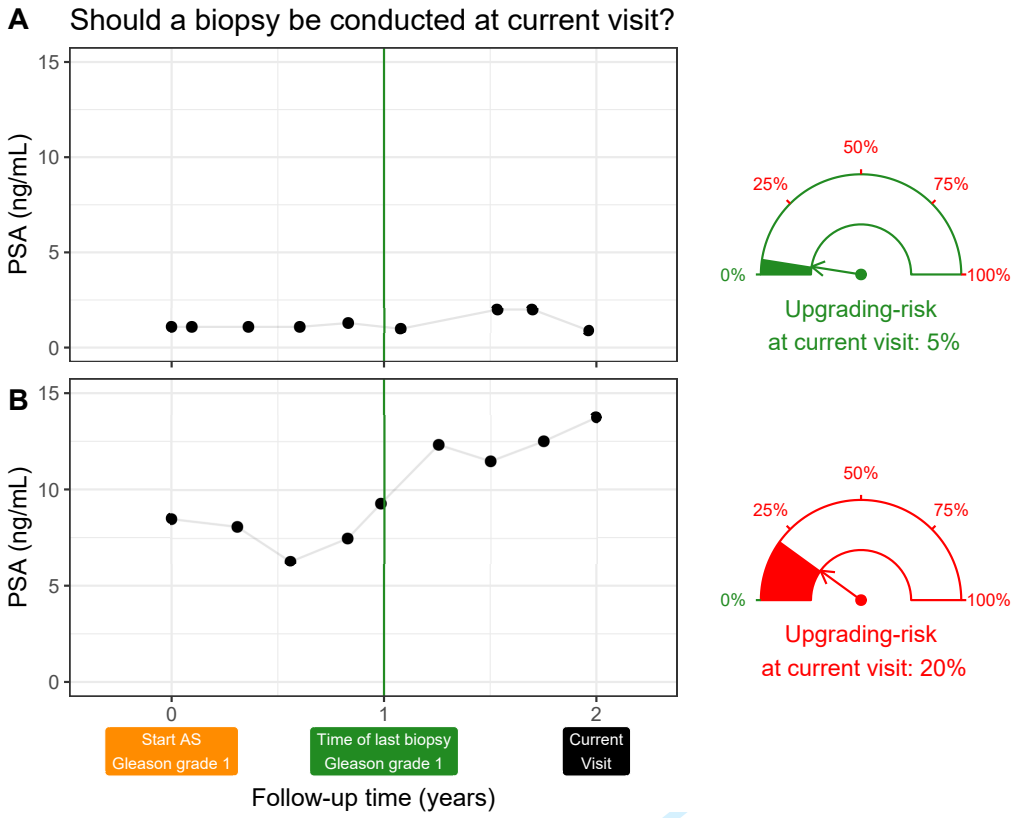

Figure 2: **Motivation for upgrading-risk based personalized biopsy decisions**: To utilize patients' complete longitudinal data and results from previous biopsies in making biopsy decisions. For this purpose, we first process data using a statistical model and then utilize the patient-specific predictions 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.

we will externally validate it in the largest six AS cohorts from the Movember Foundation's Global Action Plan (GAP3) database [18]. Last, we aim to implement the validated model and methodology in a web-application.

2. Patients and Methods

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 [4], dated April 2019 (Table 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 [19]. 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 [10].

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 [18] version 3.1, covering nearly 73% of the GAP3 patients: the University of Toronto AS (Toronto), Johns Hopkins AS (Hopkins), Memorial Sloan Kettering Cancer 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 Supplementary A.2.

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 [11, 12, 20]. 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.

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 pa-

Table 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 [2] 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)

tient 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 [21, 14, 22].

Our joint model consisted of two sub-models. Namely, a linear mixed-effects sub-model [23] 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 [24], 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 3). Subsequently, we calculated the mathematical derivative of the patient's fitted PSA profile (Equation 2, Supplementary A), to obtain his follow-up time

specific instantaneous PSA velocity (Panel B, Figure 3). This instantaneous velocity is a stronger predictor of upgrading than the widely used average PSA velocity [25]. 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 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 Supplementary A). This approach also accommodates watchful waiting and treatment protocols that are also based on patient data. Specifically, the parameters of our two sub-models were estimated jointly under the Bayesian paradigm (Supplementary A) using the R package **JMbayes** [26].

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 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 risk predictions for a patient are not calculated only once. Rather, as illustrated in Figure 5 of Supplementary B, risk-predictions update over the follow-up, to account for additional patient data (e.g., new biopsy results, PSA measurements) that becomes available. We validated our model internally in the PRIAS cohort, and externally in the largest six GAP3 database cohorts. We employed calibration plots [27, 28] and follow-up *time-dependent* mean absolute risk prediction error or MAPE [29] to graphically and quantitatively evaluate our model's risk prediction accuracy, respectively. We assessed our

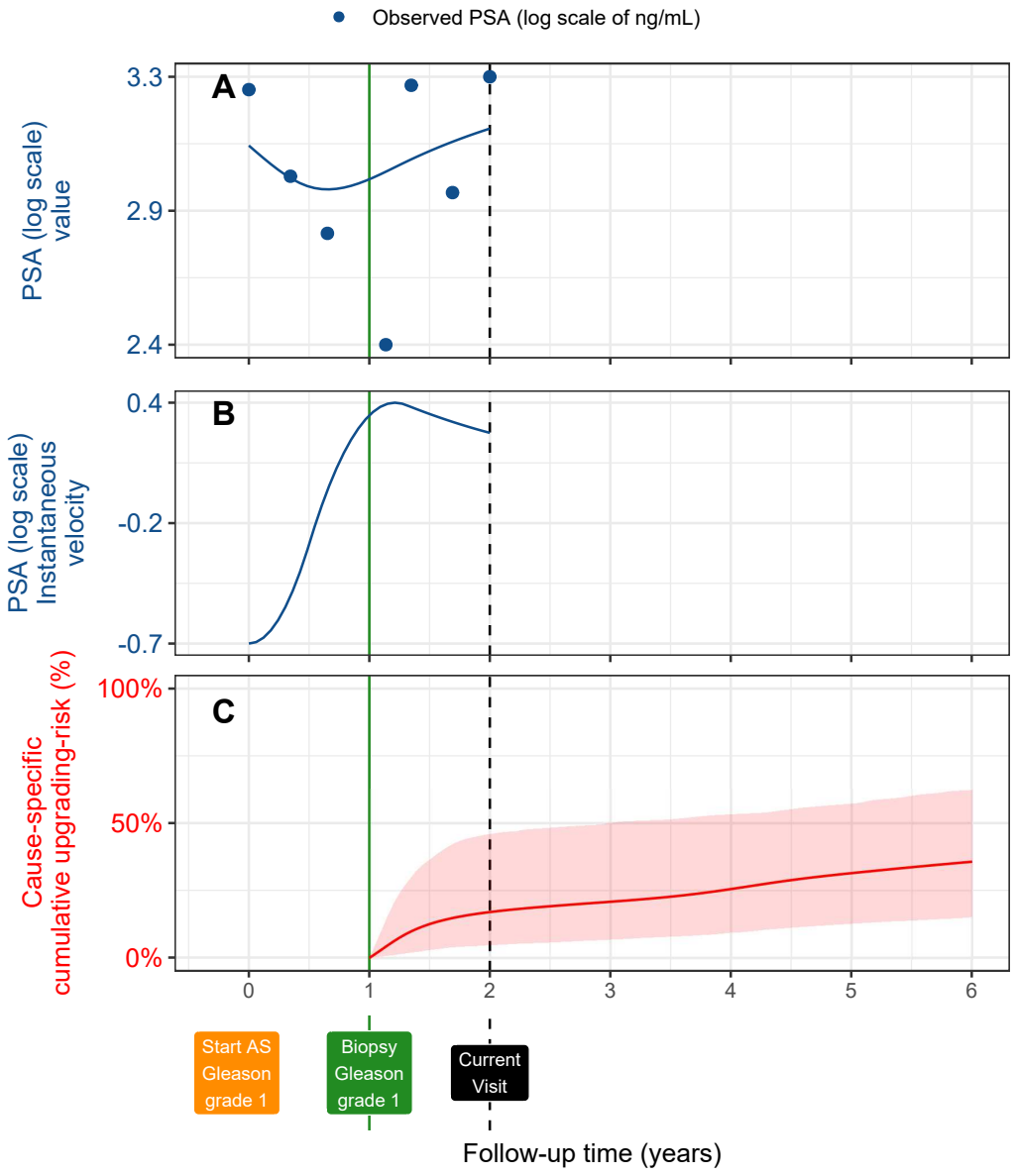

Figure 3: **Illustration of the joint model on a real PRIAS patient**. **Panel A:** Observed PSA (blue dots) and fitted PSA (solid blue line), log-transformed from ng/mL. **Panel B:** Estimated instantaneous velocity of PSA (log-transformed). **Panel C**: Predicted cause-specific cumulative upgrading-risk (95% credible interval shaded). Upgrading is defined as an increase in the Gleason grade group from group 1 [2] 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.

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 [29].

The aforementioned *time-dependent* AUC and MAPE [29] are temporal extensions of their standard versions [28] 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 (Supplementary B.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. For example, calculations for AUC and MAPE for the time interval year two to year three do not utilize data of patients who progressed before year two. Last, to resolve any potential model miscalibration in validation cohorts, we aimed to recalibrate our model's baseline hazard of upgrading (Supplementary B.1), individually for each cohort.

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 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 (Supplementary Table 8).

The time-dependent AUC, calibration plot, and time-dependent MAPE of our model are shown in Figure 4, and Supplementary Figure 8. 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, Supplementary Table 8), and large (0.2 to 0.3) otherwise. Our model was miscalibrated for validation cohorts (Panel B, Figure 4) because cohorts had differences in inclusion criteria (e.g., PSA density) and follow-up protocols [18] 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 (Supplementary Figure 6). We compared risk predictions from the recalibrated models, with predictions from separately fitted cohort-specific joint models (Supplementary Figure 7). The difference in predictions was lowest in the Johns Hopkins cohort (impact of PSA on upgrading-risk similar to PRIAS). Comprehensive results are in Supplementary A.3 and Supplementary B.

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) 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 Supplementary C for

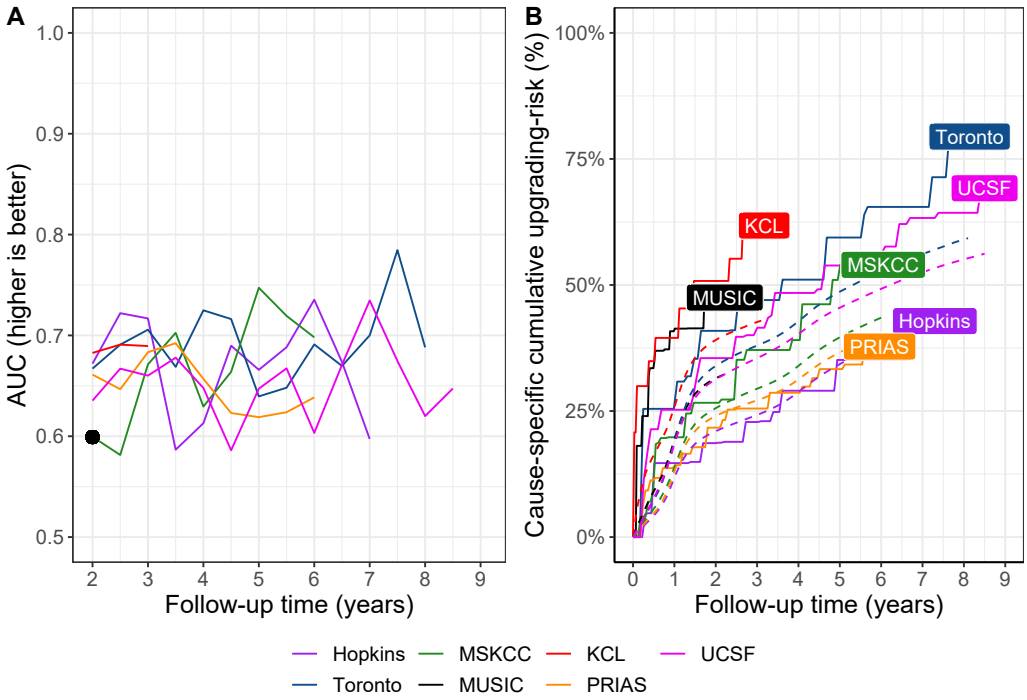

Figure 4: **Model Validation Results**. **Panel A**: time-dependent area under the receiver operating characteristic curve or AUC (measure of discrimination). AUC at year one is not shown because we do not intend to replace the confirmatory biopsy at year one. **Panel B**: calibration-at-large indicates model miscalibration. This is because solid lines depicting the non-parameteric estimate of the cause-specific cumulative upgrading-risk [30], 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 (Supplementary Figure 6). 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.

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, and in Supplementary Figure 10–12.

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, 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 all available clinical data of the patient (see Supplementary C). 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.

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. Users must first choose the cohort to which the patient belongs (left panel), and then upload

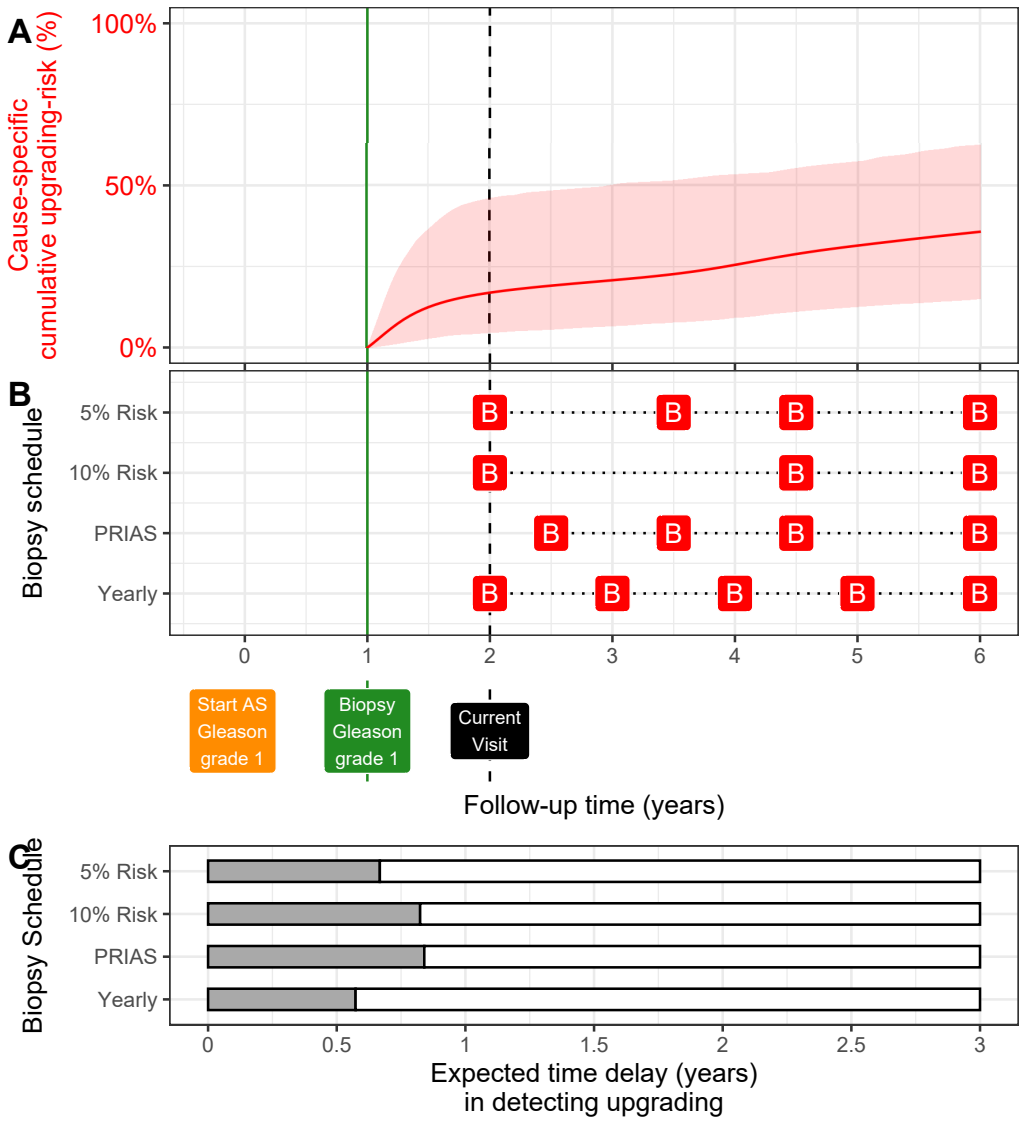

Figure 5: **Illustration of personalized and fixed schedules of biopsies for patient from Figure 3**. **Panel A:** Predicted cumulative upgrading-risk (95% credible interval shaded). **Panel B:** Different biopsy schedules with a red 'B' indicating a future biopsy. Risk: 5% and Risk: 10% are personalized schedules in which a biopsy is planned whenever the conditional cause-specific cumulative upgrading-risk is above 5% or 10% risk, respectively. Smaller risk thresholds lead to more frequently planned biopsies. 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. The maximum time delay with limited adverse consequences is three years [13]. A compulsory biopsy was scheduled at year six (maximum biopsy scheduling time in PRIAS, Supplementary C) in all schedules for a meaningful comparison between delays.

patient data in Microsoft Excel format. Internally, the web-application loads the appropriate validated and recalibrated model for that cohort. The maximum follow-up time up to which predictions can be obtained depends on each cohort (Supplementary Table 9). 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.

4. Discussion

We successfully developed and externally validated a statistical model for predicting upgrading-risk [3] in prostate cancer AS, and providing risk-based personalized biopsy decisions. Our work has four novel features over earlier risk calculators [14, 15]. 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 [18]. 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 following 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. A benefit of our model is that it allows the biopsy schedule, schedule of longitudinal measurements, and loss to follow-up in each cohort to depend on patient age and PSA characteristics. Consequently, in future, when MRI data is included in the model, our model will also accommodate biopsy schedules dependent on MRI data or MRI schedules dependent on previous biopsy results, PSA characteristics, and age of the patient (mathematical proof in Appendix A.2). An additional advantage of our model and resulting personalized schedules is that they update as more patient data becomes available over follow-up. 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 [28] 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 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 [7], the availability of our methodology as a web-application may encourage patients/doctors to consider upgrading-risk based personalized schedules instead. 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 personalized schedules is quite low because of the low rate of metastases and prostate cancer specific mortality in AS patients (Table 1). Second, studies [31, 8] 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 [32]. 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 (Supplementary Table 9). 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. In this regard, the current model can be extended to include both MRI and PSA density data as predictors when they become available in the future. Our model schedules biopsies in a personalized manner, but the patient burden can be decreased even more if we also personalize the schedule of PSA measurements. A caveat of doing so is that reduction in the number of PSA measurements can also lead to an increase in the variance of risk estimates, and also affect the performance of personalized schedules. Although we have done a simulation study to conclude that personalized schedules may not be any more harmful than PRIAS or annual schedule [32], with an infrequent PSA schedule, these conclusions may not hold. Hence, we do not recommend any changes in the schedule of PSA measurements from the current protocol of PSA measurements every six months. At the same time, personalizing the schedule of both biopsies and PSA measurements together

is a research problem we intend to tackle in the near 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 [14, 33] will be interesting to investigate further. Even with an enhanced risk prediction model, the methodology for personalized scheduling and calculation of expected time delay (Supplementary C) 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. 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.

Author Contributions

Anirudh Tomer had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Tomer, Nieboer, Roobol, Bjartell, and Rizopoulos

Acquisition of data: Tomer, Nieboer, and Roobol

Analysis and interpretation of data: Tomer, Nieboer, and Rizopoulos

Drafting of the manuscript: Tomer, and Rizopoulos

Critical revision of the manuscript for important intellectual content: Tomer, Nieboer, Roobol, Bjartell, Steyerberg, and Rizopoulos

Statistical analyses: Tomer, Nieboer, Steyerberg, and Rizopoulos

Obtaining funding: Roobol, Steyerberg, and Rizopoulos

Administrative, technical or material support: Nieboer

Supervision: Roobol, and Rizopoulos

Other: none

Acknowledgments

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 Bi-

ology 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.

Appendix A. 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 Uni-

versity 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).

References

[1] Briganti A, Fossati N, Catto JW, Cornford P, Montorsi F, Mottet N, et al. Active surveillance for low-risk prostate cancer: the European Association of Urology position in 2018. European urology 2018;74(3):357–68.

[2] Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. The American journal of surgical pathology 2016;40(2):244–52.

[3] Bruinsma SM, Roobol MJ, Carroll PR, Klotz L, Pickles T, Moore CM, et al. Expert consensus document: semantics in active surveillance for

men with localized prostate cancer—results of a modified Delphi consensus procedure. Nature reviews urology 2017;14(5):312.

[4] Bul M, Zhu X, Valdagni R, Pickles T, Kakehi Y, Rannikko A, et al. Active surveillance for low-risk prostate cancer worldwide: the PRIAS study. European urology 2013;63(4):597–603.

[5] Loeb S, Carter HB, Schwartz M, Fagerlin A, Braithwaite RS, Lepor H. Heterogeneity in active surveillance protocols worldwide. Reviews in urology 2014;16(4):202–3.

[6] Loeb S, Vellekoop A, Ahmed HU, Catto J, Emberton M, Nam R, et al. Systematic review of complications of prostate biopsy. European urology 2013;64(6):876–92.

[7] Bokhorst LP, Alberts AR, Rannikko A, Valdagni R, Pickles T, Kakehi Y, et al. Compliance rates with the Prostate Cancer Research International Active Surveillance (PRIAS) protocol and disease reclassification in noncompliers. European Urology 2015;68(5):814–21.

[8] Inoue LY, Lin DW, Newcomb LF, Leonardson AS, Ankerst D, Gulati R, et al. Comparative analysis of biopsy upgrading in four prostate cancer active surveillance cohorts. Annals of internal medicine 2018;168(1):1–9.

[9] Bratt O, Carlsson S, Holmberg E, Holmberg L, Johansson E, Josefsson A, et al. 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 2013;47(5):347–55.

[10] Nieboer D, Tomer A, Rizopoulos D, Roobol MJ, Steyerberg EW. Active surveillance: a review of risk-based, dynamic monitoring. Translational andrology and urology 2018;7(1):106–15.

[11] Kasivisvanathan V, Giganti F, Emberton M, Moore CM. Magnetic resonance imaging should be used in the active surveillance of patients with localised prostate cancer. European urology 2020;77(3):318–9.

[12] Chesnut GT, Vertosick EA, Benfante N, Sjoberg DD, Fainberg J, Lee T, et al. 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 2020;77(4):501–7.

[13] de Carvalho TM, Heijnsdijk EA, de Koning HJ. Estimating the risks and benefits of active surveillance protocols for prostate cancer: a microsimulation study. BJU international 2017;119(4):560–6.

[14] Coley RY, Zeger SL, Mamawala M, Pienta KJ, Carter HB. Prediction of the pathologic Gleason score to inform a personalized management program for prostate cancer. European urology 2017;72(1):135–41.

[15] Ankerst DP, Xia J, Thompson Jr IM, Hoefler J, Newcomb LF, Brooks JD, et al. Precision medicine in active surveillance for prostate cancer: development of the canary–early detection research network active surveillance biopsy risk calculator. European urology 2015;68(6):1083–8.

[16] Partin AW, Yoo J, Carter HB, Pearson JD, Chan DW, Epstein JI, et al. 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 1993;150(1):110–4.

[17] Makarov DV, Trock BJ, Humphreys EB, Mangold LA, Walsh PC, Epstein JI, et al. 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 2007;69(6):1095–101.

[18] Bruinsma SM, Zhang L, Roobol MJ, Bangma CH, Steyerberg EW, Nieboer D, et al. The Movember foundation's GAP3 cohort: a profile of the largest global prostate cancer active surveillance database to date. BJU international 2018;121(5):737–44.

[19] Mottet N, Bellmunt J, Bolla M, Briers E, Cumberbatch MG, De Santis M, et al. Eau-estro-siog guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent. European Urology 2017;71(4):618–29.

[20] Schoots IG, Petrides N, Giganti F, Bokhorst LP, Rannikko A, Klotz L, et al. Magnetic resonance imaging in active surveillance of prostate cancer: a systematic review. European urology 2015;67(4):627–36.

[21] Tomer A, Nieboer D, Roobol MJ, Steyerberg EW, Rizopoulos D. Personalized schedules for surveillance of low-risk prostate cancer patients. Biometrics 2019;75(1):153–62.

[22] Rizopoulos D. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. CRC Press; 2012. ISBN 9781439872864.

[23] Laird NM, Ware JH, et al. Random-effects models for longitudinal data. Biometrics 1982;38(4):963–74.

[24] Nixon RG, Wener MH, Smith KM, Parson RE, Strobel SA, Brawer MK. 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 1997;157(6):2183–90.

[25] Cooperberg MR, Brooks JD, Faino AV, Newcomb LF, Kearns JT, Carroll PR, et al. Refined analysis of prostate-specific antigen kinetics to predict prostate cancer active surveillance outcomes. European urology 2018;74(2):211–7.

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

[27] Royston P, Altman DG. External validation of a cox prognostic model: principles and methods. BMC medical research methodology 2013;13(1):33.

[28] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass) 2010;21(1):128.

[29] Rizopoulos D, Molenberghs G, Lesaffre EM. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biometrical Journal 2017;59(6):1261–76.

[30] Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society Series B (Methodological) 1976;38(3):290–5.

[31] de Carvalho TM, Heijnsdijk EA, de Koning HJ. When should active surveillance for prostate cancer stop if no progression is detected? The Prostate 2017;77(9):962–9.

[32] Tomer A, Rizopoulos D, Nieboer D, Drost FJ, Roobol MJ, Steyerberg EW. Personalized decision making for biopsies in prostate cancer active surveillance programs. Medical Decision Making 2019;39(5):499–508.

[33] Balasubramanian R, Lagakos SW. Estimation of a failure time distribution based on imperfect diagnostic tests. Biometrika 2003;90(1):171–82.