

# Supplementary Materials for “Risk of Upgrading Based Personalized Biopsy Schedules for Prostate Cancer Active Surveillance Patients”

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

<sup>a</sup>*Department of Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands*

<sup>b</sup>*Department of Public Health, Erasmus University Medical Center, Rotterdam, the Netherlands*

<sup>c</sup>*Department of Urology, Erasmus University Medical Center, Rotterdam, the Netherlands*

<sup>d</sup>*Department of Urology, Skåne University Hospital, Malmö, Sweden*

<sup>e</sup>*Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands*

<sup>f</sup>*The Movember Foundation’s Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium members presented in Appendix F*

---

## 1 **Appendix A. A Joint Model for the Longitudinal PSA, and Time** 2 **to Gleason Upgrading**

3 Let  $T_i^*$  denote the true time of upgrading (increase in biopsy Gleason  
4 grade group from 1 to 2 or higher) for the  $i$ -th patient included in PRIAS.  
5 Since biopsies are conducted periodically,  $T_i^*$  is observed with interval cen-  
6 soring  $l_i < T_i^* \leq r_i$ . 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](mailto:a.tomer@erasmusmc.nl) (Anirudh Tomer, MSc),  
[d.nieboer@erasmusmc.nl](mailto:d.nieboer@erasmusmc.nl) (Daan Nieboer, MSc), [m.roobol@erasmusmc.nl](mailto:m.roobol@erasmusmc.nl) (Monique J. Roobol, PhD), [anders.bjartell@med.lu.se](mailto:anders.bjartell@med.lu.se) (Anders Bjartell, MD, PhD),  
[e.w.steyerberg@lumc.nl](mailto:e.w.steyerberg@lumc.nl) (Ewout W. Steyerberg, PhD), [d.rizopoulos@erasmusmc.nl](mailto:d.rizopoulos@erasmusmc.nl) (Dimitris Rizopoulos, PhD)

7 biopsy time  $r_i$ , then  $l_i$  denotes the time of the second latest biopsy. Oth-  
 8 erwise,  $l_i$  denotes the time of the latest biopsy and  $r_i = \infty$ . Let  $\mathbf{y}_i$  denote  
 9 his observed PSA longitudinal measurements. The observed data of all  $n$   
 10 patients is denoted by  $\mathcal{A}_n = \{l_i, r_i, \mathbf{y}_i; i = 1, \dots, n\}$ .

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):

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

11 where,  $m_i(t)$  denotes the measurement error free value of  $\log_2(\text{PSA}+1)$  trans-  
 12 formed [2, 3] measurements at time  $t$ . We model it non-linearly over time us-  
 13 ing B-splines [4]. To this end, our B-spline basis function  $B_k\{(t-2)/2, (\mathcal{K}-2)/2\}$   
 14 has three internal knots at  $\mathcal{K} = \{0.5, 1.3, 3\}$  years, which are the three quar-  
 15 tiles of the observed follow-up times. The boundary knots of the spline are  
 16 at 0 and 6.3 years (95-th percentile of the observed follow-up times). We  
 17 mean centered (mean 2 years) and standardized (standard deviation 2 years)  
 18 the follow-up time  $t$  and the knots of the B-spline  $\mathcal{K}$  during parameter esti-  
 19 mation for better convergence. The fixed effect parameters are denoted by  
 20  $\{\beta_0, \dots, \beta_5\}$ , and  $\{b_{0i}, \dots, b_{4i}\}$  are the patient specific random effects. The  
 21 random effects follow a multivariate normal distribution with mean zero and  
 22 variance-covariance matrix  $\mathbf{W}$ . The error  $\varepsilon_i(t)$  is assumed to be t-distributed  
 23 with three degrees of freedom (see Appendix B.1) and scale  $\sigma$ , and is inde-  
 24 pendent 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):

$$\begin{aligned} h_i(t) &= h_0(t) \exp \left( \gamma \text{age}_i + \alpha_1 m_i(t) + \alpha_2 \frac{dm_i(t)}{dt} \right), \\ R_i(t) &= \exp \left\{ - \int_0^t h_i(s) ds \right\}, \end{aligned} \quad (2)$$

where,  $\gamma$  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

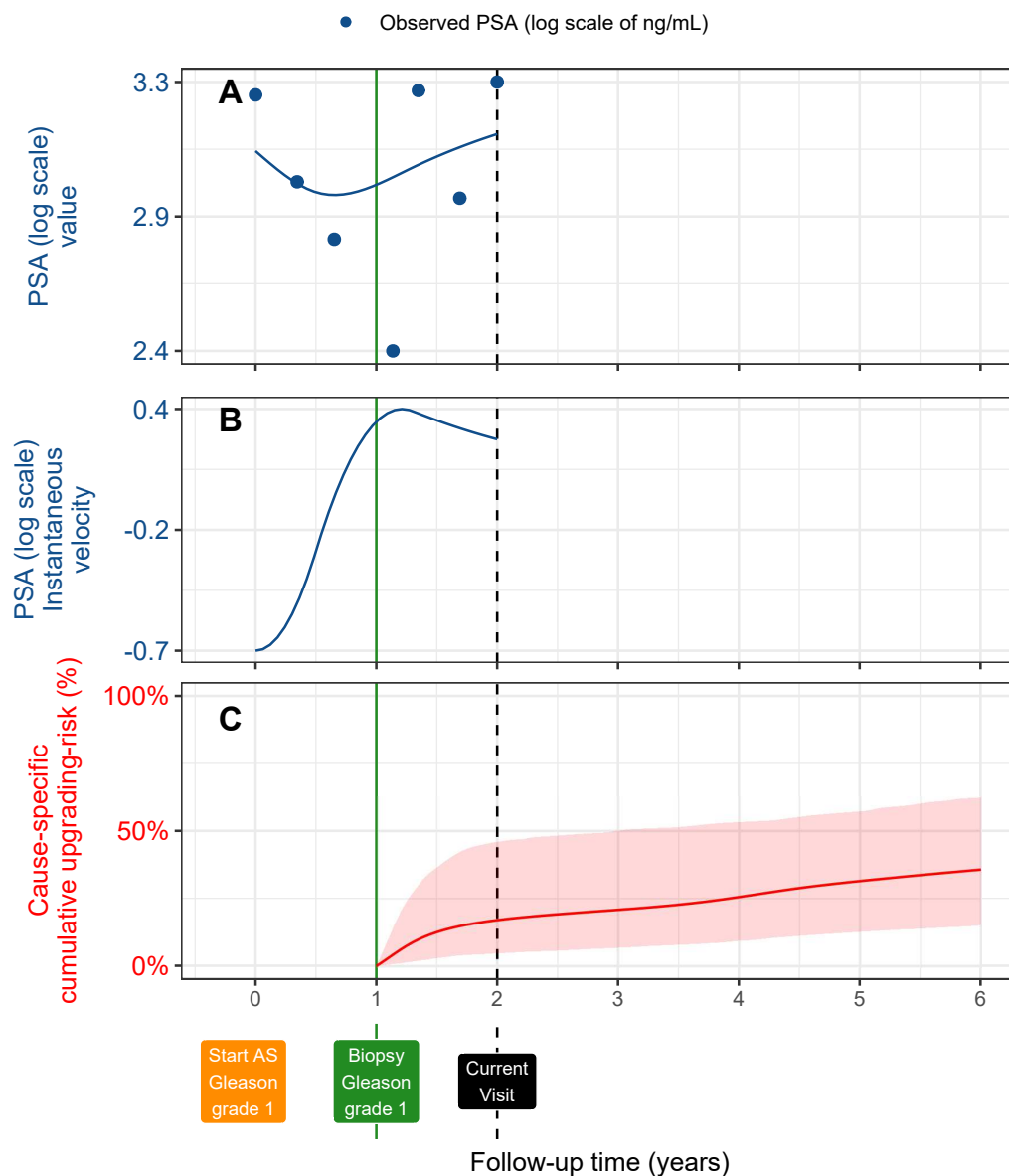


Figure 1: **Illustration of the joint model on a real PRIAS dataset patient.** **Panel A:** Observed (blue dots) and fitted PSA (solid blue line) measurements, log-transformed. **Panel B:** Estimated instantaneous velocity of PSA (log-transformed). **Panel C:** Predicted cumulative-risk of upgrading (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  $dm_i(t)/dt$  (see Panel B, Figure 1). The corresponding parameters are  $\alpha_1$  and  $\alpha_2$ , respectively. Lastly,  $h_0(t)$  is the baseline hazard at time  $t$ , and is modeled flexibly using P-splines [5]. More specifically:

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

where  $B_q(t, \mathbf{v})$  denotes the  $q$ -th basis function of a B-spline with knots  $\mathbf{v} = v_1, \dots, v_Q$  and vector of spline coefficients  $\gamma_{h_0}$ . 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  $\gamma_{h_0}$  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:

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

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

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

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:

$$p(l_i, r_i \mid \mathbf{b}_i, \boldsymbol{\theta}) = \exp \left\{ -\int_0^{l_i} h_i(s) ds \right\} - \exp \left\{ -\int_0^{r_i} h_i(s) ds \right\}. \quad (3)$$

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

32 We use independent normal priors with zero mean and variance 100 for  
 33 the fixed effects  $\{\beta_0, \dots, \beta_5\}$ , and inverse Gamma prior with shape and rate  
 34 both equal to 0.01 for the parameter  $\sigma^2$ . For the variance-covariance matrix  
 35  $\mathbf{W}$  of the random effects, we take inverse Wishart prior with an identity scale  
 36 matrix and degrees of freedom equal to 5 (number of random effects). For  
 37 the relative risk model's parameter  $\gamma$  and the association parameters  $\alpha_1, \alpha_2$ ,  
 38 we use independent normal priors with zero mean and variance 100.

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

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

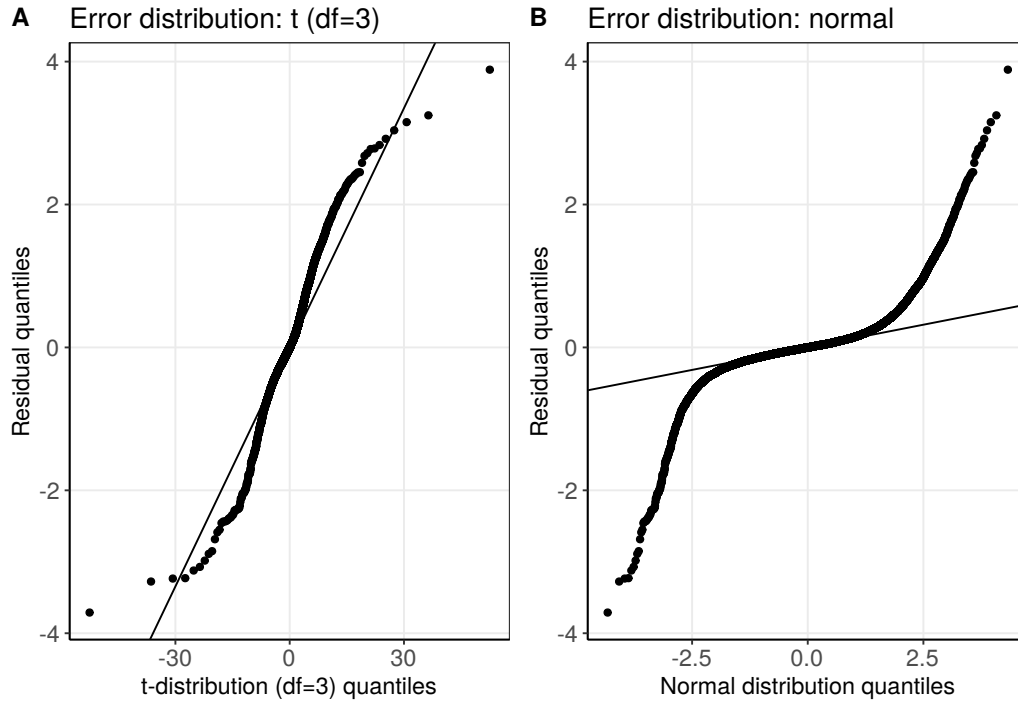


Figure 2: **Quantile-quantile plot** of subject-specific PSA residuals from two different joint models fitted to the PRIAS dataset. **Panel A:** model assuming a t-distribution ( $df=3$ ) for the error term  $\varepsilon$  (see Equation 1). **Panel B:** model assuming a normal distribution for the error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

49 *Appendix A.2. Results*

50 Characteristics of the six validation cohorts from the GAP3 database [6]  
 51 are shown in Table 1, Table 2, and Table 3. The cause-specific cumulative  
 52 upgrading-risk in these cohorts is shown in Figure 3.

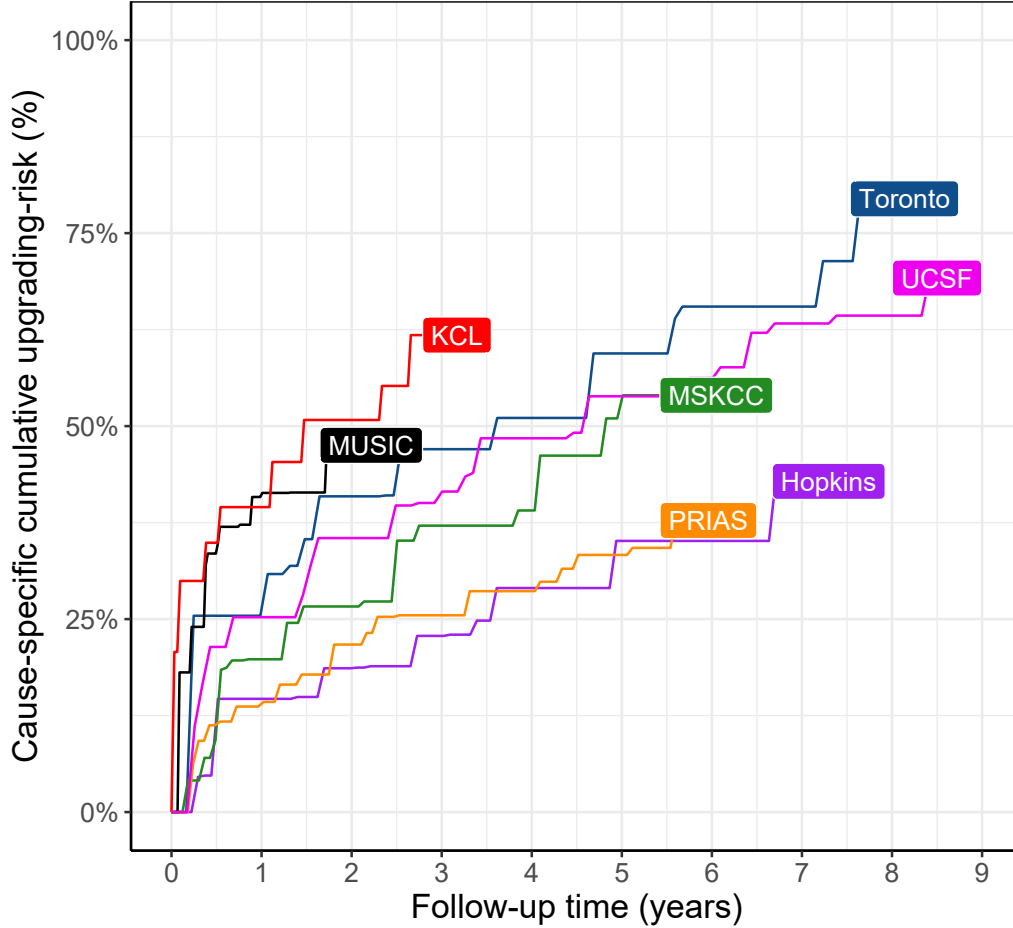


Figure 3: **Nonparametric estimate [7] of the cause-specific cumulative upgrading-risk** in the world’s largest AS cohort PRIAS, and largest six AS cohorts from the GAP3 database [6]. Abbreviations are *Hopkins*: Johns Hopkins Active Surveillance, *PRIAS*: Prostate Cancer International Active Surveillance, *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 [6]**. 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 [6]**. 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 [6].** 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  $\mathbf{W}$**  of the random effects  $\mathbf{b} = (b_0, b_1, b_2, b_3, b_4)$  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	$b_0$	$b_1$	$b_2$	$b_3$	$b_4$
$b_0$	<b>0.229</b>	0.030	0.023	0.073	0.007
$b_1$	0.030	<b>0.149</b>	0.098	0.171	0.085
$b_2$	0.023	0.098	<b>0.276</b>	0.335	0.236
$b_3$	0.073	0.171	0.335	<b>0.560</b>	0.359
$b_4$	0.007	0.085	0.236	0.359	<b>0.351</b>

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 5 (longitudinal sub-model for PSA outcome) and Table 6 (relative risk sub-model). The parameter estimates for the variance-covariance matrix  $\mathbf{W}$  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
$\sigma$	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

65 PSA profiles for nine randomly selected patients.

66 For the relative risk sub-model (see Equation 2), the parameter estimates  
 67 in Table 6 show that  $\log_2(\text{PSA} + 1)$  velocity and age of the patient were  
 68 significantly associated with the hazard of upgrading.

69 It is important to note that since age, and  $\log_2(\text{PSA} + 1)$  value and ve-  
 70 locity are all measured on different scales, a comparison between the cor-  
 71 responding parameter estimates is not easy. To this end, in Table 7, we  
 72 present the hazard ratio of upgrading, for an increase in the aforementioned  
 73 variables from their 25-th to the 75-th percentile. For example, an increase  
 74 in fitted  $\log_2(\text{PSA} + 1)$  velocity from -0.085 to 0.308 (fitted 25-th and 75-th  
 75 percentiles) corresponds to a hazard ratio of 2.433. The interpretation of the  
 76 rest is similar.

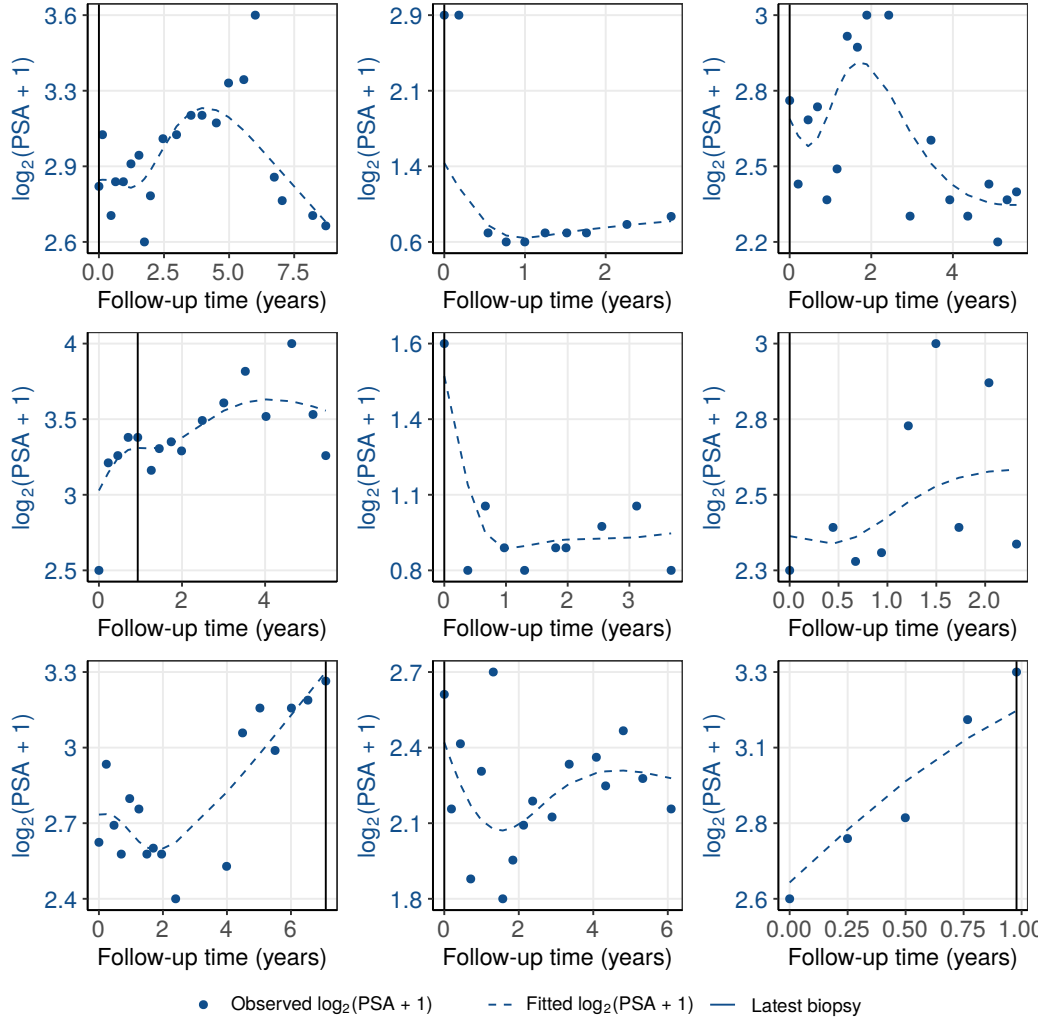


Figure 4: **Fitted versus observed  $\log_2(\text{PSA} + 1)$  profiles** for nine randomly selected PRIAS patients. The fitted profiles utilize information from the observed PSA measurements, and time of the latest biopsy.

Table 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]

## 77 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:

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

78 The distribution  $g(T_j^*)$  depends not only depends on the observed data of the  
 79 patient  $T_j^* > t, \mathcal{Y}_j(v)$ , but also depends on the information from the PRIAS  
 80 dataset  $\mathcal{A}_n$ . To this the the posterior distribution of random effects  $\mathbf{b}_j$  and  
 81 posterior distribution of the vector of all parameters  $\boldsymbol{\theta}$  are utilized, respec-  
 82 tively. The distribution  $g(T_j^*)$  can be estimated as detailed in Rizopoulos  
 83 et al. [9]. Since, many prostate cancer patients may not obtain upgrading  
 84 in the current follow-up period of PRIAS,  $g(T_j^*)$  can only be estimated for a  
 85 currently limited follow-up period.

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

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

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

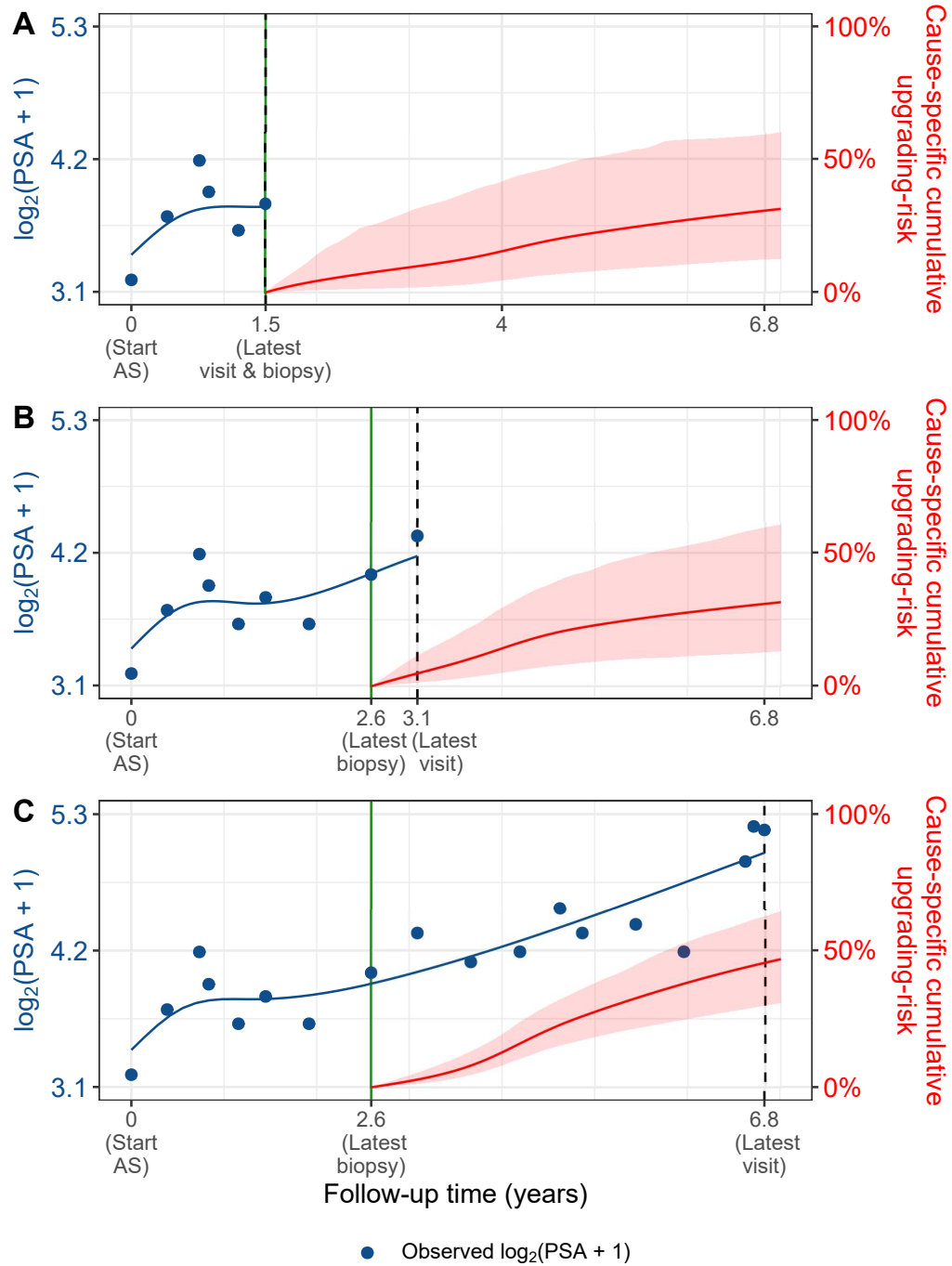


Figure 5: **Cause-specific cumulative upgrading-risk changing dynamically over follow-up** as more patient data is gathered. The three **Panels A,B and C**: are ordered by the time of the latest visit (dashed vertical black line) of a new patient. At each of the latest follow-up visits, we combine the 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.

### 88 *Appendix B.1. Validation of Risk Predictions*

89 We wanted to check the usefulness of our model for not only the PRIAS  
 90 patients but also for patients from other cohorts. To this end, we validated  
 91 our model in the PRIAS dataset (internal validation) and the largest six co-  
 92 horts from the GAP3 database [6]. These are the University of Toronto AS  
 93 (Toronto), Johns Hopkins AS (Hopkins), Memorial Sloan Kettering Can-  
 94 cer Center AS (MSKCC), University of California San Francisco Active  
 95 Surveillance (UCSF), King’s College London AS (KCL), Michigan Urological  
 96 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 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 com-  
 pared the averaged cumulative-risk profile with a non-parametric estimate [7]  
 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 base-  
 line hazard of the joint model fitted to the PRIAS dataset, individually for  
 each of the cohorts except the Hopkins cohort. More specifically, given the  
 data of an external cohort  $\mathcal{A}^c$ , where  $c$  denotes the cohort, the recalibrated  
 parameters  $\gamma_{h_0}^c$  (Appendix A) of the log baseline hazard are given by:

$$p(\gamma_{h_0}^c \mid \mathcal{A}^c, \mathbf{b}^c, \boldsymbol{\theta}) \propto \prod_{i=1}^{n^c} p(l_i^c, r_i^c \mid \mathbf{b}_i^c, \boldsymbol{\theta}) p(\gamma_{h_0}^c) \quad (5)$$

97 where  $n^c$  are the number of patients in the  $c$ -th cohort, and  $\boldsymbol{\theta}$  is the vector of  
 98 all parameters of the joint model fitted to the PRIAS dataset. The interval in  
 99 which upgrading is observed for the  $i$ -th patient is given by  $l_i^c, r_i^c$ , with  $r_i^c = \infty$   
 100 for right-censored patients. The symbol  $\mathbf{b}_i^c$  denotes patient-specific random  
 101 effects (Appendix A) in the  $c$ -th cohort. The random effects are obtained  
 102 using the joint model fitted to the PRIAS dataset before recalibration. We  
 103 re-evaluated the calibration-in-the-large of our model after the recalibration  
 104 of the baseline hazard individually for each cohort. The improved calibration-  
 105 in-the-large is shown in Panel B of Figure 6.

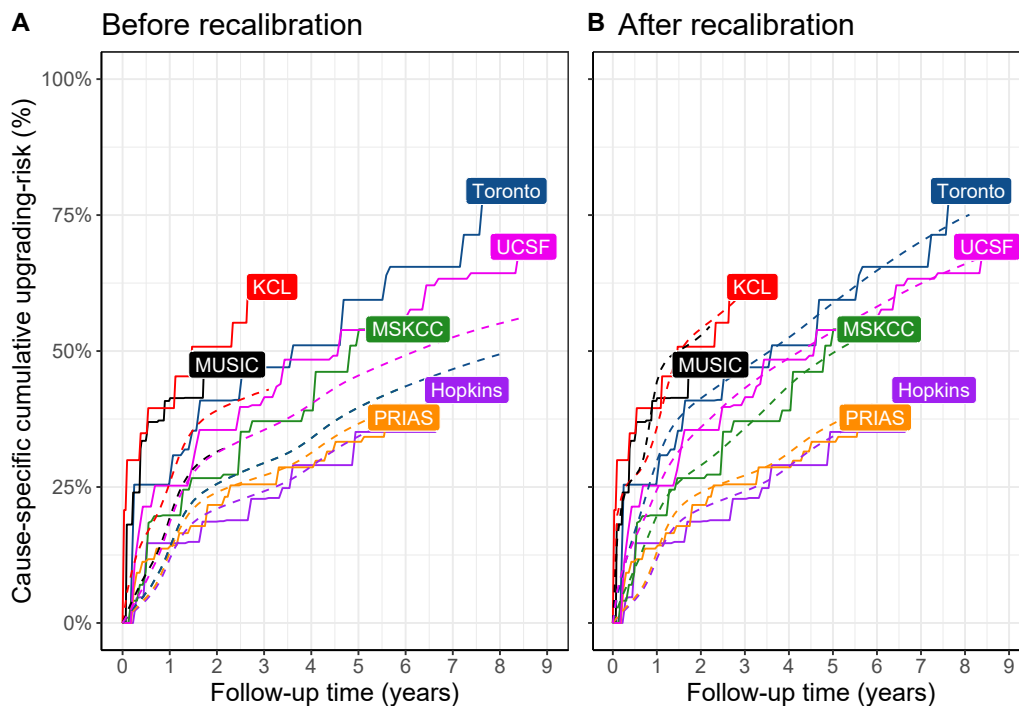


Figure 6: **Calibration-in-the-large of our model:** In **Panel A** we can see that our model is not well calibrated for use in KCL, MUSIC, Toronto and MSKCC. In **Panel B** we can see that calibration of model predictions improved in KCL, MUSIC, Toronto and MSKCC cohorts after recalibrating our model.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



106     ***Recalibrated PRIAS Model Versus Individual Joint Models***  
107     ***For Each Cohort*** We wanted to check if our recalibrated PRIAS model  
108 performed as good as a new joint model that could be fitted to the external  
109 cohorts. To this end, we predicted cause-specific cumulative upgrading-risk  
110 for each patient from each cohort using two sets of models, namely the recal-  
111 ibrated PRIAS model for each cohort, and a new joint model fitted to each  
112 cohort. The difference in predicted cause-specific cumulative upgrading-risk  
113 from these models is shown in Figure 7. We can see that the difference is  
114 smaller in those cohorts in which the effects of  $\log_2(\text{PSA} + 1)$  value and ve-  
115 locity were similar to that of PRIAS (Table 8). For example, the Hopkins  
116 cohort had parameter estimates similar to that of PRIAS, and consequently,  
117 the difference in predicted risks for this cohort is smallest. The opposite of  
118 this phenomenon holds for the MUSIC and KCL cohorts.

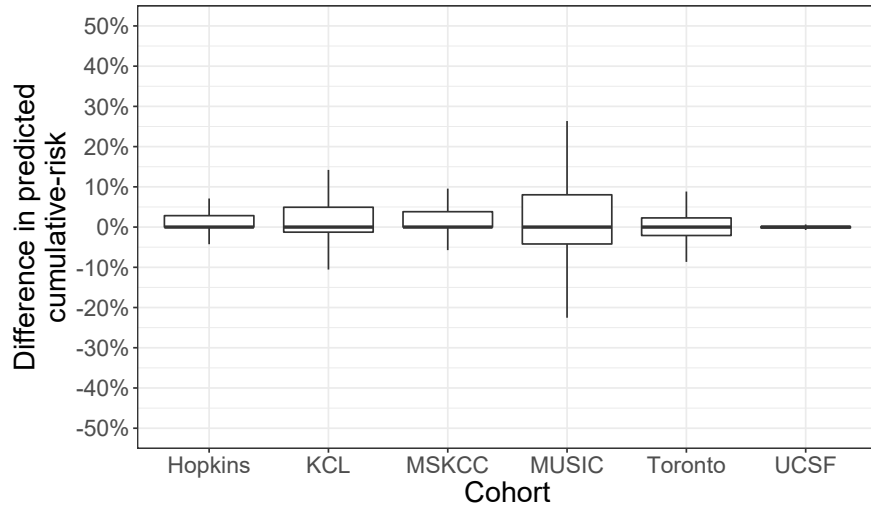


Figure 7: **Comparison of predictions from recalibrated PRIAS model with individual joint models fitted to external cohorts:** On Y-axis we show the difference between predicted cause-specific cumulative upgrading-risk for individual patients using two models, namely the recalibrated PRIAS model for 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 [9]. For time-varying prediction error, we assessed the mean absolute prediction error or MAPE [9]. 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  $\text{AUC} > 0.5$  because an  $\text{AUC} \leq 0.5$  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,  $v \in \{1, 1.5, \dots\}$  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

157 who probably had Gleason grade group 2 at inclusion but were misclassified  
 158 by the urologist as Gleason grade group 1 patients. To remedy this problem,  
 159 a biopsy for all patients at year one is commonly recommended in all AS  
 160 programs [11].

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]

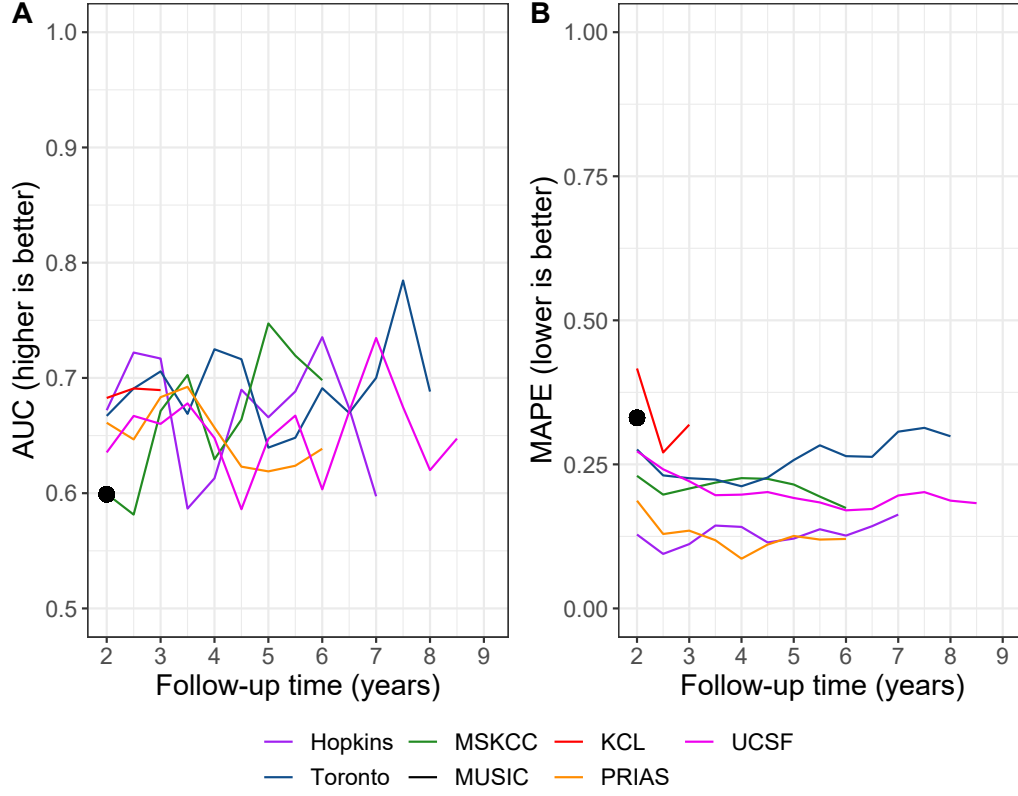


Figure 8: **Validation of dynamic predictions of cause-specific cumulative upgrading-risk.** In **Panel A** we can see that the time dependent area under the receiver operating characteristic curve or AUC (measure of discrimination) is above 0.5 in PRIAS (internal validation), and in Toronto, Hopkins, 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 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]



## 161 Appendix C. Personalized Biopsies Based on Cause-Specific Cu- 162 mulative Upgrading-Risk

163 Consider some real patients from the PRIAS database shown in Fig-  
164 ure 10– 12. In line with the protocols of most AS cohorts [12], we first  
165 schedule a compulsory biopsy at year one of follow-up. This promises early  
166 detection of Gleason upgrade for patients misdiagnosed as low-grade cancer  
167 patients or patients who chose AS despite having a higher grade at diagnosis.  
168 We also maintain a recommended minimum gap of one year between consec-  
169 utive biopsies [11]. That is, we intend to develop a personalized schedule of  
170 biopsies for these patients starting from the second year. The added benefit  
171 of planning biopsies year two onwards is that due to the longitudinal mea-  
172 surements accumulated over two years, and year one biopsy results, we are  
173 able to make reasonably accurate predictions of the cause-specific cumulative  
174 upgrading-risk.

Using the joint model fitted to the PRIAS dataset, we first obtain a pa-  
tient’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  
 $U = u_1, \dots, u_L$  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:

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

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

The cumulative-risk  $R_j(u_l | t_l, v)$  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  $Q_j^\kappa(u_l | t_l, v)$ . 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:

$$S_j^\kappa(U | t, v) = \{u_l \in U | Q_j^\kappa(u_l | t_l, v) = 1\}. \quad (6)$$

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

#### 177 *Appendix C.1. Expected Time Delay in Detecting Upgrading*

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

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:

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

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

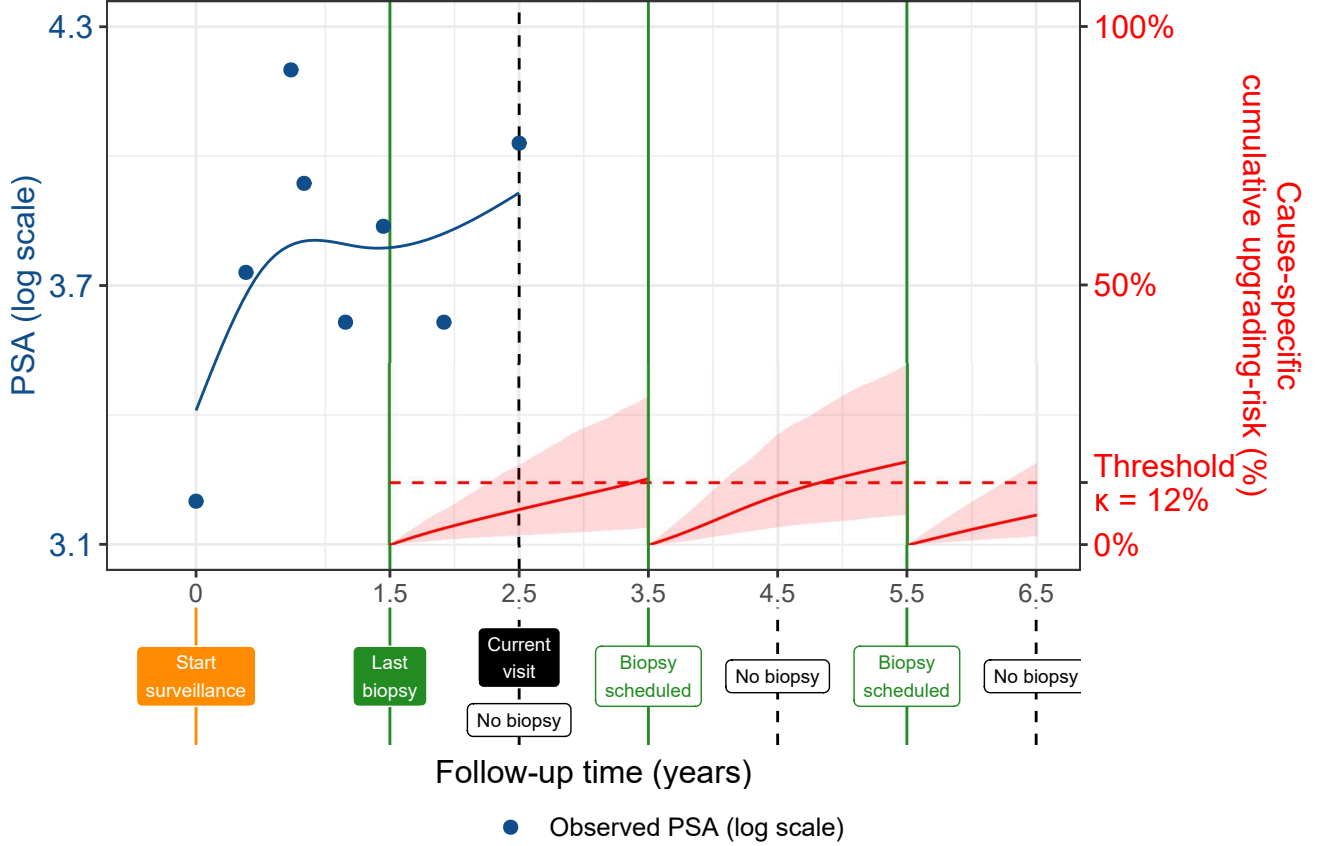


Figure 9: **Illustration of Personalized Biopsy Decisions Using Patient-specific Conditional Cumulative Upgrading-risk.** The last biopsy on which upgrading was not observed was conducted at  $t = 1.5$  years. The current visit time of the patient is  $v = 2.5$  years. Decisions for biopsy need to be 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  $s_{n-1} < T_j^* \leq s_n$ .

$$\begin{aligned}
 E\{D_j(S | t)\} &= \sum_{n=1}^N \left\{ s_n - E(T_j^* | s_{n-1}, s_n, v) \right\} \\
 &\quad \times \Pr\left\{ s_{n-1} < T_j^* \leq s_n \mid T_j^* \leq s_N, \mathcal{Y}_j(v), \mathcal{A}_n \right\}, \quad s_0 = t \\
 E(T_j^* | s_{n-1}, s_n, v) &= s_{n-1} + \int_{s_{n-1}}^{s_n} \Pr\left\{ T_j^* \geq u \mid s_{n-1} < T_j^* \leq s_n, \mathcal{Y}_j(v), \mathcal{A}_n \right\} du,
 \end{aligned}$$

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

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

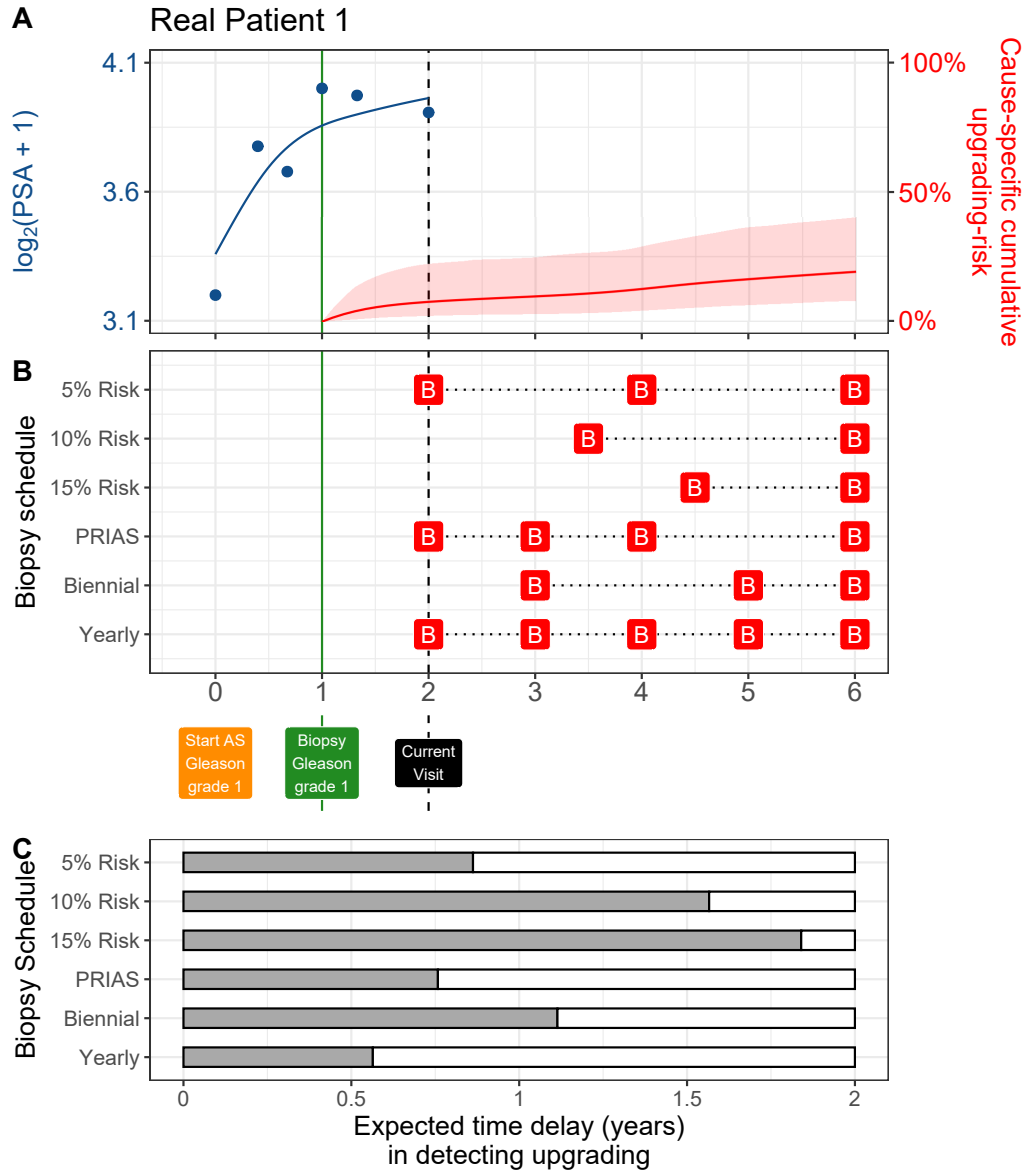


Figure 10: **Personalized and fixed schedules of biopsies for patient 1.** **Panel A:** shows the observed and fitted  $\log_2(\text{PSA} + 1)$  measurements (Equation 1), and the dynamic cause-specific cumulative upgrading-risk (see Appendix B) over follow-up period. **Panel B** shows the personalized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

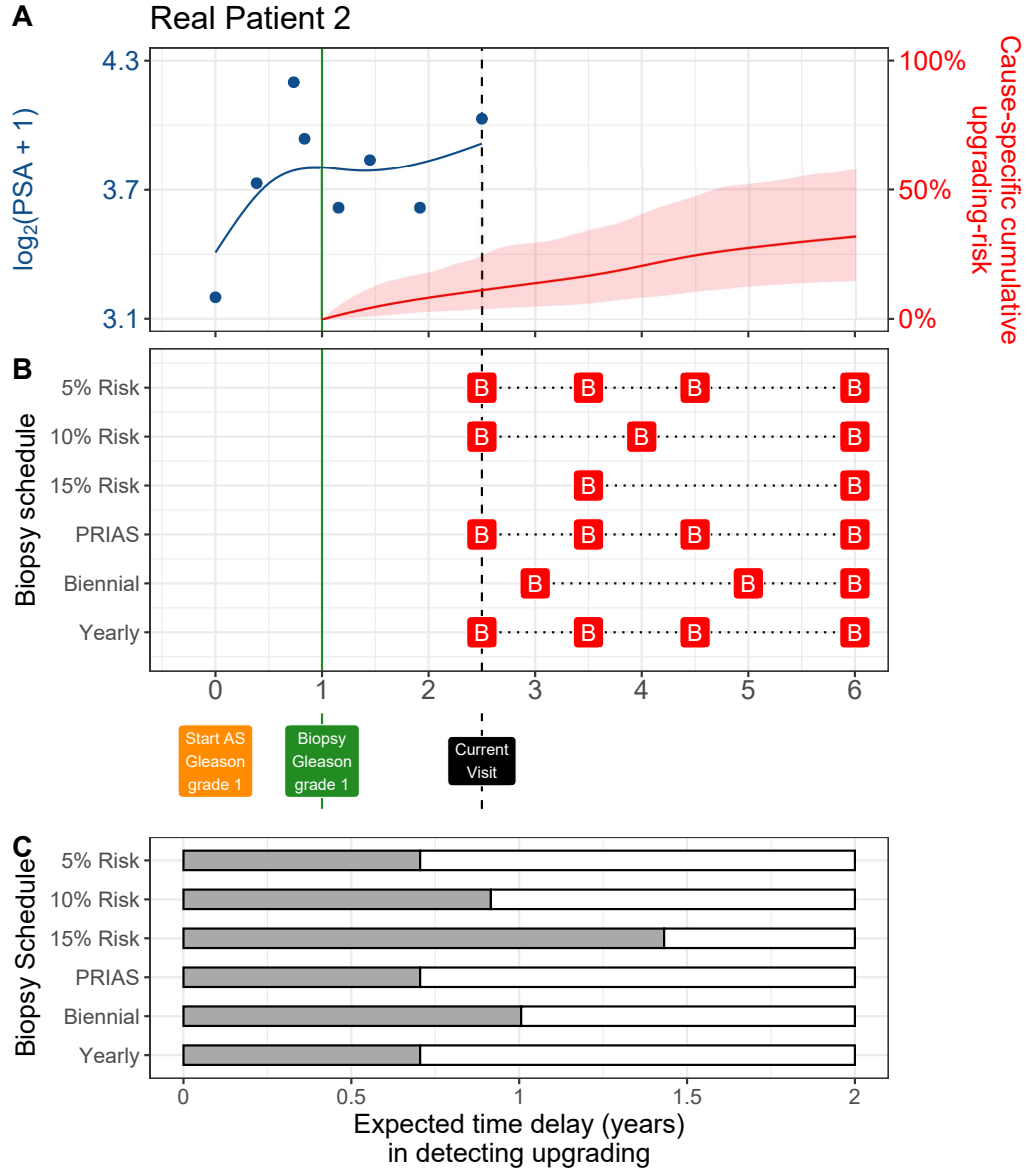


Figure 11: **Personalized and fixed schedules of biopsies for patient 2.** **Panel A:** shows the observed and fitted  $\log_2(\text{PSA} + 1)$  measurements (Equation 1), and the dynamic cause-specific cumulative upgrading-risk (see Appendix B) over follow-up period. **Panel B** shows the personalized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

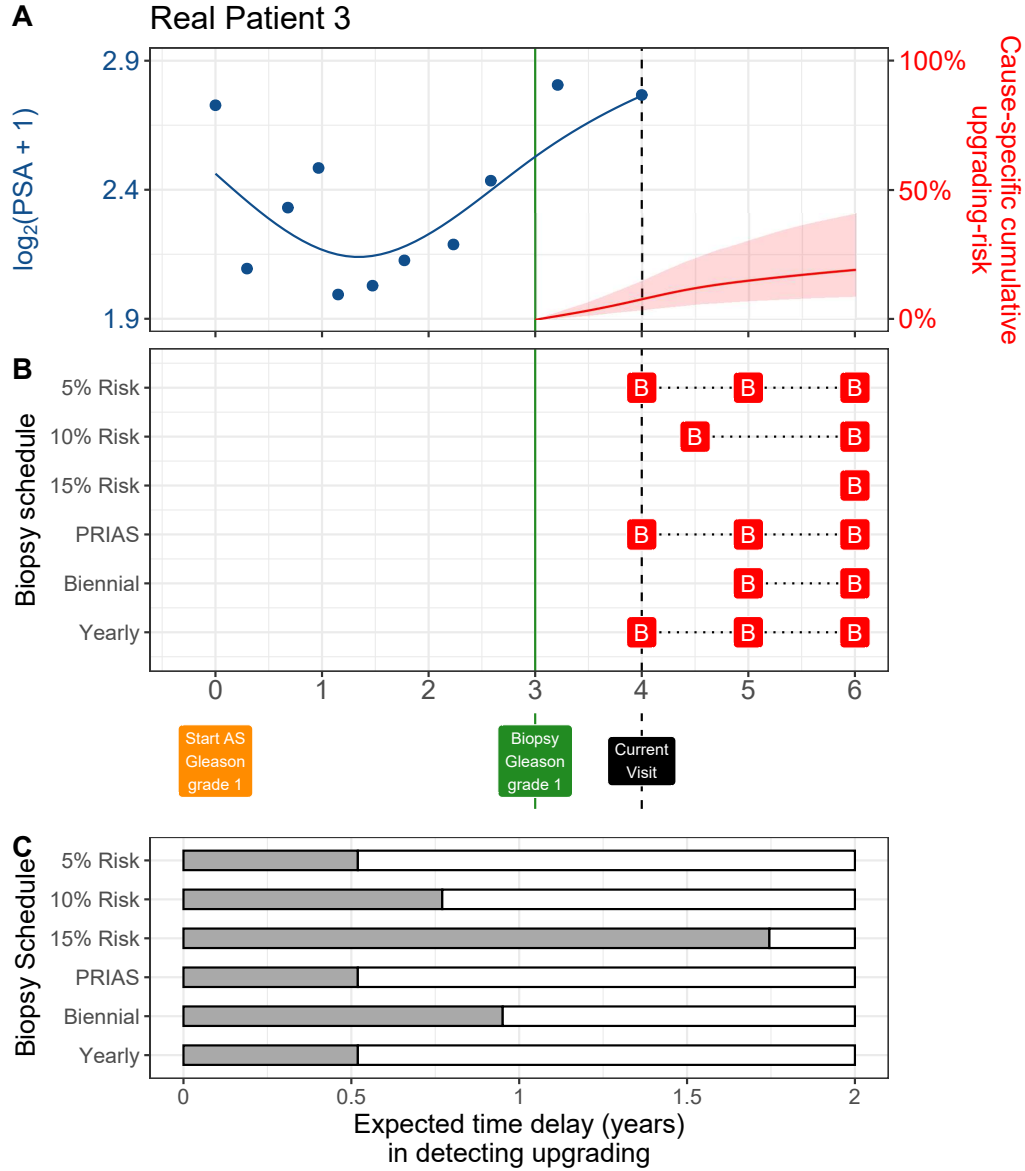


Figure 12: **Personalized and fixed schedules of biopsies for patient 3.** **Panel A:** shows the observed and fitted  $\log_2(\text{PSA} + 1)$  measurements (Equation 1), and the dynamic cause-specific cumulative upgrading-risk (see Appendix B) over follow-up period. **Panel B** shows the personalized and fixed schedules of biopsies with a 'B' indicating times of biopsies. **Panel C** various schedules are compared in terms of the expected time delay in detecting upgrading (years) if patient progresses before year six. A compulsory biopsy was scheduled at year six (maximum biopsy scheduling time in PRIAS, Table 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 [6]. 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/](https://emcbiostatistics.shinyapps.io/prias_biopsy_recommender/).

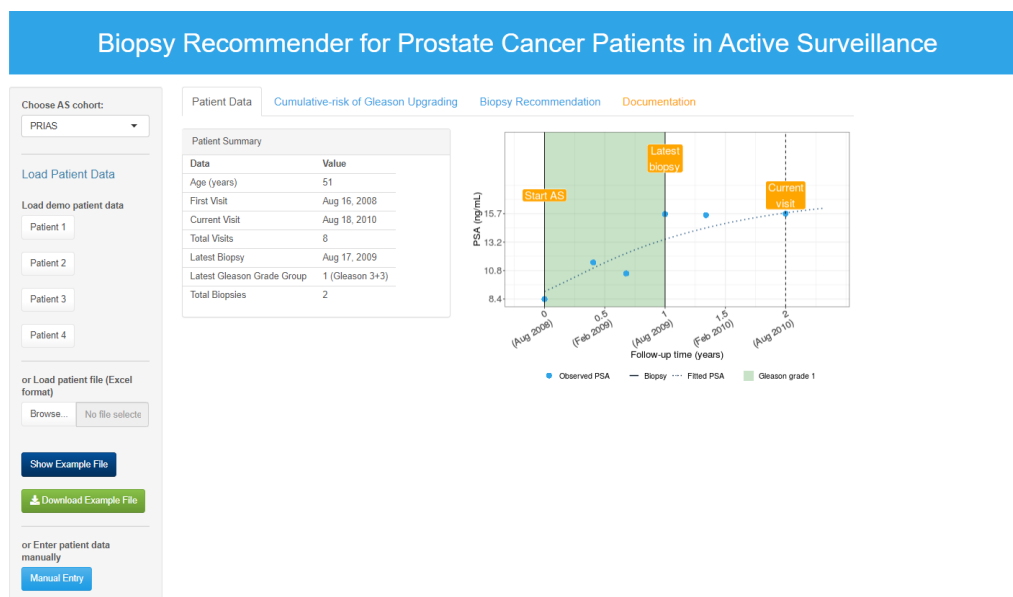


Figure 13: Landing page of the web-application. Panel on the left allows users to load patient data and panel on the right provides information. Patient data can be entered manually, or via Excel files. In addition, demo patient data is already uploaded to assist users in understanding the web-application.

## 208 Appendix E. Source Code

209 The R code for fitting the joint model to the PRIAS dataset, is at [https://github.com/anirudhtomer/prias/tree/master/src/clinical\\_gap3](https://github.com/anirudhtomer/prias/tree/master/src/clinical_gap3). We  
 210 refer to this location as ‘R\_HOME’ in the rest of this document.  
 211

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

213 **Accessing the dataset:** The PRIAS dataset is not openly accessible.  
 214 However, access to the database can be requested via the contact links at  
 215 <https://www.prias-project.org>.  
 216

217 **Formatting the dataset:** This dataset, however, is in the so-called wide  
 218 format and also requires the removal of incorrect entries. This can be done  
 219 via the R script `R_HOME/dataset_cleaning.R`. This will lead to two R ob-  
 220 jects, namely ‘`prias_final.id`’ and ‘`prias_long_final`’. The ‘`prias_final.id`’ object  
 221 contains information about the time of upgrading for PRIAS patients. The  
 222 ‘`prias_long_final`’ object contains longitudinal PSA measurements, the time  
 223 of biopsies and results of biopsies.  
 224

225 **Fitting the joint model:** We use a joint model for time-to-event and  
 226 longitudinal data to model the evolution of PSA measurements over time,  
 227 and to simultaneously model their association with the risk of upgrading.  
 228 The R package we use for this purpose is called **JMbayes** ([https://cran.r-](https://cran.r-project.org/web/packages/JMbayes/JMbayes.pdf)  
 229 [project.org/web/packages/JMbayes/JMbayes.pdf](https://cran.r-project.org/web/packages/JMbayes/JMbayes.pdf)). The API we use, how-  
 230 ever, is currently not hosted on CRAN, and can be found here: <https://github.com/anirudhtomer/JMbayes>. The joint model can be fitted via  
 231 the script `R_HOME/analysis.R`. It takes roughly 6 hours to run on an Intel  
 232 Core-i5 machine with four cores and 8GB of RAM.  
 233

234 The graphs presented in the main manuscript, and the supplementary  
 235 material can be generated by the scripts in `R_HOME/plots/`.

### 236 Appendix E.2. Validation of Predictions of Upgrading

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

240 *Appendix E.3. Creating Personalized Schedules of Biopsies*

241     Once a joint model is fitted to the PRIAS dataset, personalized schedules  
242 of biopsies based on the risk of upgrading for new patients can be developed as  
243 shown in the script `R_HOME/plots/demo_schedule_supplementary.R` or di-  
244 rectly using the script `https://raw.githubusercontent.com/anirudhtomer/`  
245 `prias/master/src/lastpaper/pers\_schedule\_api.R`.

246 *Appendix E.4. Source Code for Web Application*

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

249 **Appendix F. Appendix A. Members of The Movember Founda-**  
 250 **tion’s Global Action Plan Prostate Cancer Active**  
 251 **Surveillance (GAP3) consortium**

252 *Principle Investigators:* Bruce Trock (Johns Hopkins University, The  
 253 James Buchanan Brady Urological Institute, Baltimore, USA), Behfar Ehdaie  
 254 (Memorial Sloan Kettering Cancer Center, New York, USA), Peter Car-  
 255 roll (University of California San Francisco, San Francisco, USA), Christo-  
 256 pher Filson (Emory University School of Medicine, Winship Cancer Insti-  
 257 tute, Atlanta, USA), Jeri Kim / Christopher Logothetis (MD Anderson  
 258 Cancer Centre, Houston, USA), Todd Morgan (University of Michigan and  
 259 Michigan Urological Surgery Improvement Collaborative (MUSIC), Michi-  
 260 gan, USA), Laurence Klotz (University of Toronto, Sunnybrook Health Sci-  
 261 ences Centre, Toronto, Ontario, Canada), Tom Pickles (University of British  
 262 Columbia, BC Cancer Agency, Vancouver, Canada), Eric Hyndman (Uni-  
 263 versity of Calgary, Southern Alberta Institute of Urology, Calgary, Canada),  
 264 Caroline Moore (University College London & University College London  
 265 Hospital Trust, London, UK), Vincent Gnanapragasam (University of Cam-  
 266 bridge & Cambridge University Hospitals NHS Foundation Trust, Cam-  
 267 bridge, UK), Mieke Van Hemelrijck (King’s College London, London, UK  
 268 & Guy’s and St Thomas’ NHS Foundation Trust, London, UK), Prokar  
 269 Dasgupta (Guy’s and St Thomas’ NHS Foundation Trust, London, UK),  
 270 Chris Bangma (Erasmus Medical Center, Rotterdam, The Netherlands/ rep-  
 271 resentative of Prostate cancer Research International Active Surveillance  
 272 (PRIAS) consortium), Monique Roobol (Erasmus Medical Center, Rotter-  
 273 dam, The Netherlands/ representative of Prostate cancer Research Interna-  
 274 tional Active Surveillance (PRIAS) consortium), The PRIAS study group,  
 275 Arnauld Villers (Lille University Hospital Center, Lille, France), Antti Ran-  
 276 nikko (Helsinki University and Helsinki University Hospital, Helsinki, Fin-  
 277 land), Riccardo Valdagni (Department of Oncology and Hemato-oncology,  
 278 Università degli Studi di Milano, Radiation Oncology 1 and Prostate Cancer  
 279 Program, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy),  
 280 Antoinette Perry (University College Dublin, Dublin, Ireland), Jonas Hugos-  
 281 son (Sahlgrenska University Hospital, Göteborg, Sweden), Jose Rubio-Briones  
 282 (Instituto Valenciano de Oncología, Valencia, Spain), Anders Bjartell (Skåne  
 283 University Hospital, Malmö, Sweden), Lukas Hefermehl (Kantonsspital Baden,  
 284 Baden, Switzerland), Lee Lui Shiong (Singapore General Hospital, Singa-  
 285 pore, 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 Hulsén (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 Hiram (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. 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.
  7. 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.
  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 models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass)* 2010;21(1):128.
  11. 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.
  12. 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.