Supplementary Materials for "Personalized Biopsies in Prostate Cancer Active Surveillance"

Anirudh Tomer, MSc[a,*], Daan Nieboer, MSc[b], Monique J. Roobol, PhD[c], Anders Bjartell, PhD[d], Ewout W. Steyerberg, PhD[b,e], Dimitris Rizopoulos, PhD[a], Movember Foundations 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 Foundations Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium members presented in Appendix A*

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

Let T_i^* denote the true time of reclassification (increase in biopsy Gleason grade 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 reclassification is observed for the patient at his latest biopsy time r_i, then l_i denotes the time of the second latest biopsy. Otherwise, l_i denotes the time of the latest biopsy and $r_i = \infty$. Let $\boldsymbol{y}_i$ denote

*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, PhD), e.w.steyerberg@lumc.nl (Ewout W. Steyerberg, PhD), d.rizopoulos@erasmusmc.nl (Dimitris Rizopoulos, PhD)

his observed PSA longitudinal measurements. The observed data of all n patients is denoted by $\mathcal{D}_n = \{l_i, r_i, \boldsymbol{y}_i; i = 1, \ldots, n\}$.

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 3 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 $\boldsymbol{D}$. 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 reclassification, our joint model uses a relative risk sub-model. More specifically, the hazard of reclassification denoted as $h_i(t)$, and the cumulative risk of reclassification denoted as $R_i(t)$, at a time t are (see Panel C, Figure 1):

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

where, γ is the parameter for the effect of age. The impact of PSA on the hazard of reclassification is modeled in two ways, namely the impact of the error free underlying PSA value $m_i(t)$ (see Panel A, Figure 1), and the impact of the underlying PSA velocity $\partial m_i(t)/\partial t$ (see Panel B, Figure 1).

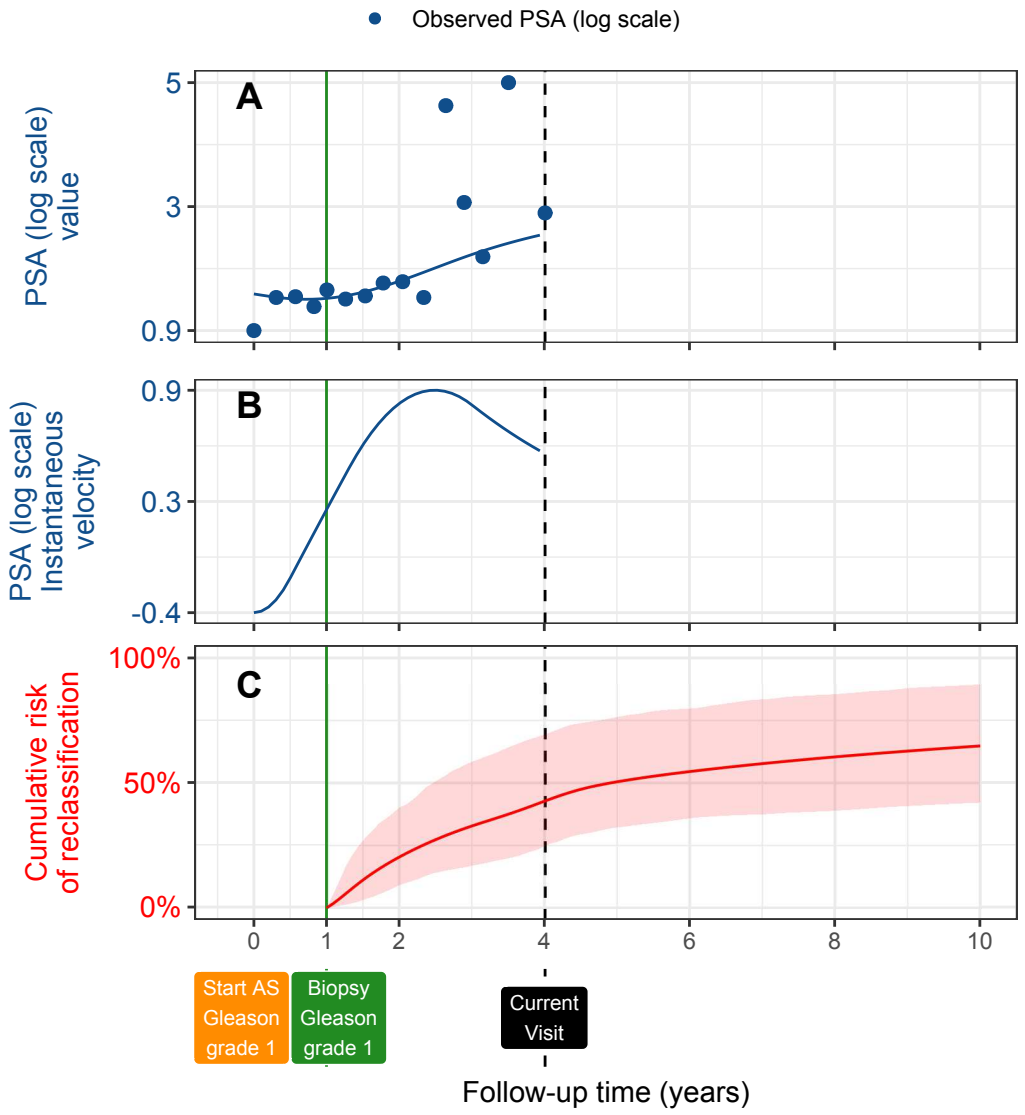

Figure 1: **Illustration of the joint model on a real PRIAS dataset patient**. **Panel A:** Observed (blue dots) and fitted PSA (solid blue line) measurements, log-transformed. **Panel B:** Estimated instantaneous velocity of PSA (log-transformed). **Panel C**: Predicted cumulative-risk of reclassification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

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

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

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

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

The likelihood contribution of the time of reclassification 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 integral in (3) does not have a closed-form solution, and therefore we use a 15-point Gauss-Kronrod quadrature rule to approximate it.

We use independent normal priors with zero mean and variance 100 for the fixed effects $\{\beta_0, \ldots, \beta_5\}$, and inverse Gamma prior with shape and rate both equal to 0.01 for the parameter σ^2. For the variance-covariance matrix $\boldsymbol{D}$ 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 were 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.

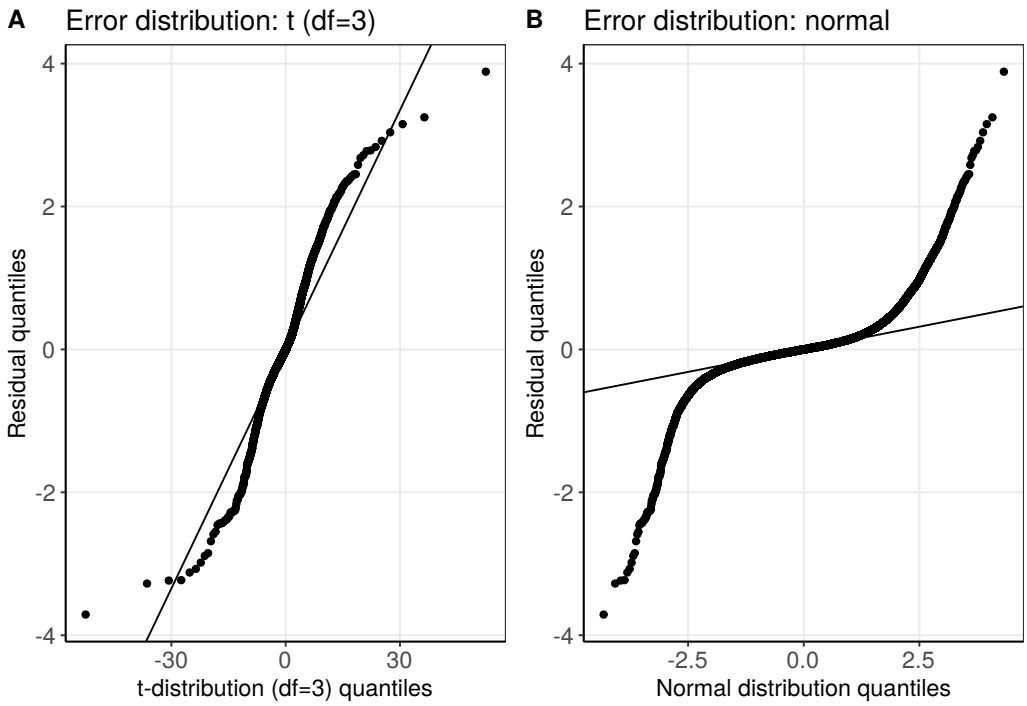

Figure 2: Quantile-quantile plot of subject-specific residuals from the joint models fitted to the PRIAS dataset. **Panel A**: model assuming a t-distribution (df=3) for the error term ε (see Equation 1). **Panel B**: model assuming a normal distribution for the error term ε.

Table 1: Estimated variance-covariance matrix $\boldsymbol{D}$ 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**

Table 2: Estimated mean and 95% credible interval for the parameters of the longitudinal sub-model (see Equation 1) for the PSA outcome.

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	

Appendix A.2. Results

The joint model was fitted using the R package **JMbayes** [8]. This package utilizes the Bayesian methodology to estimate model parameters. The corresponding posterior parameter estimates are shown in Table 2 (longitudinal sub-model for PSA outcome) and Table 3 (relative risk sub-model). The parameter estimates for the variance-covariance matrix $\boldsymbol{D}$ from the longitudinal sub-model for PSA are shown in the following Table 1:

For the PSA mixed effects sub-model parameter estimates (see Equation 1), in Table 2 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 PSA profiles for nine randomly selected patients.

For the relative risk sub-model (see Equation 2), the parameter estimates in Table 3 show that $\log_2(\text{PSA} + 1)$ velocity and age of the patient were significantly associated with the hazard of reclassification.

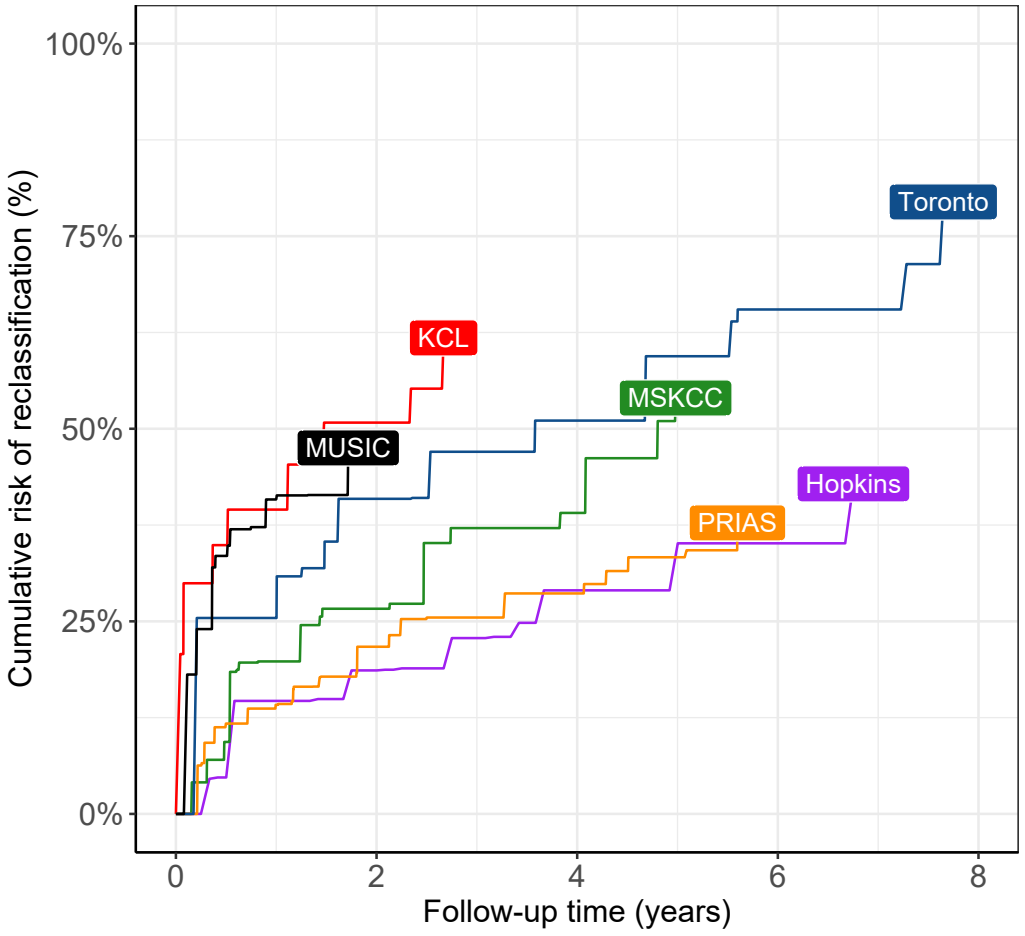

Figure 3: **Nonparametric estimate [6] of cumulative risk of reclassification** in the world's largest AS cohort PRIAS, and largest five AS cohorts from the GAP3 database [7]. Abbreviations are *Hopkins*: Johns Hopkins Active Surveillance, *PRIAS*: Prostate Cancer International Active Surveillance, *Toronto*: 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.

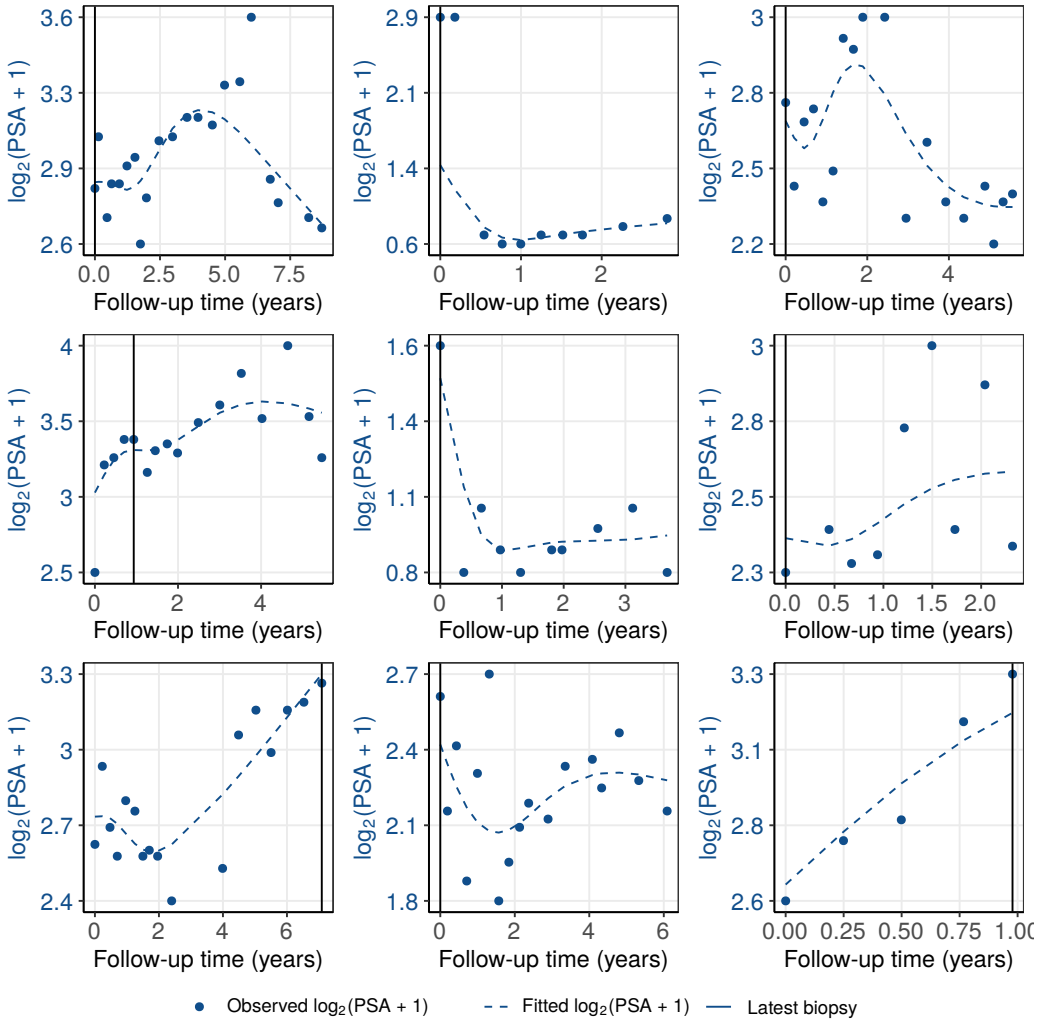

Figure 4: Fitted versus observed $\log_2(\text{PSA}+1)$ profiles for nine randomly selected PRIAS patients. The fitted profiles utilize information from the observed PSA measurements, and time of the latest biopsy.

Table 3: Estimated mean and 95% credible interval for the parameters of the relative risk sub-model (see Equation 2) of the joint model fitted to the PRIAS dataset.

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

Table 4: Hazard (of reclassification) ratio and 95% credible interval (CI), for 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]

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 4, we present the hazard ratio of reclassification, 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 for the rest is similar.

Appendix B. Risk Predictions for Reclassification

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

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

The distribution $g(T_j^*)$ depends not only depends on the observed data of the patient $T_j^* > t, \mathcal{Y}_j(s)$, but also depends on the information from the PRIAS dataset $\mathcal{D}_n$. 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. [9]. Since, majority of the prostate cancer patients may not obtain reclassification in the current follow-up period of PRIAS (thirteen years), $g(T_j^*)$ can only be estimated for time points falling within the thirteen year follow-up.

The cumulative risk of reclassification can be derived from $g(T_j^*)$ as given in [9]. It is given by:

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

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

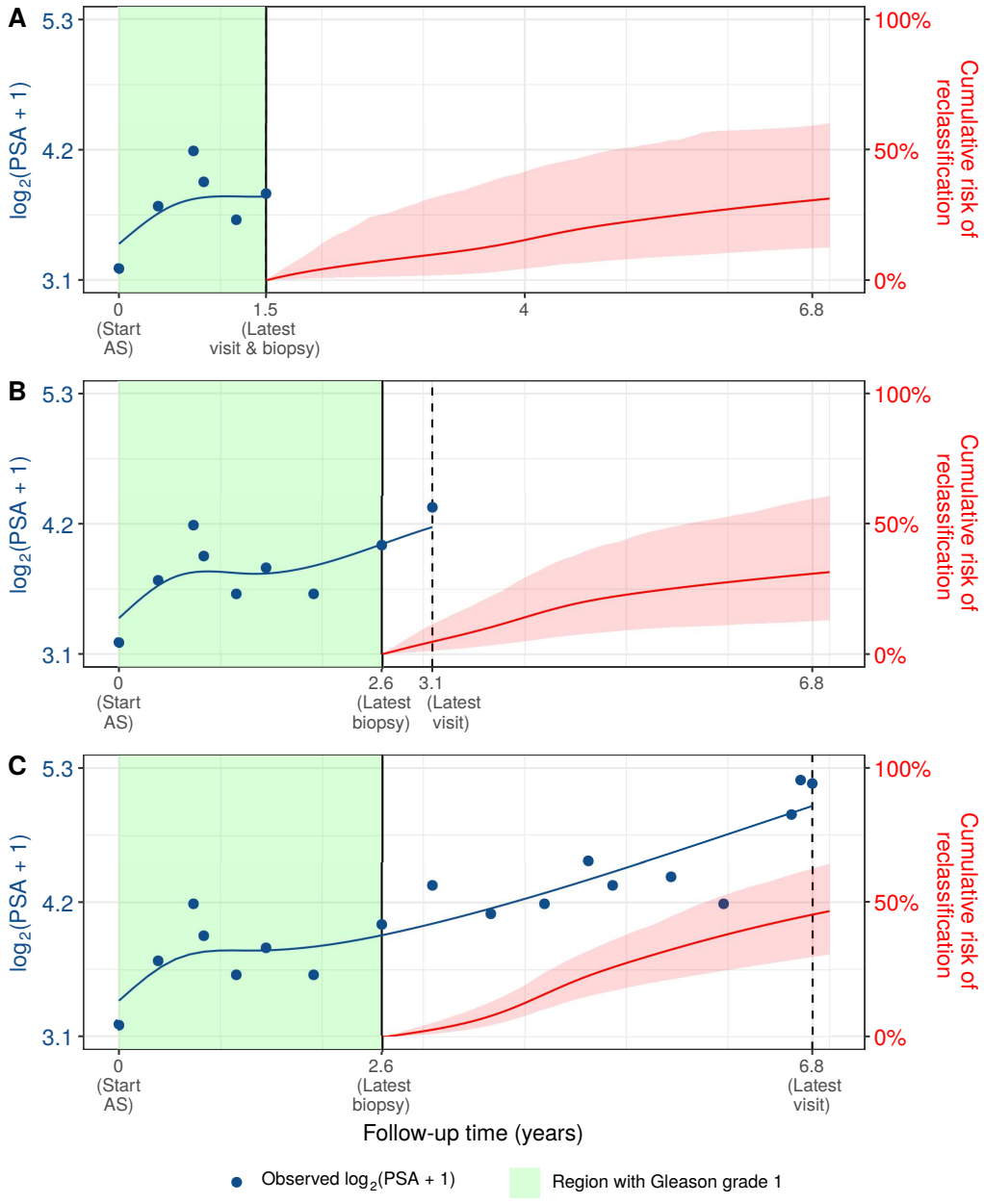

Figure 5: **Cumulative risk of (reclassification) changing dynamically over follow-up** as more patient data is gathered. The three **Panels A,B and C:** are ordered by the time of the latest visit (dashed vertical black line) of a new patient. At each of the latest follow-up visits, we combine the accumulated PSA measurements (shown in blue), and latest time of negative biopsy (solid vertical black 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 PRIAS dataset (internal validation) and in largest five cohorts from the GAP3 database [7]. 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), and Michigan Urological Surgery Improvement Collaborative AS (MUSIC).

Calibration-in-the-large We first assessed calibration-in-the-large [10] of our model in the aforementioned cohorts. To this end, we used our model to predict the cumulative risk of reclassification for each patient given their PSA measurements and biopsy results. We then averaged the resulting profiles of cumulative risk of reclassification. Subsequently we compared the averaged cumulative-risk profile with a non-parametric estimate [6] of the cumulative risk of reclassification in each of the cohorts. The results are shown in Panel A of Figure 6. We can see that our model's calibration is fine only in PRIAS and Hopkins cohorts. To improve our model's calibration in KCL, MUSIC, Toronto, and MSKCC cohorts, we recalibrated the baseline hazard of the PRIAS model individually for each of these cohorts. More specifically, given the cohort data $\mathcal{D}_{n^c}$ of the c-th cohort, the recalibrated parameters $\boldsymbol{\gamma}_{h0}^c$ (Section Appendix A) of the log baseline hazard are given by:

$$p(\boldsymbol{\gamma}_{h0}^c \mid \mathcal{D}_n^c, \boldsymbol{b^c}, \boldsymbol{\theta}) \propto \prod_{i=1}^{n^c} p(l_i^c, r_i^c \mid \boldsymbol{b_i^c}, \boldsymbol{\theta}) p(\boldsymbol{\gamma}_{h0}^c) \tag{5}$$

where n^c are the number of patients in the c-th cohort, $\boldsymbol{\theta}$ are the parameters of the joint model fitted to the PRIAS dataset, l_i^c, r_i^c are the interval in which reclassification is observed ($r_i^c = \infty$ for right censored patients) for the $i-th$ patient of the c-th cohort. The symbol $\boldsymbol{b_i^c}$ denotes his patient-specific random effects (Section Appendix A). The random effects are estimated from the original joint model fitted to the PRIAS dataset. We re-evaluated the calibration-in-the-large of our model after the recalibration of the baseline hazard. The improved calibration-in-the-large is shown in Panel B of Figure 6.

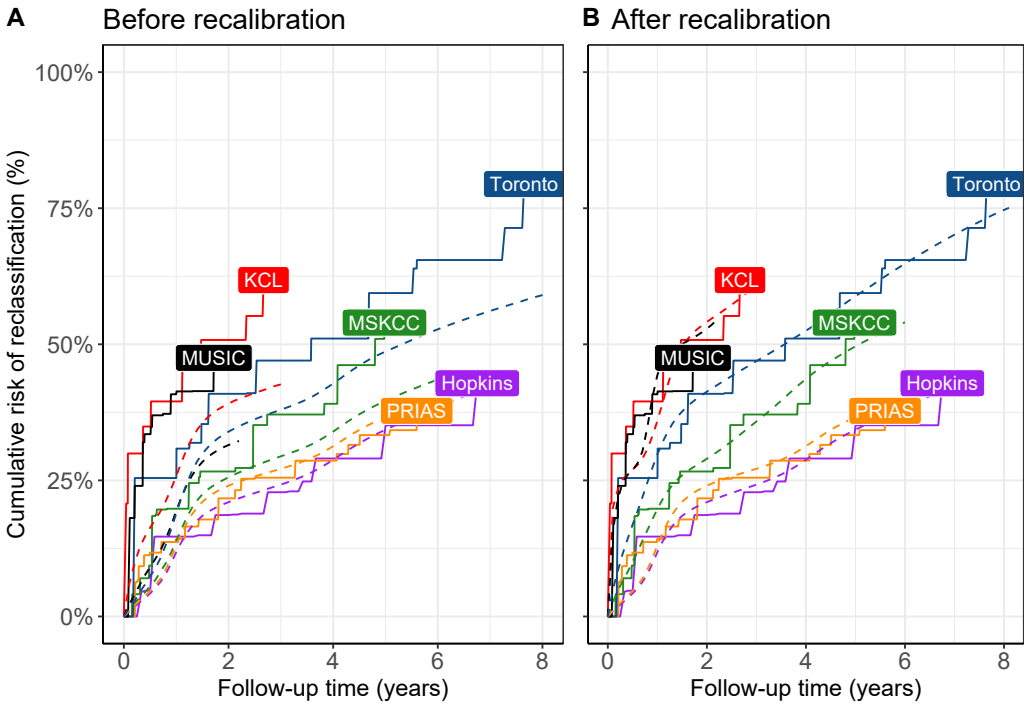

Figure 6: **Calibration-in-the-large of our model:**. In **Panel A** we can see that our model is not well calibrated for use in KCL, MUSIC, Toronto and MSKCC. In **Panel B** we can see that calibration of model predictions improved in KCL, MUSIC, Toronto and MSKCC cohorts. Recalibration was not conducted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

Recalibrated PRIAS Model Versus Individual Joint Models For Each Cohort We wanted to check if our recalibrated PRIAS model performed as well fitting an entirely new joint model to the external cohorts. To this end, we predicted cumulative-risk of reclassification for each patient from each patient using two different models, namely the recalibrated PRIAS model for that cohort, and a new joint model fitted to that cohort. The difference in predicted cumulative-risk of reclassification from these cohorts (Figure 7) is quite small. The only exception is the MUSIC cohort in which individual risk predictions obtained from the recalibrated PRIAS model may differ from predictions from a newly fitted joint model, by more than 10% in at least half of the patients.

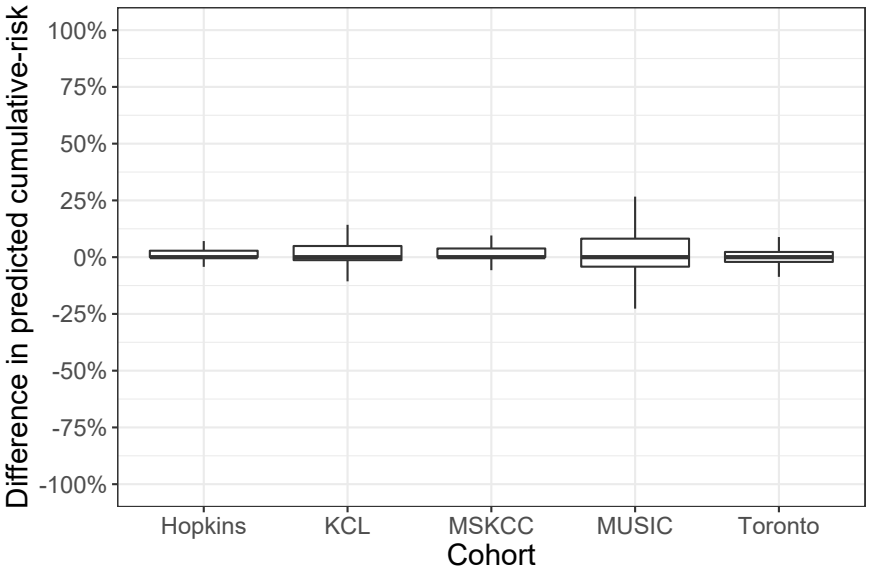

Figure 7: **Comparison of predictions from recalibrated PRIAS model with individual joint models fitted to external cohorts:** On Y-axis we show the difference between predicted cumulative-risk of reclassification for individual patients using two models, namely the recalibrated PRIAS model for each cohort, and individual joint models fitted to each cohort. The maximum differences in each direction can be 100% or -100%. The figure shows that in all cohorts except the MUSIC cohort, the recalibrated PRIAS model predicts as good as a newly fitted joint model to each of the cohorts. 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.

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 calibration of the joint model also depends on the available data. We assessed these two measures dynamically in the PRIAS cohort (interval validation) and in the largest five 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 [9]. For time-varying calibration we assessed the mean absolute prediction error or MAPE [9]. The AUC indicates how well the model discriminates between patients who experience reclassification and those do not. The MAPE indicates how well the model predicts reclassification. 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 s, we calculate AUC and MAPE for a medically relevant time frame $(t, s]$, within which the occurrence of reclassification is of interest. In the case of prostate cancer, at any point in time s it is of interest to identify patients who may have experienced reclassification in the last one year $(s-1, s]$. That is we set $t = s - 1$. We then calculate AUC and MAPE at a gap of every six months (follow-up schedule of PRIAS). That is, $s\epsilon\{1, 1.5, \ldots\}$ years. To obtain reliable estimates of AUC and MAPE, in each cohort we restrict s to a maximum time point $s_{\max}$, such that there are at least 10 patients who experience reclassification after $s_{\max}$. This maximum time point $s_{\max}$ differs between cohorts. The resulting estimates of AUC are summarized in Figure 8, and in Table 5 to Table 10. Results are based on the recalibrated PRIAS model for Toronto, MSKCC, MUSIC, and KCL cohorts, whereas original joint model fitted to the PRIAS dataset is used for Hopkins and PRIAS 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 two of follow-up. This could be because of two reasons. Firstly, MAPE at year 1 is based only on four PSA measurements gathered in first year of follow-up, whereas after year two number of PSA measurements increase. Secondly, patients in year one consist of two

sub-populations, namely patients with a correct Gleason grade 1 at the time of inclusion in AS, and patients who probably had Gleason grade 2 at inclusion but were misclassified by the urologist as Gleason grade 1 patients. To remedy this problem, a biopsy for all patients at year one is commonly recommended in all AS programs [11].

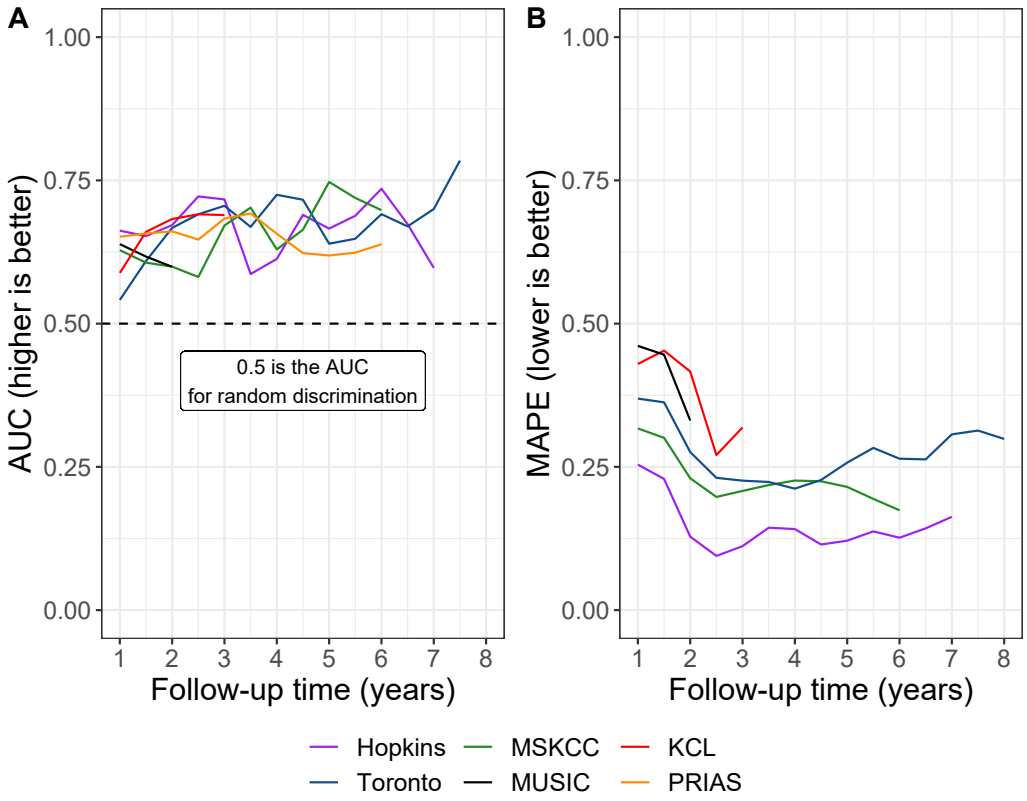

Figure 8: **Validation of Dynamic Cumulative-Risk Predictions**. In **Panel A** we can see that the time dependent area under the receiver operating characteristic curve or AUC (measure of discrimination) is above 0.5 in PRIAS (internal validation), and in Toronto, JHAS, MSKCC, KCL, and MUSIC AS cohorts (external validation). In **Panel B** we can see that the time dependent root mean squared prediction error or MAPE (measure of calibration) is similar for PRIAS, and JHAS and Toronto cohorts. The bootstrapped 95% confidence interval for these estimates are presented in Table 5 to Table 9. Full names of Cohorts are *PRIAS*: Prostate Cancer International Active Surveillance, *Toronto*: University of Toronto Active Surveillance, *JHAS*: 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.

Table 5: **Internal validation of predictions of reclassification in PRIAS cohort**. The area under the receiver operating characteristic curve or AUC (measure of discrimination) and mean absolute prediction error or MAPE (measure of calibration) 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)
0.0 to 1.0	0.652 [0.611, 0.690]	0.227 [0.223, 0.236]
0.5 to 1.5	0.657 [0.641, 0.673]	0.376 [0.371, 0.382]
1.0 to 2.0	0.661 [0.647, 0.678]	0.371 [0.364, 0.379]
1.5 to 2.5	0.647 [0.596, 0.688]	0.253 [0.245, 0.263]
2.0 to 3.0	0.683 [0.642, 0.723]	0.252 [0.241, 0.262]
2.5 to 3.5	0.692 [0.632, 0.748]	0.238 [0.224, 0.251]
3.0 to 4.0	0.657 [0.603, 0.709]	0.273 [0.263, 0.285]
3.5 to 4.5	0.623 [0.582, 0.660]	0.338 [0.326, 0.349]
4.0 to 5.0	0.619 [0.582, 0.654]	0.338 [0.325, 0.350]
4.5 to 5.5	0.624 [0.537, 0.711]	0.397 [0.355, 0.425]
5.0 to 6.0	0.639 [0.582, 0.696]	0.397 [0.355, 0.425]

Table 6: **External validation of predictions of reclassification 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 (measure of calibration) 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)
0.0 to 1.0	0.541 [0.470, 0.621]	0.369 [0.352, 0.381]
0.5 to 1.5	0.609 [0.547, 0.661]	0.363 [0.348, 0.376]
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 7: **External validation of predictions of reclassification 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 (measure of calibration) 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)
0.0 to 1.0	0.662 [0.586, 0.715]	0.254 [0.245, 0.265]
0.5 to 1.5	0.653 [0.603, 0.707]	0.229 [0.219, 0.240]
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 8: **External validation of predictions of reclassification 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 (measure of calibration) 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)
0.0 to 1.0	0.628 [0.577, 0.688]	0.317 [0.316, 0.318]
0.5 to 1.5	0.606 [0.532, 0.657]	0.301 [0.290, 0.311]
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 9: **External validation of predictions of reclassification 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 (measure of calibration) 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)
0.0 to 1.0	0.589 [0.514, 0.653]	0.430 [0.407, 0.450]
0.5 to 1.5	0.660 [0.550, 0.742]	0.453 [0.431, 0.474]
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 10: **External validation of predictions of reclassification 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 (measure of calibration) 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)
0.0 to 1.0	0.639 [0.607, 0.672]	0.461 [0.450, 0.469]
0.5 to 1.5	0.617 [0.588, 0.652]	0.446 [0.441, 0.453]
1.0 to 2.0	0.599 [0.553, 0.632]	0.331 [0.317, 0.348]

Appendix C. Personalized Biopsies Based on Risk of GS7

Consider some real patients from the PRIAS database shown in Figure 9 to Figure 12. We intend to develop personalized schedule of biopsies for these patients. Using the joint model fitted to the PRIAS dataset, we first obtain their cumulative risk of GS7 over the entire follow-up period (see Equation 4). This cumulative risk accounts for their entire history of PSA as well as the time of their latest negative biopsy. For a new patient j we suggest a personalized risk based biopsy at time s if their cumulative risk of GS7 denoted by $R_j(s \mid t, s)$ at s, given the time of their latest negative biopsy t, is above a certain threshold (e.g., 10% risk). Suppose that in this way a decision of biopsy is taken at time s. Since patients may be removed from AS upon detection of GS7, schedule of future biopsies is made by assuming that GS7 is not detected at time s. Thus, for a decision of biopsy at the next visit time $s+1$, the cumulative risk of GS7 denoted by $R_j(s+1 \mid s, s)$ that the time of latest negative biopsy is s. Similarly, if $R_j(s+1 \mid s, s) < 10\%$, then we decide for a biopsy at a subsequent time $s+2$ using the threshold $R_j(s+2 \mid s, s)$. On the other hand if $R_j(s+1 \mid s, s) \geq 10\%$ then then we decide for a biopsy at time $s+2$ using the threshold $R_j(s+2 \mid s+1, s)$. While scheduling these biopsies we always maintain a minimum gap of one year. Personalized schedules can also be made with any other risk threshold such as 5% or 15%.

To assist patients in making an informed choice for a schedule, be it personalized or fixed, we provide them patient-specific consequences of following each schedule. To this end, we first calculate the probability of occurrence of GS7 between successive biopsies of each schedule. Using these probabilities we then obtain the expected delay in detection of GS7 for following that schedule. Thus, patients have a method to compare across various schedules in terms of the personalized burden (time and total biopsies), and personalized benefit (less delay in detection of GS7 is beneficial). Suppose once again that for patient j, the time of latest negative biopsy is t, and current visit time is $s > t$. Then equation for the expected delay $D_j(\mathcal{S} \mid t, s)$ in detection of GS7 using schedule of biopsies $\mathcal{S} = \{t_1, \ldots, t_h\}$, where $t_1 \geq s$, and t_h is

the horizon time up to which we want to schedule biopsies, is given by:

$$D_j(\mathcal{S} \mid t, s) = \sum_{v=1}^{h-1} \Big\{ R_j(t_{v+1} \mid t, s) - R_j(t_v \mid t, s) \Big\} \times \left\{ t_{v+1} - t_v - \int_{t_v}^{t_{v+1}} \frac{R_j(t_{v+1} \mid t, s) - R_j(u \mid t, s)}{R_j(t_{v+1} \mid t, s) - R_j(t_v \mid t, s)} \mathrm{d}u \right\} \tag{6}$$

The personalized and fixed schedules, and their consequences for a few real patients from the PRIAS dataset are shown in Figure 9 to Figure 12. We maintained a minimum gap of one year between biopsies as advised by the PRIAS protocol. In addition, we scheduled biopsies only for the first ten years follow-up because of limited follow-up data period of PRIAS. A compulsory biopsy was done at year ten of follow-up in all schedules for meaningful comparison of their expected delays in detection of GS7.

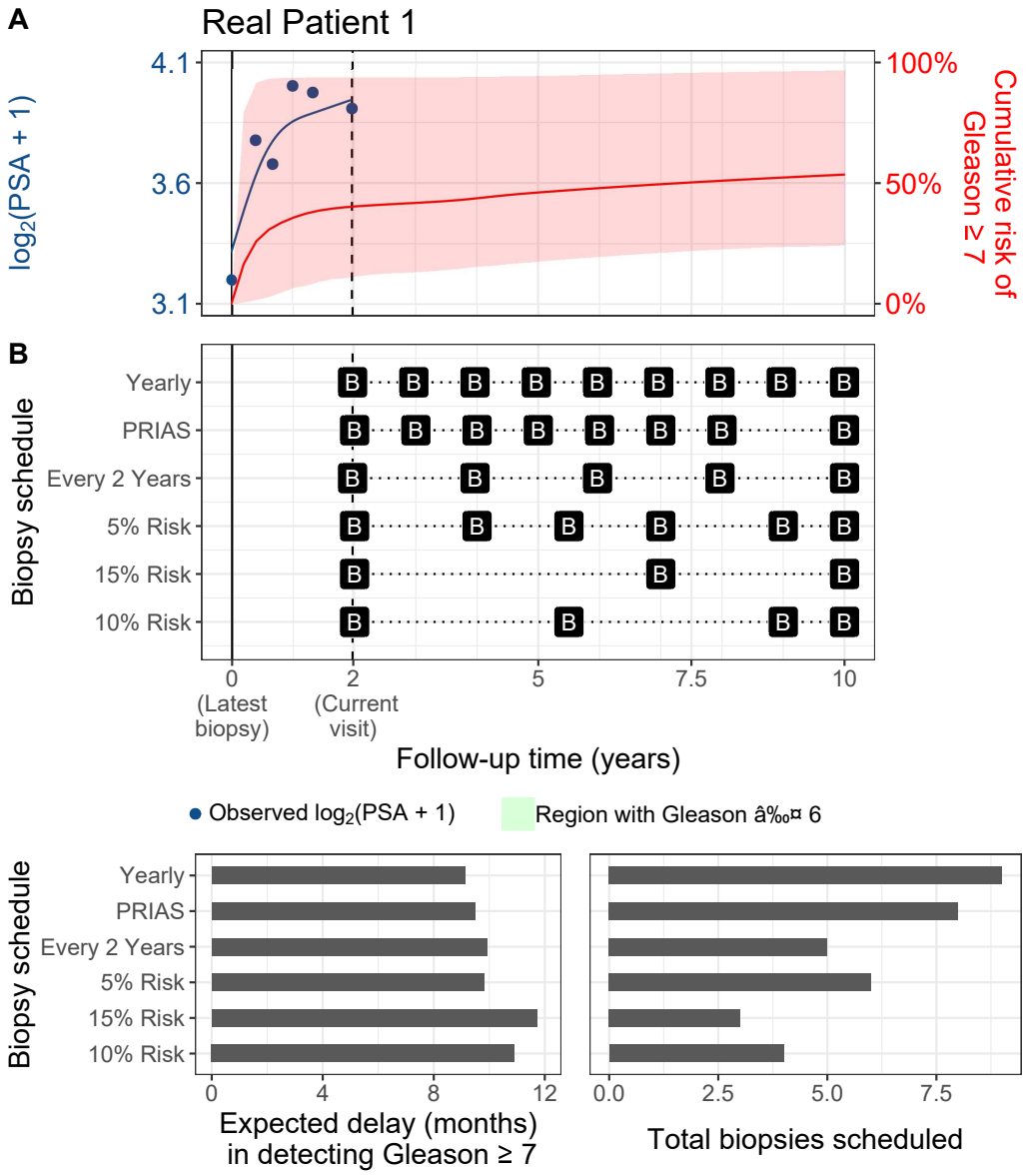

Figure 9: **Personalized and fixed schedules of biopsies for patient 1**. **Panel A:** shows the observed and fitted $\log_2(\text{PSA}+1)$ measurements (Equation 1), and the dynamic cumulative risk of Gleason ≥ 7 (see Appendix B) over follow-up period. **Panel B** shows the personalized and fixed schedules of biopsies with a 'B' indicating times of biopsies. In the bottom two panels, the various schedules are compared in terms of the number of biopsies they schedule, and the expected delay in detection of Gleason ≥ 7 if they are followed.

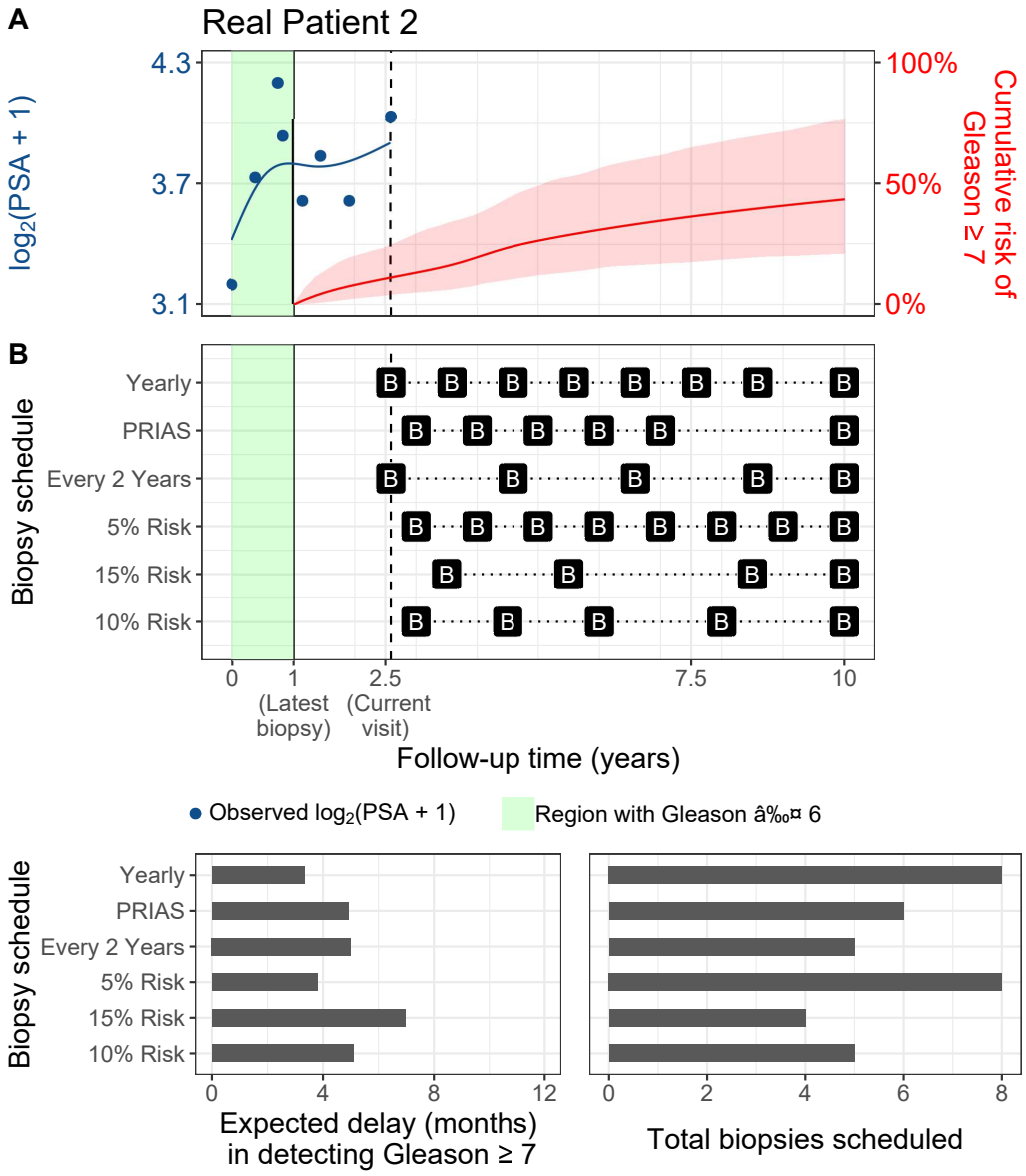

Figure 10: **Personalized and fixed schedules of biopsies for patient 2**. **Panel A:** shows the observed and fitted $\log_2(\text{PSA}+1)$ measurements (Equation 1), and the dynamic cumulative risk of Gleason ≥ 7 (see Appendix B) over follow-up period. **Panel B** shows the personalized and fixed schedules of biopsies with a 'B' indicating times of biopsies. In the bottom two panels, the various schedules are compared in terms of the number of biopsies they schedule, and the expected delay in detection of Gleason ≥ 7 if they are followed.

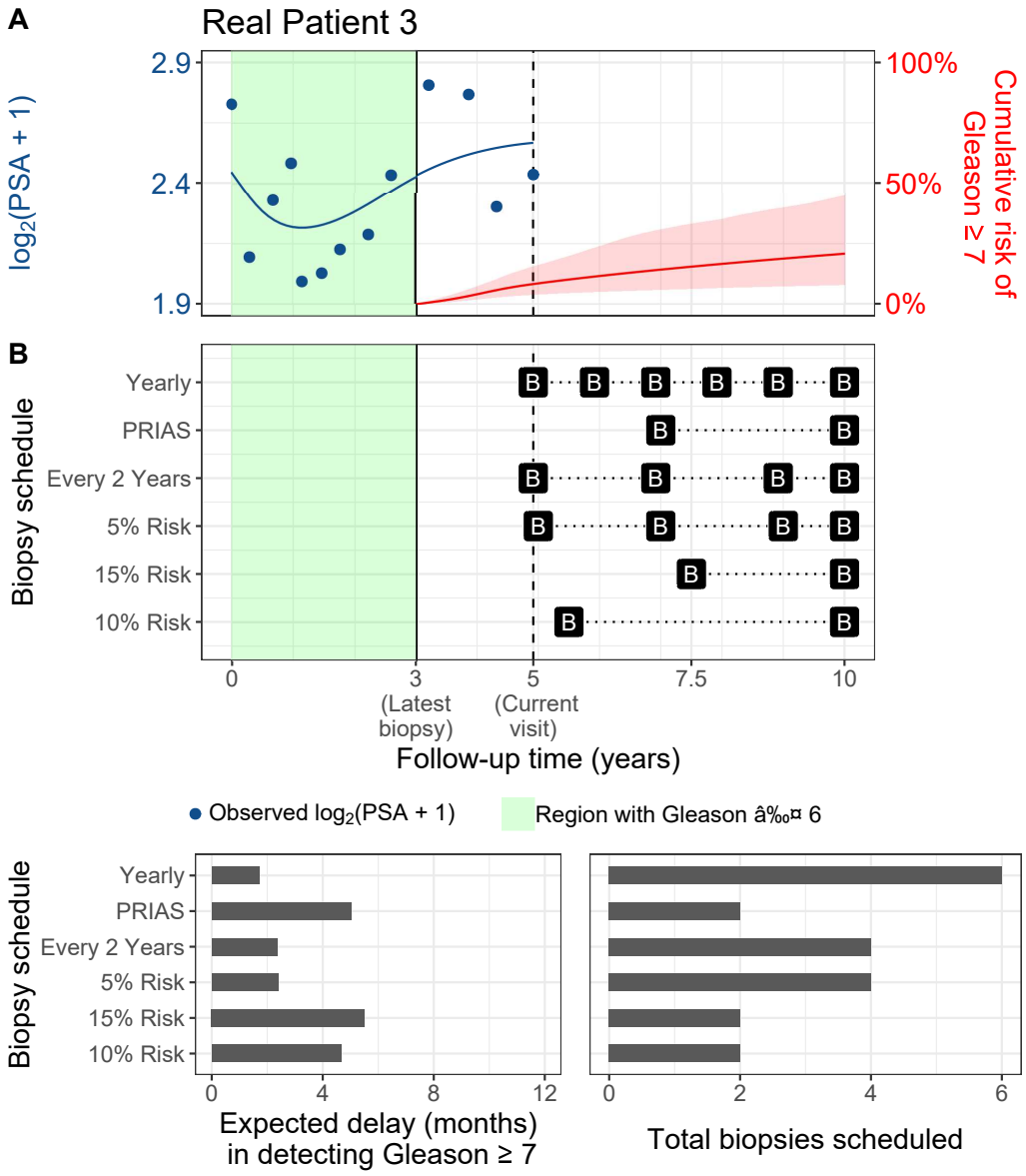

Figure 11: **Personalized and fixed schedules of biopsies for patient 3**. **Panel A:** shows the observed and fitted $\log_2(\text{PSA}+1)$ measurements (Equation 1), and the dynamic cumulative risk of Gleason ≥ 7 (see Appendix B) over follow-up period. **Panel B** shows the personalized and fixed schedules of biopsies with a 'B' indicating times of biopsies. In the bottom two panels, the various schedules are compared in terms of the number of biopsies they schedule, and the expected delay in detection of Gleason ≥ 7 if they are followed.

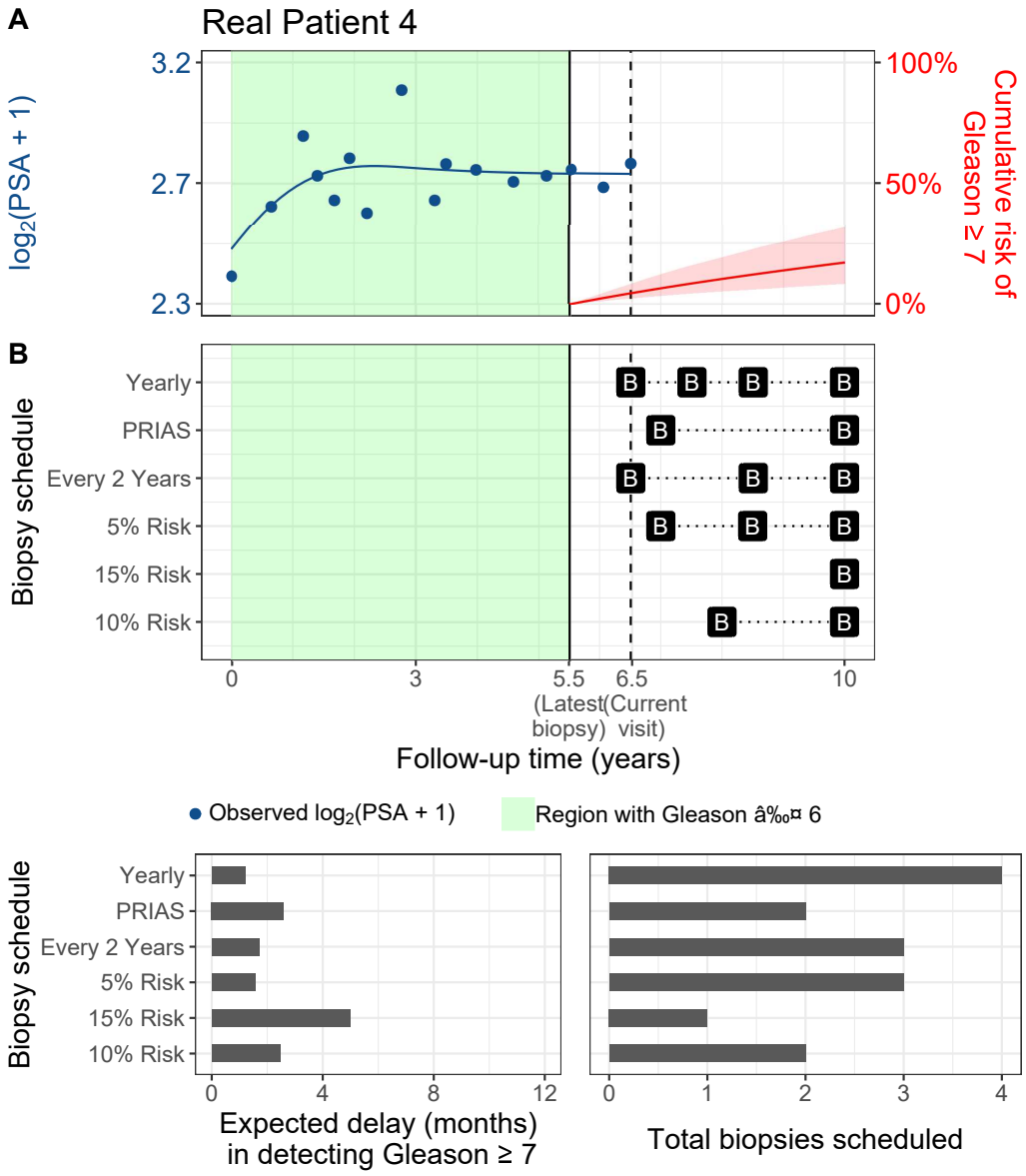

Figure 12: **Personalized and fixed schedules of biopsies for patient 4**. **Panel A:** shows the observed and fitted $\log_2(\text{PSA}+1)$ measurements (Equation 1), and the dynamic cumulative risk of Gleason ≥ 7 (see Appendix B) over follow-up period. **Panel B** shows the personalized and fixed schedules of biopsies with a 'B' indicating times of biopsies. In the bottom two panels, the various schedules are compared in terms of the number of biopsies they schedule, and the expected delay in detection of Gleason ≥ 7 if they are followed.

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. It is hosted at `https://emcbiostatistics.shinyapps.io/prias_biopsy_recommender/`.

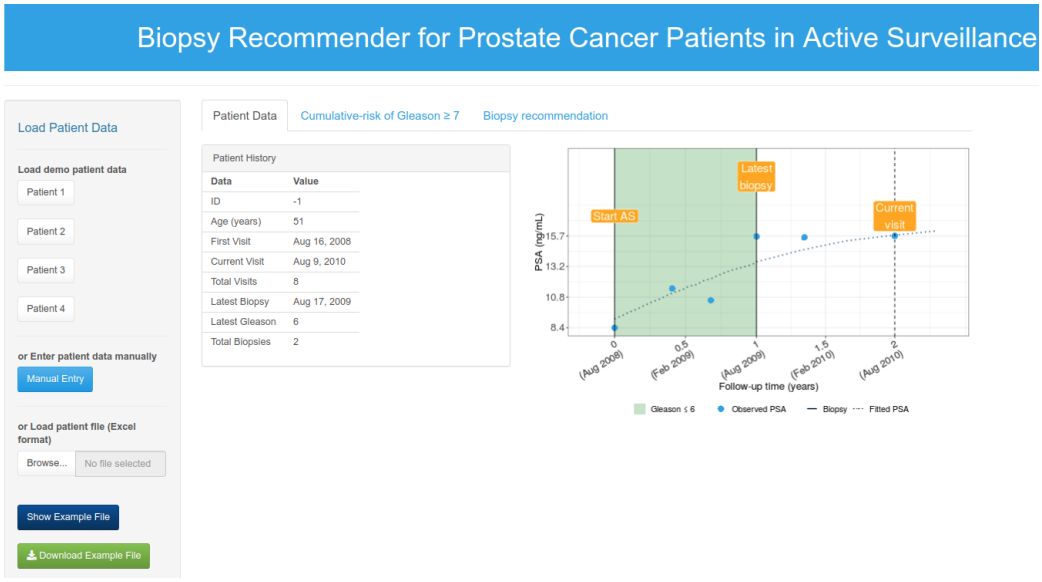

Figure 13: Landing page of the web-application. Panel on the left allows users to load patient data and panel on the right provides information. Patient data can be entered manually, or via Excel files. In addition, demo patient data is already uploaded to assist users in understanding the web-application.

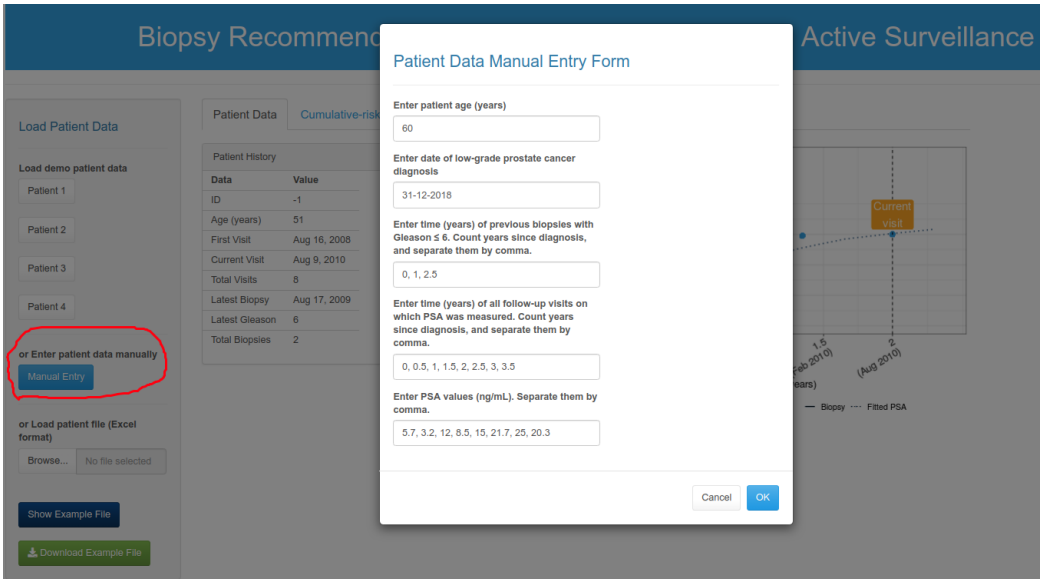

Figure 14: Patient data can be entered manually.

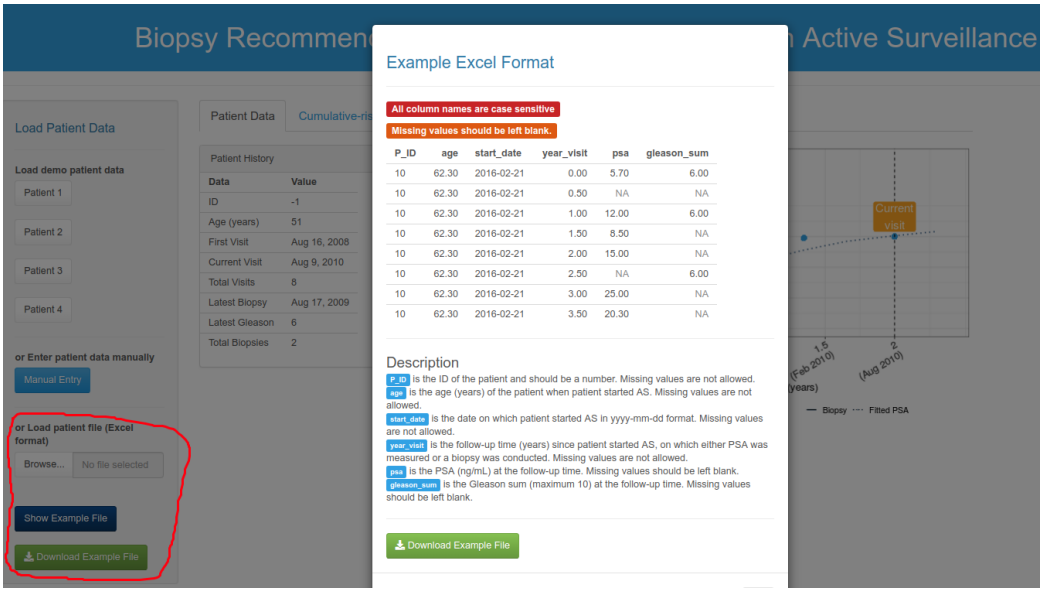

Figure 15: Patient data can be uploaded via Excel sheets. Example Excel sheet format is provided within the web-application. In addition, users can download an Excel template to fill patient data.

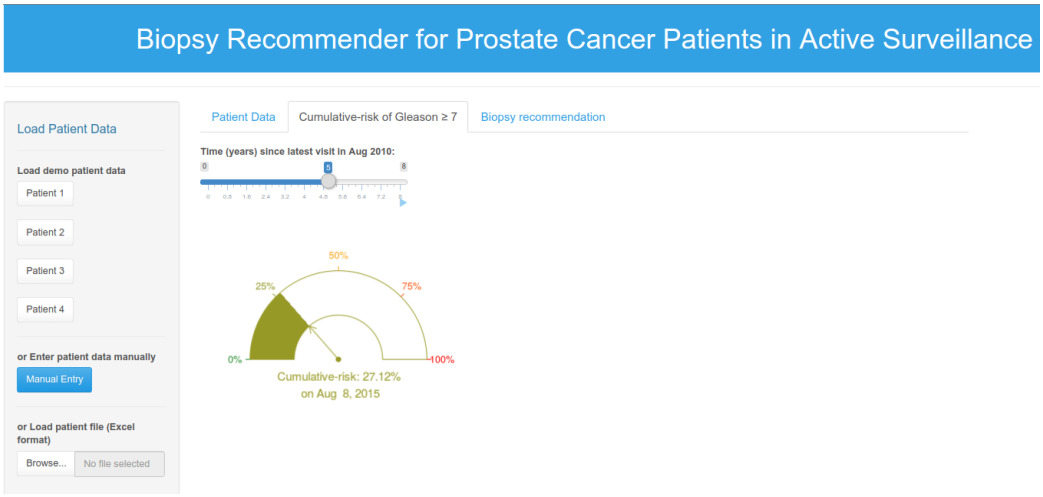

Figure 16: Second tab panel provides patient's personalized cumulative-risk of Gleason $\geq$ 7 since his latest biopsy.

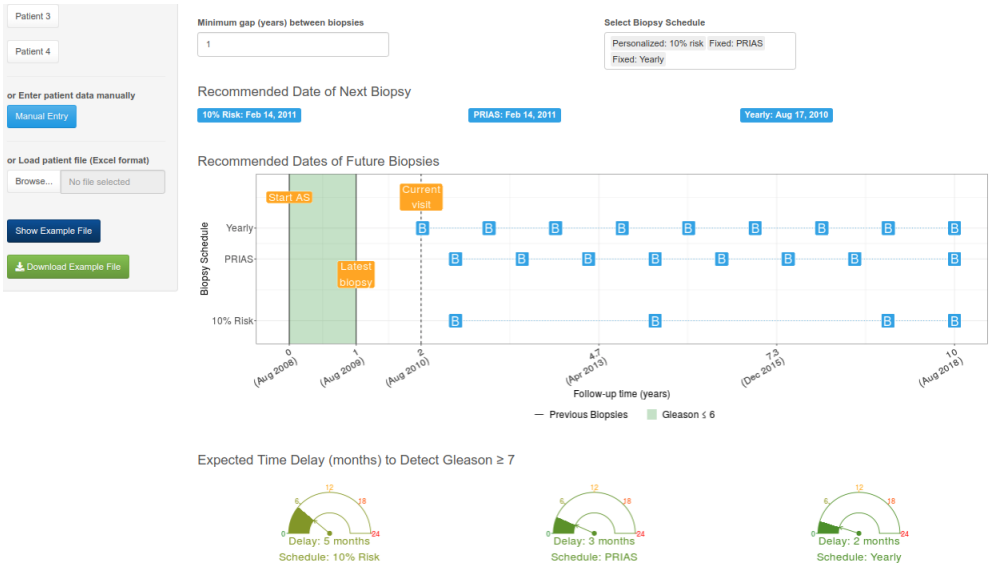

Figure 17: Third tab panel provides personalized and fixed biopsy schedule options as well as the expected time delay in detection of Gleason $\geq$ 7 for each of the schedules.

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 `www.prias-project.org`.

Formatting the dataset: This dataset however is in the so-called wide format and also requires 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 time of reclassification 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 reclassification. 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, are 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 4 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 Reclassification

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 risk of reclassification for new patients can be developed using the script `R_HOME/scheduleCreator.R`. This script also provides fixed biopsy schedules for the patients. In addition with each schedule, the expected delay in detection of reclassification is also provided.

Appendix E.4. Source Code for Web Application

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

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. 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.

7. 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.

8. 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.

9. 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.

10. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction

263 models: a framework for some traditional and novel measures. *Epidemi-*
264 *ology (Cambridge, Mass)* 2010;21(1):128.

265 11. Bokhorst LP, Alberts AR, Rannikko A, Valdagni R, Pickles T, Kakehi Y,
266 Bangma CH, Roobol MJ, PRIAS study group . Compliance rates with
267 the Prostate Cancer Research International Active Surveillance (PRIAS)
268 protocol and disease reclassification in noncompliers. *European Urology*
269 2015;68(5):814–21.