Personalized Schedules for Burdensome Surveillance Tests

Journal:	*Journal of the American Statistical Association*
Manuscript ID	Draft
Manuscript Type:	Article – Applications & Case Studies
Keywords:	Chronic NCDs, Invasive diagnostic tests, Joint models, Personalized schedules, Prostate biopsy

Personalized Schedules for Burdensome Surveillance Tests

Abstract

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

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

Keywords: Chronic NCDs, Invasive diagnostic tests, Joint models, Personalized schedules, Prostate biopsy, Surveillance

1 Introduction

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

Invasive tests are repeated until progression is observed, typically as per a one-size-fits-all *fixed schedule*, e.g., biannually, (Krist et al., 2007; McWilliams et al., 2008; Bokhorst et al., 2015). A time gap between tests causes a time delay in detecting progression (Figure 1). A shorter delay in detecting progression (*benefit*) can provide a larger window of opportunity for curative treatment. However, with fixed schedules, this means conducting tests frequently. Frequent tests are *burdensome* as they may cause pain and/or severe medical complications (Krist et al., 2007; Loeb et al., 2013). Consequently, patients may not always comply with frequent tests (Bokhorst et al., 2015; Le Clercq et al., 2015). In general, because fixed schedules do not differentiate between fast and slow/non-progressing patients, they impose disproportionate burden/benefits across the patient population.

The goal of this work (Figure 1) is to optimize the number of invasive tests (burden) and the time delay in detecting progression (shorter is beneficial) better than fixed schedules. Specifically, we intend to *personalize* test schedules using patients' clinical data accumulated over surveillance follow-up. This data includes baseline characteristics, previous test results, and longitudinal outcomes (e.g., biomarkers, medical imaging, physical exami-

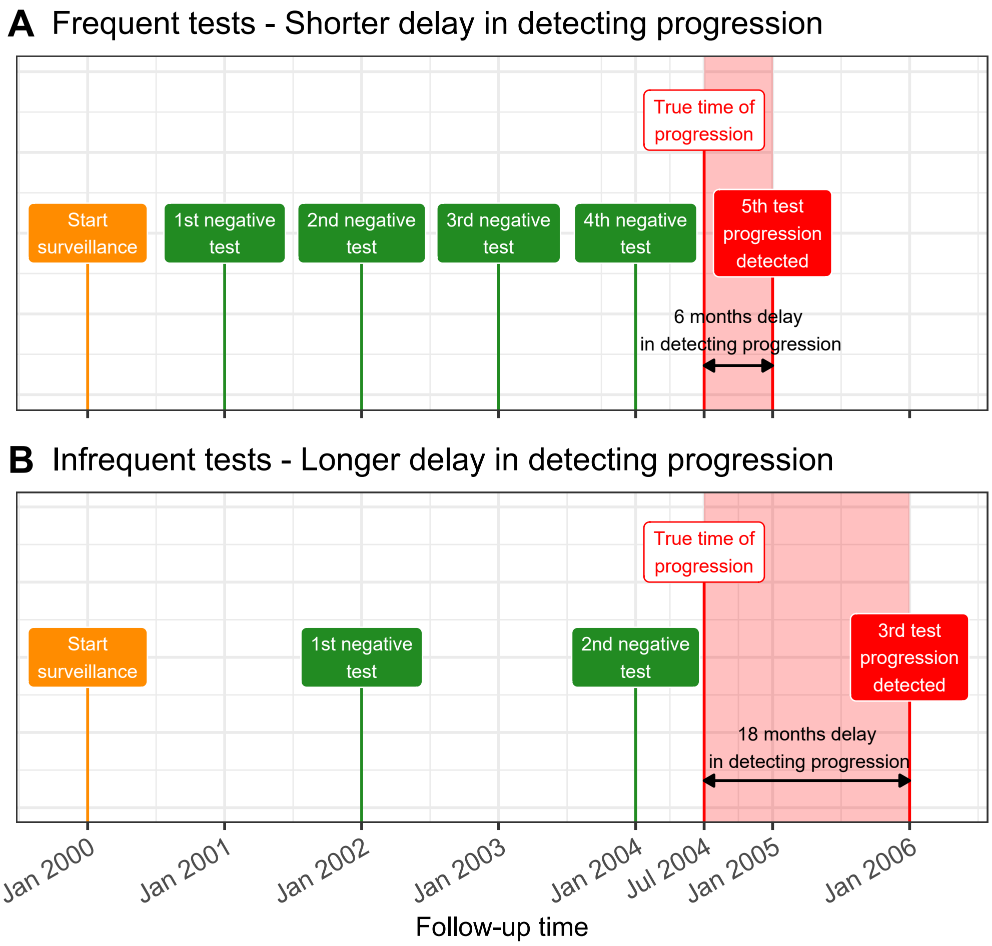

Figure 1: **Goal: Finding the optimal tradeoff between the number of invasive tests (burden) and time delay in detecting progression (shorter is beneficial)**. A progression is a non-terminal event in the surveillance of early-stage chronic non-communicable diseases. The true time of progression for the patient illustrated in this figure is July 2004. Since invasive tests are conducted repeatedly, progression is interval-censored and always observed with a delay. Frequent periodical invasive tests in **Panel A** lead to a shorter time delay in detecting progression than infrequent periodical invasive tests in **Panel B**. The interval-censored time of progression is Jan 2004–Jan 2005 in **Panel A** and between Jan 2004–Jan 2006 in **Panel B**.

nation). Many surveillance protocols currently personalize test schedules using heuristic methods such as decision flowcharts (Bokhorst et al., 2015; Weusten et al., 2017). However, flowcharts discretize continuous outcomes, often exploit only the last measurement, ignore the measurement error in observed data, and plan only one test at a time. Alternatively, a complete personalized schedule of tests can be obtained using partially observable Markov decision processes or POMDPs (Alagoz et al., 2011; Steimle and Denton, 2017). Although POMDPs typically discretize continuous longitudinal outcomes to avoid the curse of dimensionality. In scenarios such as ours, where decisions (test/no test) and disease state (low-grade disease/progressed) are both binary, POMDPs may not be necessary either. The reason is that such POMDPs give the same optimal schedule, which can be alternatively obtained by just planning a test when the probability of transition from non-progressed to progressed state is more than a certain threshold (see Vickers and Elkin, 2006, Equation 1).

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

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

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

The remaining paper is as follows. Section 2 introduces the joint modeling framework. The personalized scheduling methodology is described in Section 3, and demonstrated for prostate cancer surveillance patients in Section 4. In Section 5, we compare personalized and fixed schedules via a realistic simulation study based on a joint model fitted to the PRIAS dataset.

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

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

2.1 Longitudinal Sub-process

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

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

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

2.2 Survival Sub-process

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

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

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

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

These formulations of $f(\cdot)$ postulate that the hazard of progression at time t may depend on the underlying level $m_i(t)$ (e.g., PSA value in prostate cancer) or on both the level and velocity $m_i'(t)$ (e.g., PSA velocity) of the longitudinal outcome at t. Lastly, the baseline hazard $h_0(t)$ is modeled flexibly using P-splines (Eilers and Marx, 1996). The detailed specification of the baseline hazard, and the joint parameter estimation of the longitudinal and relative-risk sub-models using the Bayesian approach are presented in Supplementary A.

3 Personalized Schedule of Invasive Tests for Detecting Progression

3.1 Cumulative-risk of progression

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

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

The cumulative-risk function $R_j(\cdot)$ depends on patient-specific clinical data and the training dataset, via the posterior distribution of the random effects $\boldsymbol{b}_j$ and posterior distribution of the vector of all parameters $\boldsymbol{\theta}$ of the fitted joint model, respectively. This cumulative-risk function is dynamic, in the sense that it automatically updates over time as more longitudinal data become available (Figure 2).

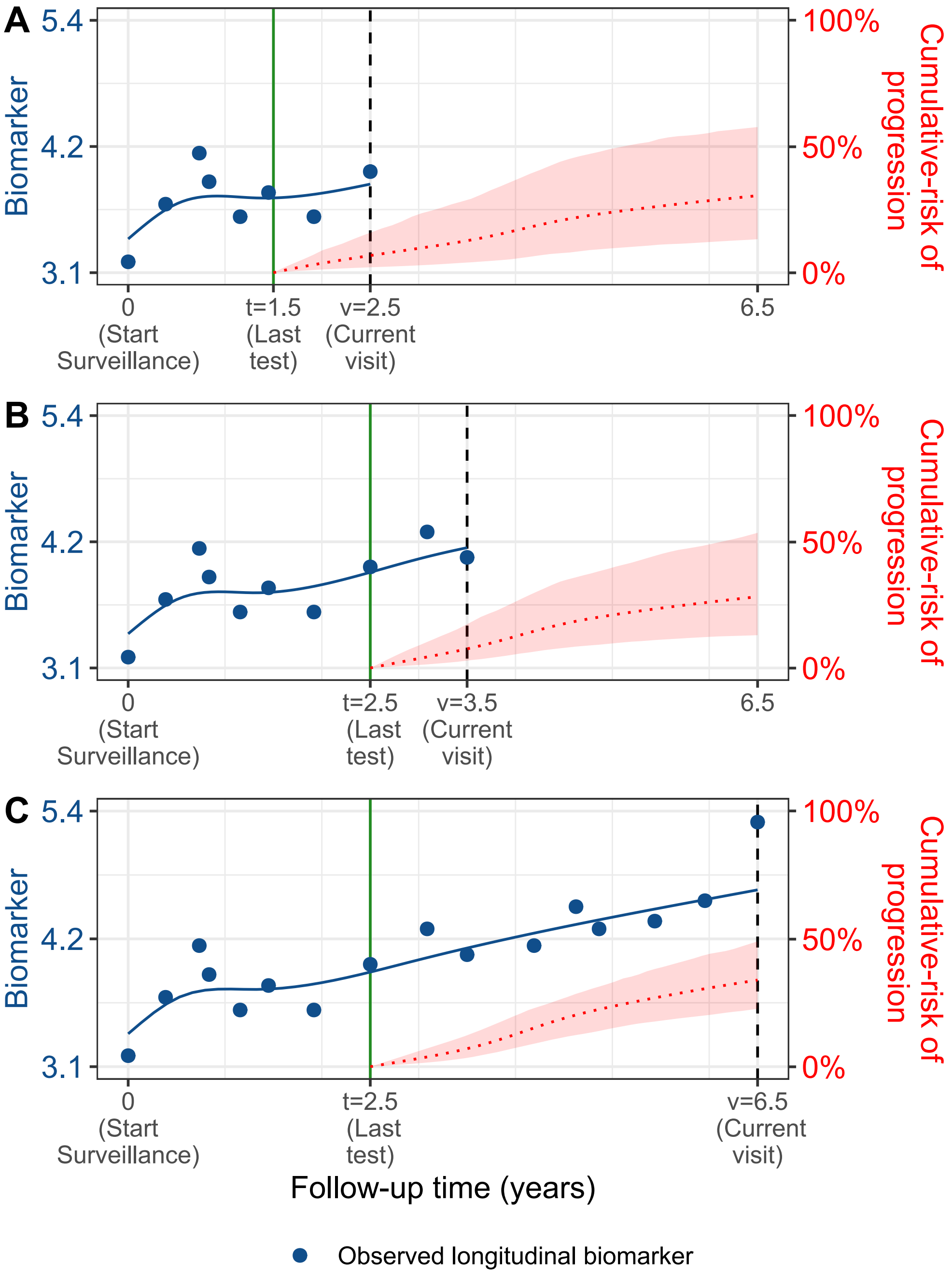

Figure 2: **Cumulative-risk of progression updated dynamically over follow-up** as more patient data is gathered. A single longitudinal outcome, namely, a continuous biomarker of disease progression, is used for illustration. **Panels A, B and C:** are ordered by the time of the current visit v (dashed vertical black line) of a new patient. At each of these visits, we combine the accumulated longitudinal data (shown in blue circles), and time of the last negative invasive test t (solid vertical green line) to obtain the updated cumulative-risk profile $R_j(u \mid t, v)$ (dotted red line with 95% credible interval shaded) of the patient defined in (1). All values are illustrative.

3.2 Personalized Test Decision Rule

We intend to exploit the cumulative-risk function $R_j(\cdot)$ to develop a risk-based personalized schedule of invasive tests for the j-th patient. Typically, invasive tests are decided on the same visit times on which longitudinal data (e.g., biomarkers) are measured. Let $U = \{u_1, \ldots, u_L\}$ represent a schedule of such visits (e.g., biannual PSA measurement in prostate cancer). Here, $u_1 = v$ is also the current visit time. The maximum future visit time u_L can be chosen based on the available information in the training dataset $\mathcal{A}_n$. That is, tests for the new patient j are planned only up to a future visit time u_L at which a sufficient number of events in $\mathcal{A}_n$ are available for making reliable risk predictions (e.g., up to the 80% or 90% percentile of progression times).

We propose to take the decision of conducting a test at a future visit time $u_l \in U$ if the cumulative-risk of progression at time u_l exceeds a certain risk threshold κ (Figure 3). In particular, the test decision at time u_l is given by,

$$Q_j^{\kappa}(u_l \mid t_l, v) = I\big\{R_j(u_l \mid t_l, v) \geq \kappa\big\}, \quad 0 \leq \kappa \leq 1, \tag{2}$$

where $I(\cdot)$ is the indicator function, $R_j(u_l \mid t_l, v)$ is the cumulative-risk of progression at the current decision time u_l, and $t_l < u_l$ is the time of the last test conducted before u_l. Thus, the future time at which a test will be planned, depends on both the threshold κ and the cumulative-risk of the patient. Moreover, when a test gets planned at time u_l, i.e., $Q_j^{\kappa}(u_l \mid t_l, v) = 1$, then the cumulative-risk profile is updated before making the next test decision at time u_{l+1} (Figure 3). Specifically, the cumulative-risk at time u_{l+1} is updated by setting the corresponding time of the last test $t_{l+1} = u_l$. This accounts for the possibility that progression may occur after time $u_l < T_j^*$. Hence, the time of last test t_l is defined as,

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

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

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

To facilitate shared-decision making of invasive tests, we translate our proposed decision rule, i.e., the choice of a specific risk threshold κ, into two clinically relevant quantities. First, the number of tests (burden) we expect to perform for patient j, and second, if the patient progresses, the time delay (shorter is beneficial) expected in detecting progression. To calculate these two quantities, we first suppose that patient j does not progress between his last negative test at time t and the maximum future visit time u_L. Under this assumption, the subset of future visit times in U on which a test is planned using (2) results into a personalized schedule of future tests (Figure 3), given by,

$$\{s_1, \ldots, s_{N_j}\} = \big\{u_l \in U : Q_j^\kappa(u_l \mid t_l, v) = 1\big\}, \quad N_j \leq L. \tag{3}$$

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

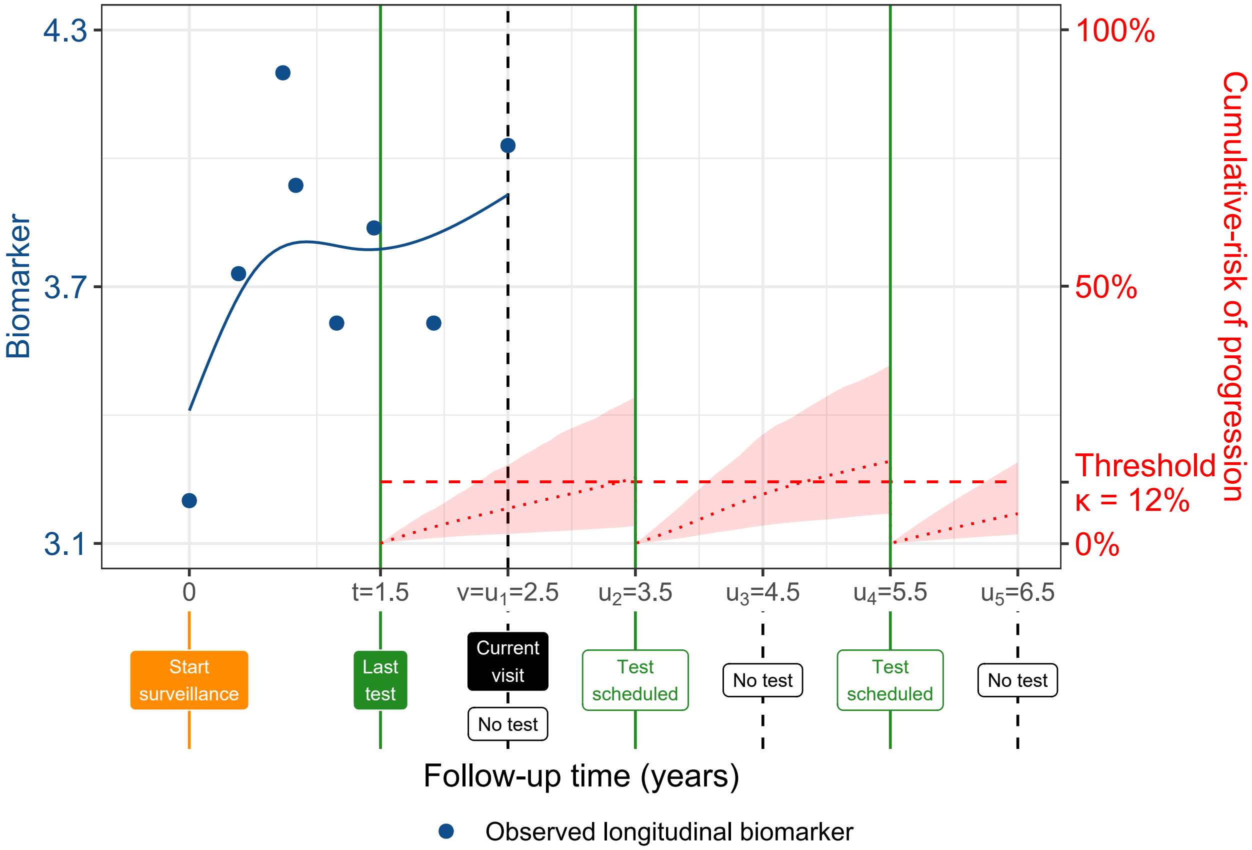

Figure 3: **Successive personalized test decisions based on patient-specific cumulative-risk of progression (2)**. Time of current visit: $v = 2.5$ years (dashed vertical black line). Time of the last test on which progression was not observed: $t = 1.5$ years. Longitudinal data up to current visit: $\mathcal{Y}_j(v)$ is a continuous biomarker (blue circles). Example risk threshold: $\kappa = 0.12$ (12%). Grid of future visits on which future tests are planned: $U = \{2.5, 3.5, 4.5, 5.5, 6.5\}$ years. The cumulative-risk profiles $R_j(u_l \mid t_l, v)$ employed in (2) are shown with dotted red lines (95% credible intervals shaded), and are updated each time a test is planned (solid vertical green lines). Future test decisions $Q_j(u_l \mid t_l, v)$ defined in (2) are: $Q_j^\kappa(u_1 = 2.5 \mid t_1 = 1.5, v) = 0$, $Q_j^\kappa(u_2 = 3.5 \mid t_2 = 1.5, v) = 1$, $Q_j^\kappa(u_3 = 4.5 \mid t_2 = 3.5, v) = 0$, $Q_j^\kappa(u_4 = 5.5 \mid t_2 = 3.5, v) = 1$, and $Q_j^\kappa(u_5 = 6.5 \mid t_5 = 4.5, v) = 0$. All values are illustrative.

as,

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

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

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

where

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

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

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

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

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

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

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

The personalized schedule in (3), and the corresponding personalized expected number of tests and time delay, have the advantage of getting updated with newly collected data over follow-up. Also, the expected number of tests and time delay can be calculated for any schedule, fixed or personalized. Hence, patients/doctors can use them to compare different schedules. Although, a fair comparison of time delays between different schedules for the same patient, requires a compulsory test at a common horizon time point in all schedules.

3.4 How to Select the Risk Threshold κ

The risk threshold κ controls the timing and the total number of invasive tests in the personalized schedule S_j^κ. Through the timing and the total number of planned tests, κ also indirectly affects the potential time delay (Figure 1) in detecting progression if a particular schedule is followed. Hence, κ should be chosen while balancing both the number of invasive tests (burden) and the time delay in detecting progression (shorter is beneficial).

To facilitate the choice of κ in practice, following our developments in the previous section, we translate the different choices for threshold κ into the expected number of tests and time delay. In particular, for a patient j having data available up to his current visit time v, we can construct a bi-dimensional Euclidean space of his expected total number of tests (x-axis) and expected time delay in detecting progression (y-axis), for different personalized test schedules obtained by varying κ in $[0, 1]$, e.g., Figure 4.

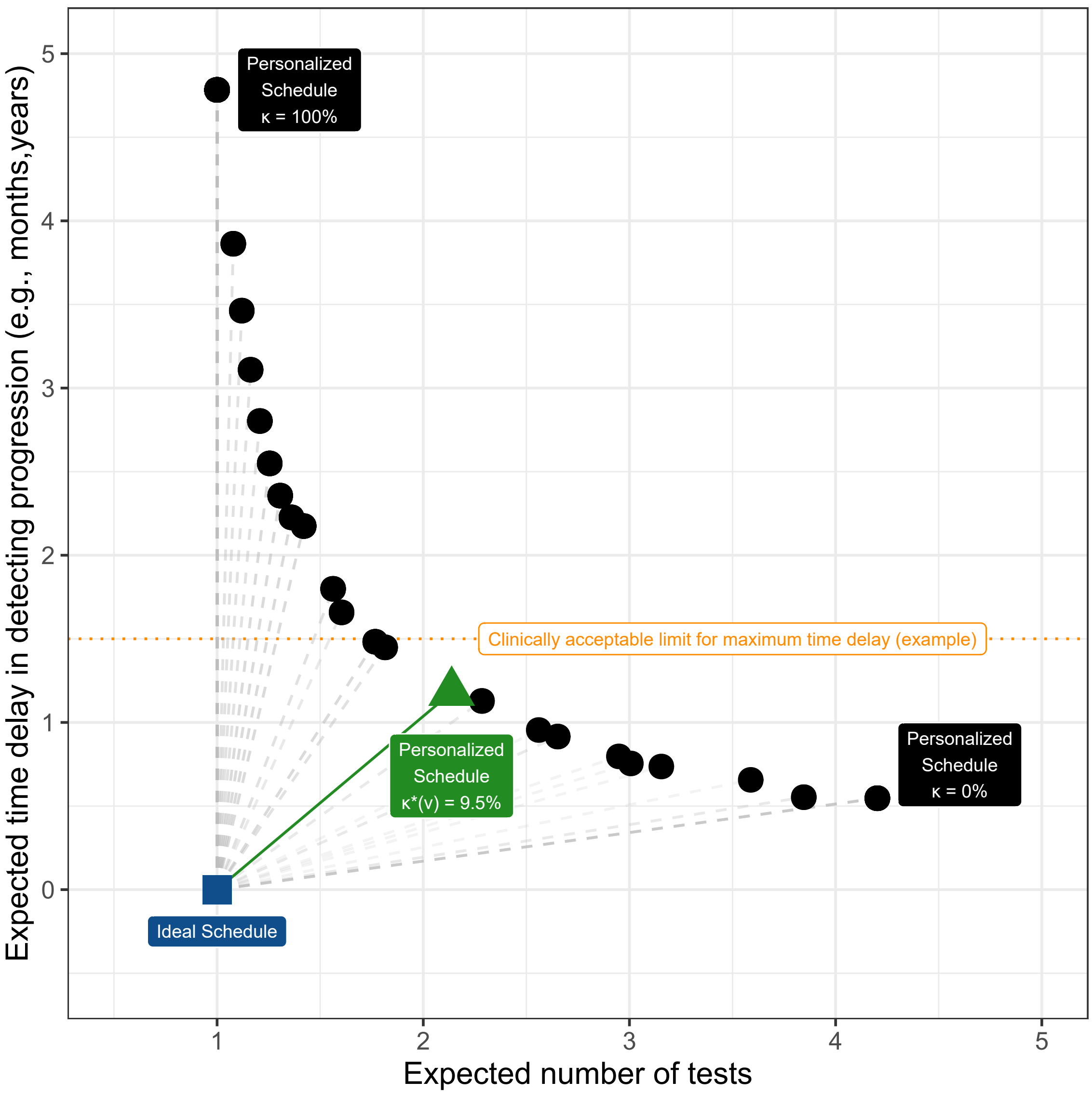

Figure 4: **Optimal current-visit time v specific risk threshold $\kappa^*(v)$ obtained using (4)** for the patient shown in Figure 3. Ideal schedule of tests: point (1,0) shown as a blue square. It plans exactly one invasive test at the true time of progression T_j^* of a patient. Hence, the time delay in detecting progression is zero. Various personalized schedules based on a grid of thresholds κ in $[0, 1]$ are shown with black circles. Higher thresholds lead to fewer tests, but also higher expected time delay. The personalized schedule based on $\kappa^*(v) = 9.5\%$ threshold (green triangle) has the least Euclidean distance (solid green line) to the ideal schedule. It is also possible to optimize the least distance under a certain clinically acceptable limit on the time delay (dotted horizontal orange line).

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

$$\kappa^*(v) = \arg\min_{0 \leq \kappa \leq 1} \sqrt{\Big[E\big\{\mathcal{N}_j(S^\kappa_j)\big\} - 1\Big]^2 + \Big[E\big\{\mathcal{D}_j(S^\kappa)\big\} - 0\Big]^2}. \tag{4}$$

In certain scenarios, patients/doctors may be apprehensive about undergoing more than a maximum expected number of future tests, or having an expected time delay higher than certain months. For such purposes, the Euclidean distance in (4) can be optimized under constraints on the expected number of tests or expected time delay (Figure 4). Doing so alleviates two problems, namely, that the time delay and the number of tests have different units of measurement, and that in (4) they are weighted equally (Cook and Wong, 1994).

We considered shorter delays in detecting progression as the benefit of repeated tests. However, in the literature, decision-theoretic measures such as quality-adjusted life-years/expectancy (QALY/QALE) gained (Sassi, 2006) have also been used to quantify the benefit of testing. Optimizing (4) with QALE needs, setting the optimal point in a Euclidean space with QALE as a dimension, and obtaining expected QALEs for different schedules. For estimating the expected QALE in a personalized manner, a mathematical definition of QALE in terms of time delay $\mathcal{D}_j$ in detecting progression (de Carvalho et al., 2017b) is required.

4 Application of Personalized Schedules in Prostate Cancer Surveillance

We next demonstrate personalized schedules for scheduling biopsies in prostate cancer active surveillance. To this end, we use results from a joint model fitted to the PRIAS dataset introduced in Section 1. The model definition (Supplementary B) utilized a linear mixed sub-model for biannually measured PSA (continuous: log-transformed from ng/mL), and a logistic mixed sub-model for biannually measured DRE (binary: tumor palpable or not). In the survival sub-model, fitted PSA value, fitted instantaneous PSA velocity (defined in Section 2.2), and log-odds of having a DRE indicating a palpable tumor, were included as time-dependent predictors. The model parameters were estimated under the Bayesian framework using the R package **JMbayes** (Rizopoulos, 2016), and are presented in Supplementary B. We next briefly present the key results relevant for personalized scheduling.

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

4.1 Personalized Biopsy Schedules for a Demonstration Patient

We utilized the joint model fitted to the PRIAS dataset to schedule biopsies in a demonstration prostate cancer patient shown in Figure 5. The time of his last negative biopsy was $t = 3.5$ years, and the time of the current visit was $v = 5$ years. We made biopsy decisions over his future visits for PSA measurement $U = \{u_1 = 5, u_2 = 5.5, \ldots, u_L = 10\}$ years using four different schedules. Two of the fixed schedules are annual biopsy schedule and the PRIAS schedule. The PRIAS schedule has compulsory biopsies at year one, four, seven, and ten of follow-up, and additional annual biopsies if PSA doubling-time (Bokhorst et al., 2015) is high. Remaining two schedules are personalized, namely, with a fixed threshold $\kappa = 10\%$ risk, and an automatically chosen current visit time v specific risk $\kappa^*(v)$ (Section 3.4). Since the demonstration patient's time of last negative biopsy $t = 3.5$ is after year one of follow-up, a time delay in detecting progression up to three years may not lead to adverse downstream outcomes (de Carvalho et al., 2017a).

The cumulative-risk of progression of the demonstration patient increases 3% yearly on average, up to 19% at the maximum study period of ten years. Hence, the patient may progress slowly. Consequently, risk-based personalized approaches plan fewer biopsies than the annual schedule (Panel B, Figure 5). Also, the time delay in detecting progression for personalized schedules (Panel D, Figure 5) is below the safe limit of three years mentioned earlier. Thus, personalized schedules can be a suitable alternative to the annual schedule.

5 Simulation Study

Although we evaluated personalized schedules for a demonstration patient, we also intend to analyze and compare personalized and fixed schedules in a full cohort. Our criteria for

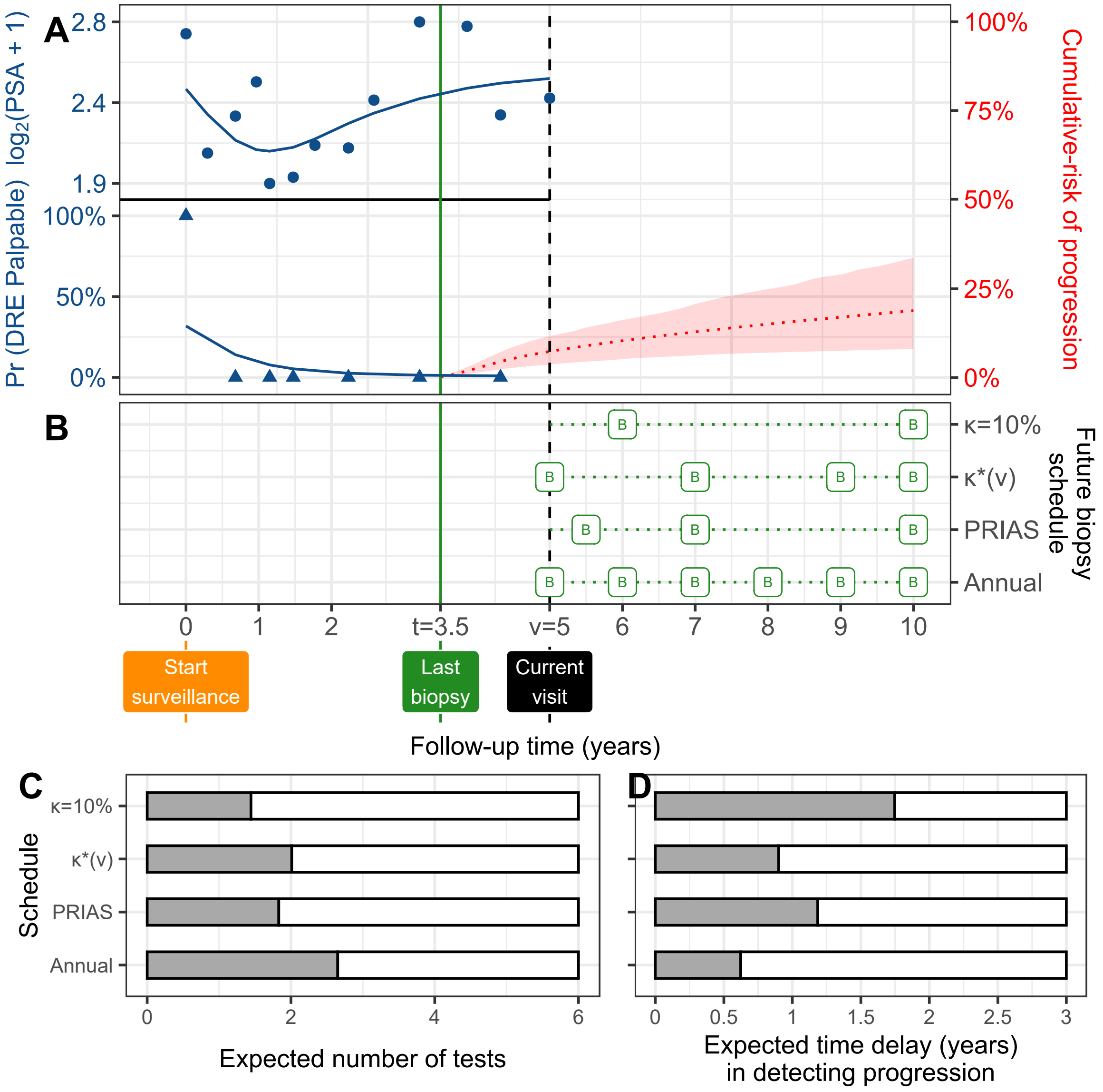

Figure 5: **Personalized schedules for a demonstration prostate cancer patient**. **Panel A**: Time of current visit: $v = 5$ years (black dashed line). Time of last negative biopsy: $t = 3.5$ years (vertical green solid line). Longitudinal data: $\log_2(\text{PSA} + 1)$ transformed PSA (observed: blue dots, fitted: solid blue line), and binary DRE (observed: blue triangles, fitted probability: solid blue line). Cumulative-risk profile: dotted red line (95% credible interval shaded). **Panel B**: Biopsy indicated with a 'B', and $\kappa = 10\%$ and $\kappa^*(v)$ are personalized biopsy schedules using a risk threshold of 10%, and a visit time v specific automatically chosen threshold (4), respectively. PRIAS and Annual denote the PRIAS biopsy schedule (Section 4.1) and annual biopsy schedule. **Panel C,D**: For all schedules we calculate the expected number of tests and expected time delay in detecting progression if the patient progresses before year ten. Since a recommended minimum gap of one year is maintained between biopsies, maximum possible number of tests are six. A delay in detecting progression of up to three years may not lead to adverse outcomes (de Carvalho et al., 2017a).

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

5.1 Simulation Setup

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

For each test patient, we conduct hypothetical biopsies using two fixed (PRIAS and annual schedule) and three personalized biopsy schedules. Personalized schedules are based on, a fixed risk threshold $\kappa = 10\%$, an optimal current visit time v specific threshold $\kappa^*(v)$ chosen via (4), and an optimal threshold obtained under the constraint that expected time delay in detecting progression is less than 0.75 years (9 months), denoted $\kappa^*\{v \mid E(\mathcal{D}) \leq 0.75\}$. The choice of 0.75 years delay constraint is arbitrary and is only used to illustrate that applying the constraint limits the average delay at 0.75 years. Successive personalized biopsy decisions are made only on the standard PSA follow-up visits, utilizing clinical data accumulated only until the corresponding current visit time (2). We maintain a minimum recommended gap of one year between consecutive prostate biopsies (Bokhorst et al., 2015) as well. Biopsies are conducted until progression is detected, or the maximum follow-up period at year ten (horizon) is reached. The actual time delay in detecting progression is equal to the difference in time at which progression is detected and the actual (simulated) time of progression of a patient.

5.2 Simulation Results

In the simulation study, nearly 50% of the patients observed progression during the ten year study period (*progressing*) and 50% did not (*non-progressing*). While we can calculate the total number of biopsies scheduled in all 500 × 250 test patients, the actual time delay in detecting progression is available only for progressing patients. Hence, we show the simulation results separately for progressing and non-progressing patients (Figure 6).

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

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

6 Discussion

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

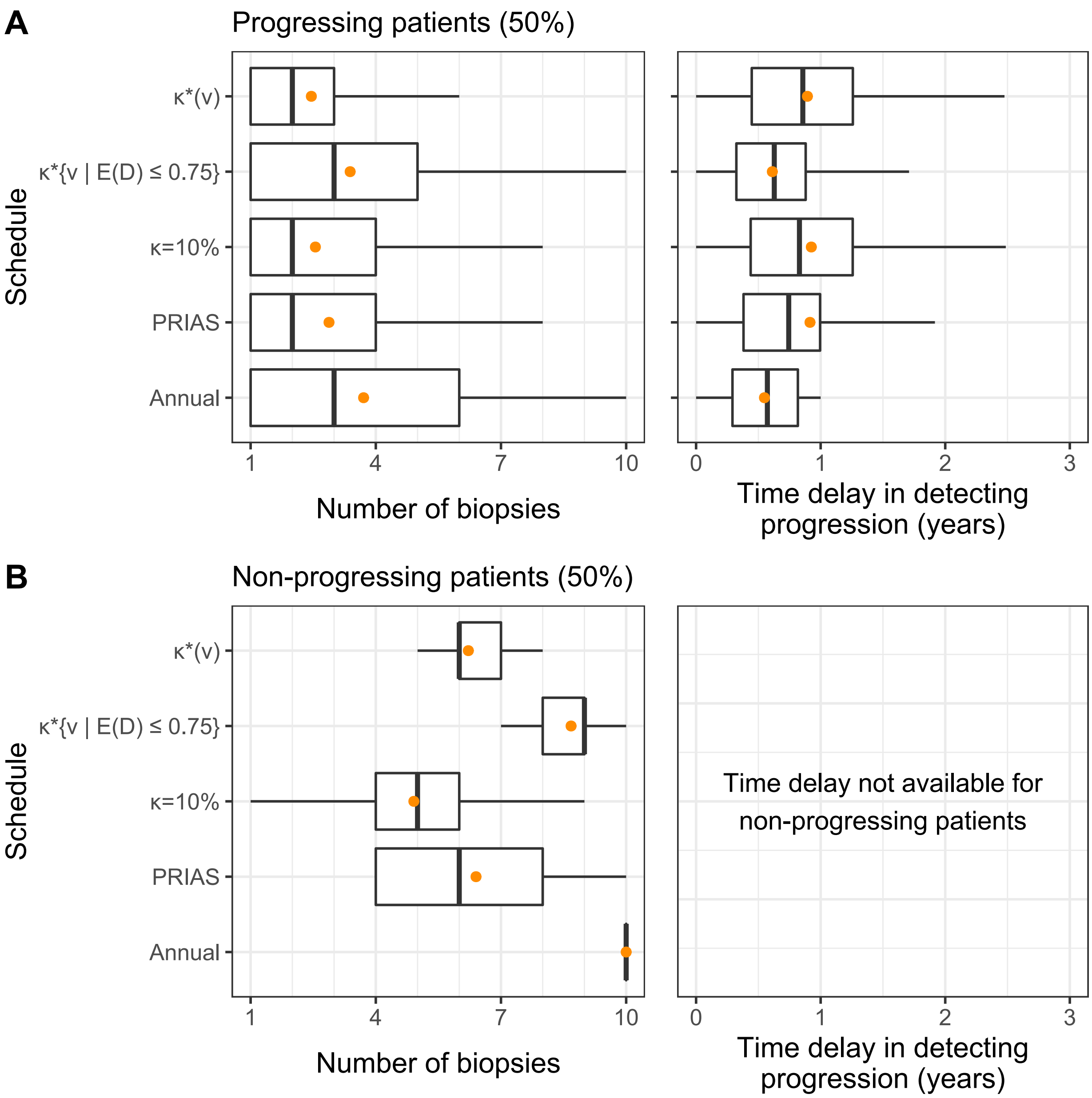

Figure 6: **Number of biopsies and the time delay in detecting cancer progression for various biopsy schedules** obtained via a simulation study. **Mean** is indicated by the orange circle. Time delay (years) is calculated as (time of positive biopsy - the actual simulated time of cancer progression). Biopsies are conducted until cancer progression is detected. **Panel A:** simulated patients who obtained cancer progression in the ten year study period (progressing). **Panel B:** simulated patients who did not obtain cancer progression in the ten year study period (non-progressing). Types of schedules: $\kappa = 10\%$ and $\kappa^*(v)$ schedule a biopsy if the cumulative-risk of cancer progression at the current visit time v is more than 10%, and an automatically chosen threshold (4), respectively. Schedule $\kappa^*\{v \mid E(\mathcal{D}) \leq 0.75\}$ is similar to $\kappa^*(v)$ except that the euclidean distance in (4) is minimized under the constraint that expected delay in detecting progression is at most 9 months (0.75 years). Annual corresponds to a schedule of yearly biopsies, and PRIAS corresponds to biopsies as per PRIAS protocol (Section 4).

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

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

Personalized schedules that we proposed require a risk threshold. We optimized the threshold choice using a generic utility function based on the expected number of biopsies and time delay in detecting progression. We used only these two measures because they are easy to interpret but simultaneously critical for deciding the timing of invasive tests.

Also, the time delay in detecting progression is an easily-quantifiable surrogate for the window of opportunity for curative treatment and additional benefits of observing progression early. Practitioners may extend/modify our utility function by adding to/replacing time delay with commonly used decision-theoretic measures such as quality-adjusted life-years/expectancy (QALY/QALE).

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

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

Data Availability

This simulation study utilized results from a statistical model fitted to the PRIAS dataset. The PRIAS database is not openly accessible. However, access to the database can be requested on the basis of a study proposal approved by the PRIAS steering committee. The website of the PRIAS program is `www.prias-project.org`. Instructions for generating a synthetic dataset are provided in the README file along with the source code.

Source Code and Supplementary Material

Supplementary sections referenced in this paper are available in the file titled ‘supplementary.pdf’. Source code is available at `https://anonymous.4open.science/r/d862487e-9a1a-4472-9564` and is also uploaded in a zip file along with the manuscript.

References

Alagoz, O., Ayer, T., and Erenay, F. S. (2011). *Operations Research Models for Cancer Screening*. American Cancer Society.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Steimle, L. N. and Denton, B. T. (2017). *Markov Decision Processes for Screening and Treatment of Chronic Diseases*. Springer International Publishing.

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

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

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

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

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

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

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

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

WHO, W. H. O. et al. (2014). *Global Status Report on Noncommunicable Diseases 2014*. Number WHO/NMH/NVI/15.1. World Health Organization.

Supplementary Materials for 'Personalized Schedules for Burdensome Surveillance Tests'

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

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

A.1 Longitudinal Sub-process

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

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

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

A.2 Survival Sub-process

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

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

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

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

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

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

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

A.3 Parameter Estimation

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

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

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

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

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

$$p(l_i, r_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \exp\Big[-\int_0^{l_i} h_i\{s \mid \mathcal{M}_i(t), \boldsymbol{w}_i(t)\}\mathrm{d}s\Big] - \exp\Big[-\int_0^{r_i} h_i\{s \mid \mathcal{M}_i(t), \boldsymbol{w}_i(t)\}\mathrm{d}s\Big]. \tag{1}$$

The integral in (1) does not have a closed-form solution, and therefore we use a 15-point Gauss-Kronrod quadrature rule to approximate it.

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

$$\alpha_s \sim \mathcal{N}(0, \tau\psi_s), \quad \tau^{-1} \sim \text{Gamma}(0.1, 0.1), \quad \psi_s^{-1} \sim \text{Gamma}(1, 0.01). \tag{2}$$

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

$$p(\gamma_{h_0} \mid \tau_h) \propto \tau_h^{\rho(\boldsymbol{K})/2} \exp\Big(-\frac{\tau_h}{2}\gamma_{h_0}^\top \boldsymbol{K} \gamma_{h_0}\Big) \tag{3}$$

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

B Joint Model for the PRIAS Dataset Used in Simulation Study

B.1 Dataset

This work uses a joint model fitted to the PRIAS dataset (Table 1). The PRIAS database is not openly accessible. However, access to the database can be requested on the basis of a study proposal approved by the PRIAS steering committee. The website of the PRIAS program is `www.prias-project.org`. We have presented the PRIAS based model's definition and parameter estimates below.

Table 1: **Summary of the PRIAS dataset**. The primary event of interest is cancer progression (increase in biopsy Gleason grade group from grade group 1 to 2 or higher). Abbreviations: PSA is prostate-specific antigen; DRE is digital rectal examination, with level T1c (Schröder et al., 1992) indicating a clinically inapparent tumor which is not palpable or visible by imaging, whereas tumors with DRE > T1c are palpable; IQR is interquartile range.

Characteristic	**Value**
Total patients	7813
progression (primary event)	1134
Treatment	2250
Watchful waiting	334
Lost to follow-up	203
Discontinued on request	46
Death (other)	95
Death (prostate cancer)	2
Total DRE measurements	37326
Total PSA measurements	67578
Total biopsies	15686
Median age at diagnosis (years)	66 (IQR: 61–71)
Median PSA (ng/mL)	5.7 (IQR: 4.1–7.7)
DRE = T1c (%)	34883/37326 (94%)
Median number of PSA per patient	6 (IQR: 4–12)
Median number of DRE per patient	4 (IQR: 2–7)
Median number of biopsies per patient	2 (IQR: 1–2)

B.2 Model Specification

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

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

$$\begin{aligned}\text{logit}\big[\Pr\{y_{di}(t) > \text{T1c}\}\big] &= \beta_{0d} + b_{0di} + (\beta_{1d} + b_{1di})t\\ &+ \beta_{2d}(\text{Age}_i - 65) + \beta_{3d}(\text{Age}_i - 65)^2 \end{aligned} \tag{4}$$

where, t denotes the follow-up visit time, and Age_i is the age of the i-th patient at the time of inclusion in AS. The fixed effect parameters are denoted by $\{\beta_{0d}, \ldots, \beta_{3d}\}$, and $\{b_{0di}, b_{1di}\}$ are the patient specific random effects. With this definition, we assume that the patient-specific log odds of obtaining a DRE measurement larger than T1c (Schröder et al., 1992), i.e., palpable tumor, remain linear over time.

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

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

where, $m_{pi}(t)$ denotes the underlying measurement error free value of $\log_2(\text{PSA} + 1)$ transformed (Pearson et al., 1994; Lin et al., 2000) measurements at time t. We model it nonlinearly over time using B-splines (De Boor, 1978). To this end, the B-spline basis function $B_k(t, \mathcal{K})$ has two internal knots at $\mathcal{K} = \{0.75, 2.12\}$ years (33.34 and 66.67 percentile of observed follow-up times), and boundary knots at 0 and 6.4 years (95-th percentile of the observed follow-up times). The fixed effect parameters are denoted by $\{\beta_{0p}, \ldots, \beta_{5p}\}$, and $\{b_{0pi}, \ldots, b_{3pi}\}$ are the patient specific random effects. The error $\varepsilon_{pi}(t)$ is assumed to be t-distributed with three degrees of freedom (Appendix B.4) and scale σ, and is independent of the random effects.

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

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

$$\begin{aligned} h_i(t) = h_0(t) \exp\Big(&\gamma_1(\text{Age}_i - 65) + \gamma_2(\text{Age}_i - 65)^2 \\ &+ \alpha_{1d} \text{logit}\big[\Pr\{y_{di}(t) > \text{T1c}\}\big] + \alpha_{1p} m_{pi}(t) + \alpha_{2p} \frac{\partial m_{pi}(t)}{\partial t}\Big), \end{aligned} \tag{6}$$

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

B.3 Parameter Estimates

Figure 1 shows the cumulative-risk of progression over the follow-up period. The posterior parameter estimates for the PRIAS based joint model are given in Table 4 (longitudinal sub-model for DRE outcome), Table 5 (longitudinal sub-model for PSA outcome) and Table 6 (relative risk sub-model). The parameter estimates for the variance-covariance matrix $\boldsymbol{W}$ from the longitudinal sub-model are shown in Table 3. As described in Appendix A the baseline hazard of the joint model model utilized a cubic P-spline. The knots of this P-spline were placed at the following time points: 0.000, 0.000, 0.000, 0.000, 0.401, 0.801, 1.202, 1.603, 2.003, 2.404, 2.805, 3.205, 3.606, 4.007, 4.407, 4.808, 5.209, 12.542, 12.542, 12.542, 12.542 The parameters of the fitted spline function are given in Table 2.

We present plots of observed DRE versus fitted probabilities of obtaining a DRE measurement larger than T1c, for nine randomly selected patients in Figure 2. Similarly observed versus fitted PSA profiles for nine randomly selected patients are shown in Figure 3.

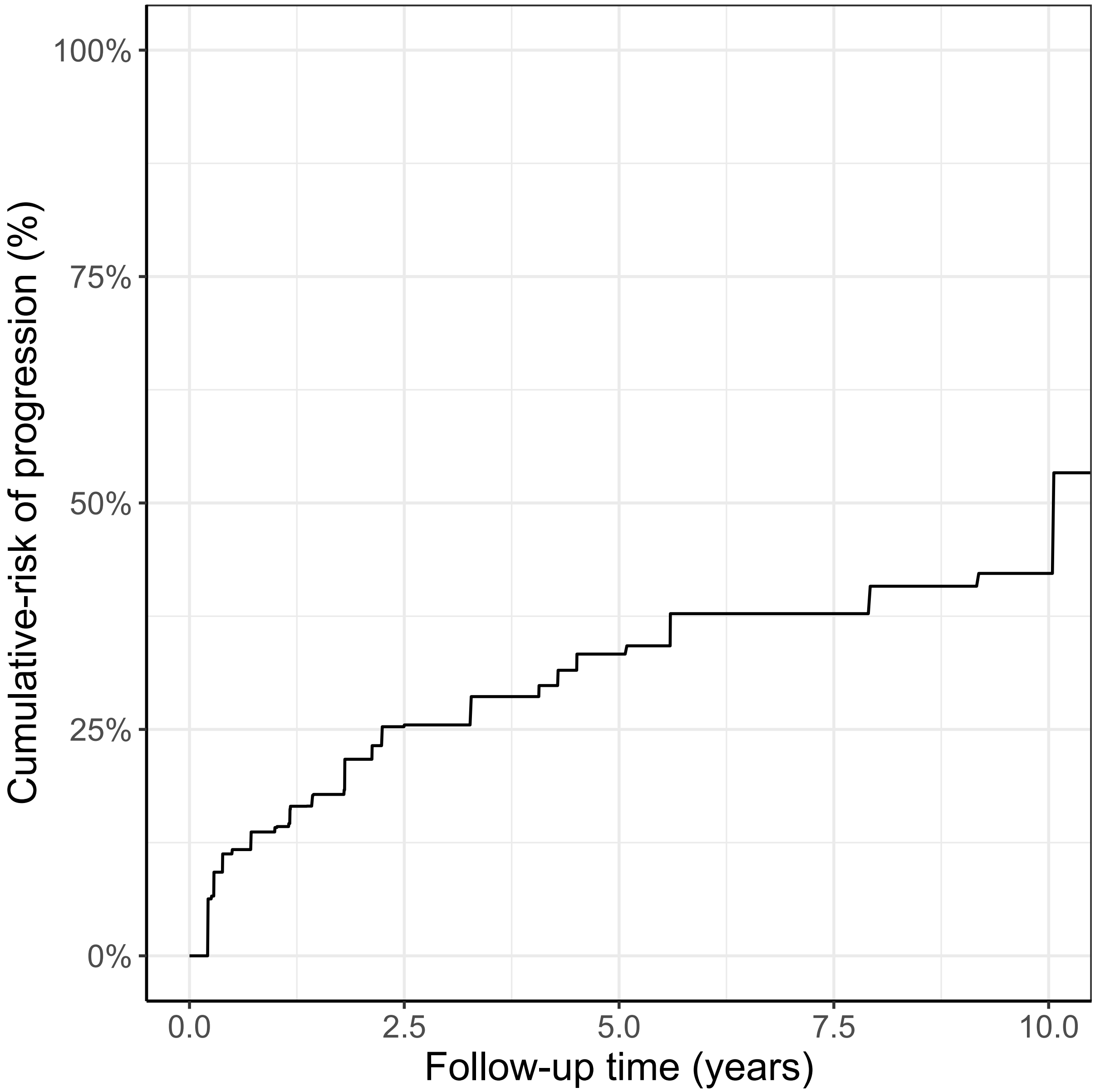

Figure 1: **Estimated cumulative-risk of cancer progression** for patients in the Prostate Cancer Research International Active Surveillance (PRIAS) dataset. Nearly 50% patients (*slow progressing*) do not progress in the ten year follow-up period. Cumulative risk is estimated using nonparametric maximum likelihood estimation (Turnbull, 1976), to account for interval censored progression times observed in the PRIAS dataset. Censoring includes death, removal from surveillance on the basis of observed longitudinal data, and patient dropout.

Table 2: Estimated parameters of the P-spline function utilized to model the baseline hazard $h_0(t)$ in joint model fitted to the PRIAS dataset. Parameters are named with the prefix ‘ps’ indicating P-spline parameter.

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

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

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

Table 4: Estimated mean and 95% credible interval for the parameters of the longitudinal sub-model (4) for the DRE outcome.

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

Table 5: Estimated mean and 95% credible interval for the parameters of the longitudinal sub-model (5) for the PSA outcome.

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

Table 6: Estimated mean and 95% credible interval for the parameters of the relative risk sub-model (6) of the joint model fitted to the PRIAS dataset.

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

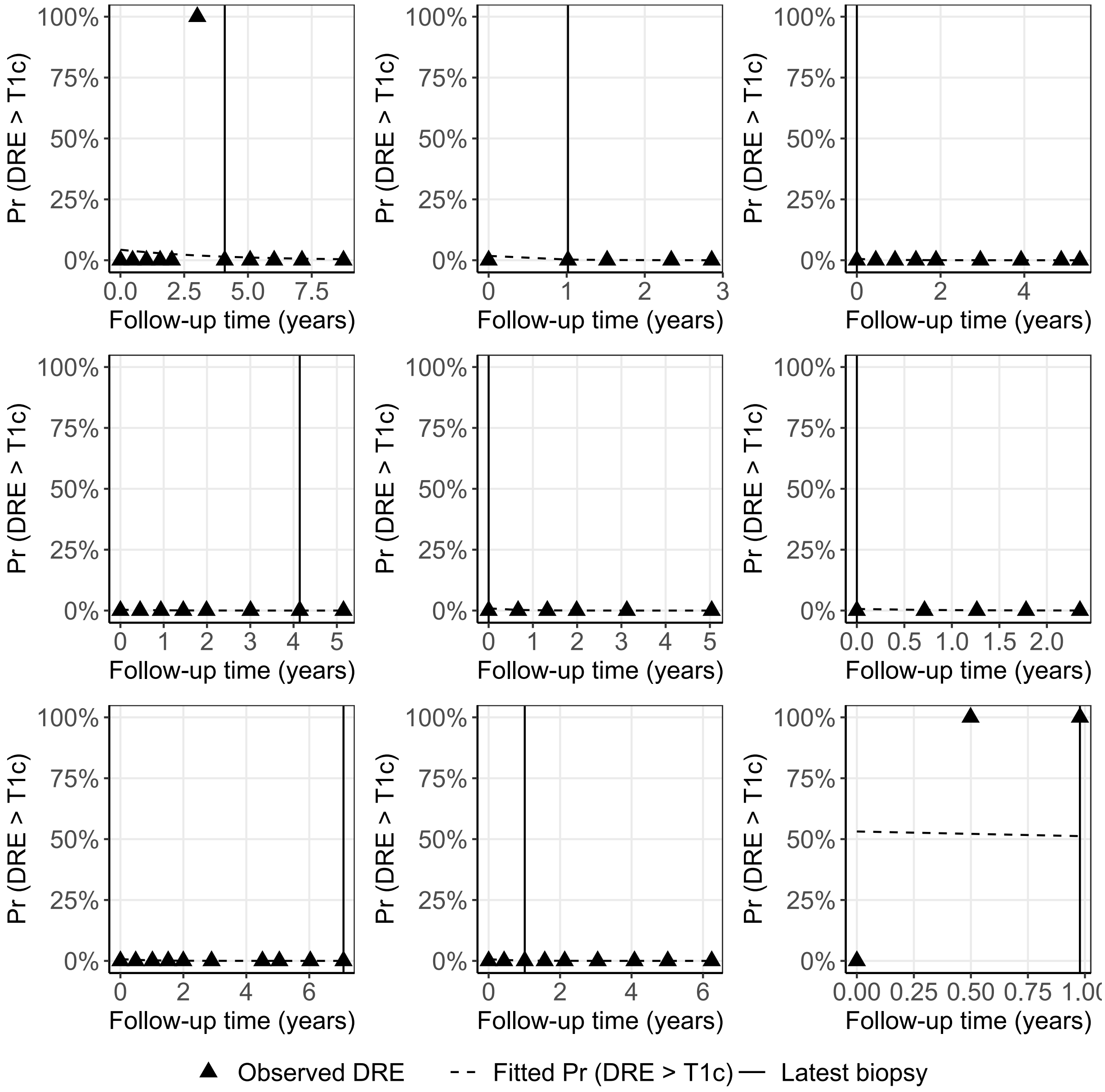

Figure 2: Observed DRE versus fitted probabilities of obtaining a DRE measurement larger than T1c, for nine randomly selected PRIAS patients. The fitted profiles utilize information from the observed DRE measurements, PSA measurements, and time of the latest biopsy. Observed DRE measurements plotted against 0% probability are equal to T1c. Observed DRE measurements plotted against 100% probability are larger than T1c.

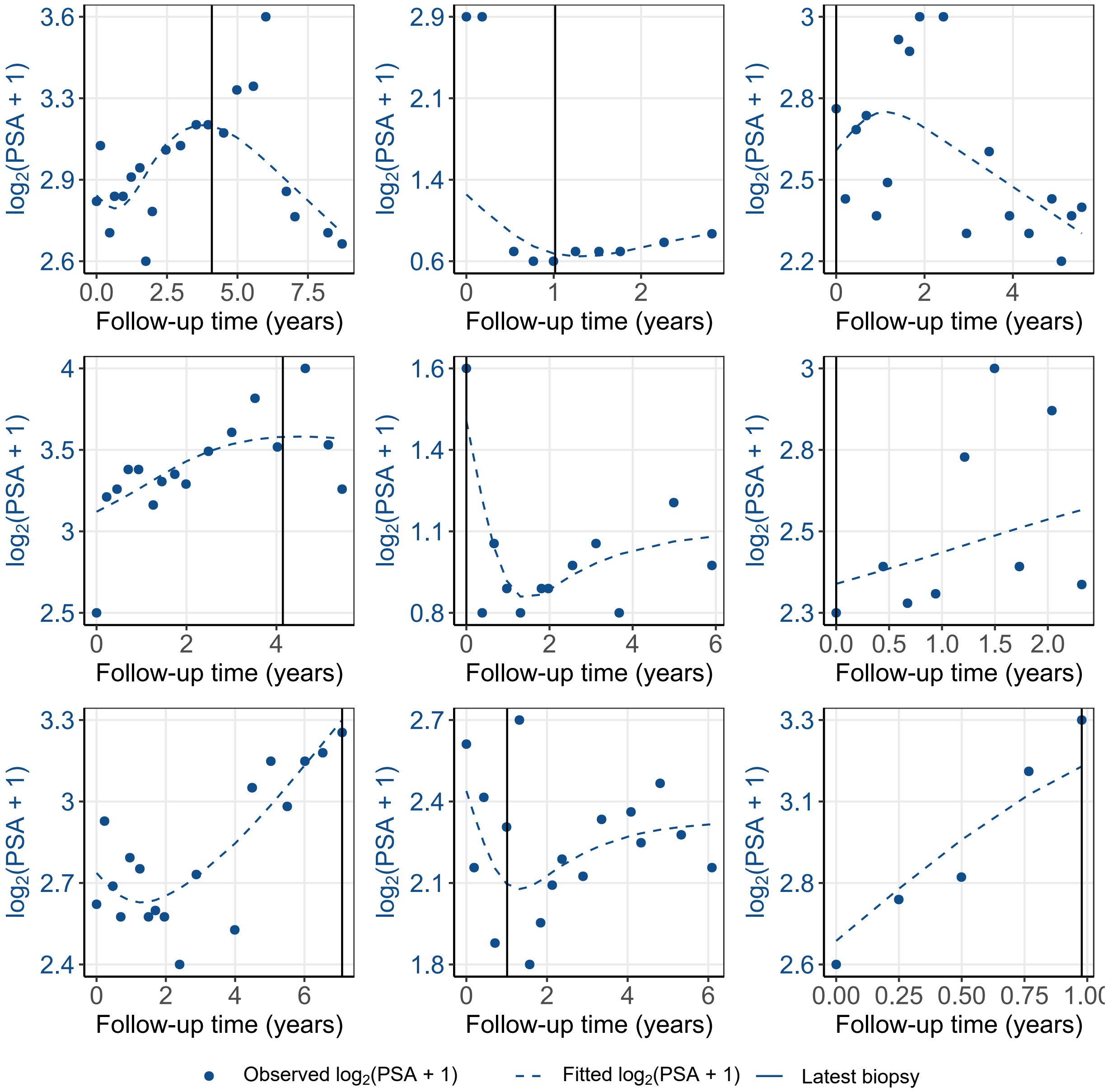

Figure 3: Fitted versus observed $\log_2(\text{PSA} + 1)$ profiles for nine randomly selected PRIAS patients. The fitted profiles utilize information from the observed PSA measurements, DRE measurements, and time of the latest biopsy.

Table 7: Data of the demonstration patient in Figure 5 of the main manuscript. Age of the patient at baseline was 60 years and time of last negative biopsy was 3.5 years. DRE: digital rectal examination.

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

B.4 Assumption of t-distributed (df=3) Error Terms

With regards to the choice of the distribution for the error term ε_p for the PSA measurements (5), 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 4, shows that the assumption of t-distributed (df=3) errors is reasonably met by the fitted model.

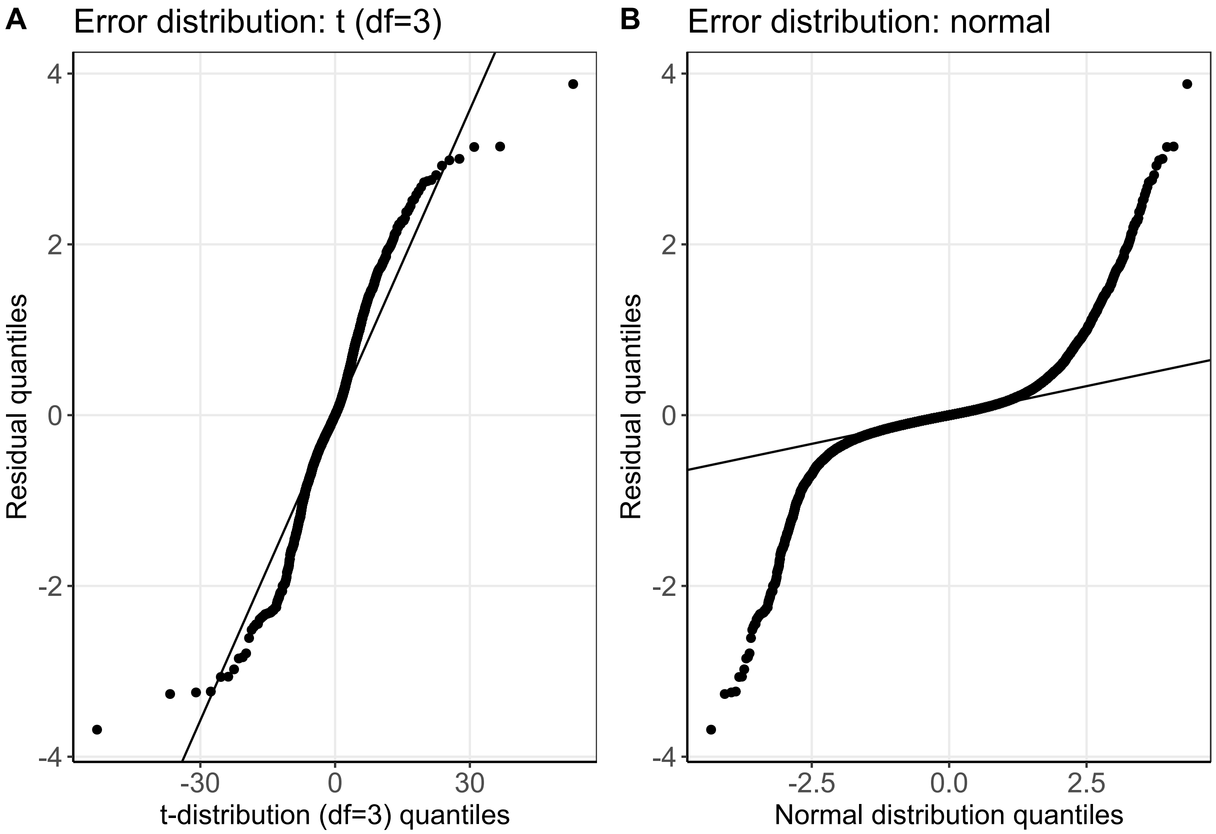

Figure 4: Quantile-quantile plot of subject-specific residuals from the joint models fitted to the PRIAS dataset. **Panel A**: model assuming a t-distribution (df=3) for the error term ε_p. **Panel B**: model assuming a normal distribution for the error term ε_p.

B.5 Predictive Performance of the PRIAS based Model

We calculate the predictive performance of the PRIAS based joint model using time-dependent area under the receiver operating characteristic curve or AUC (measure of discrimination), and the mean absolute prediction error or MAPE. Mathematical calculations for these in the joint modeling framework are detailed in Rizopoulos et al. (2017). Because these are temporal extensions of their standard versions (Steyerberg et al., 2010) in a longitudinal setting, at every six months of follow-up (standard visit times in PRIAS), we calculated a unique AUC and MAPE for predicting risk of progression in the subsequent one year (recommended time gap between subsequent biopsies). For emulating a realistic situation, we calculated the AUC and MAPE at each follow-up using only the validation data available until that follow-up. For example, calculations for AUC and MAPE for the time interval year two to year three do not utilize data of patients who progressed before year two. The AUC and MAPE for our model are shown in Table 8.

Table 8: Follow-up time dependent, area under the receiver operating characteristic curves (AUC), and mean absolute prediction error (MAPE), with bootstrapped 95% confidence interval in brackets. The choice of year six as the maximum follow-up period is based on the reasoning that it is roughly the 95-percentile of observed follow-up times.

Follow-up period (years)	AUC (95% CI)	MAPE (95%CI)
0.0 to 1.0	0.658 [0.620, 0.693]	0.234 [0.229, 0.240]
0.5 to 1.5	0.648 [0.631, 0.663]	0.220 [0.213, 0.226]
1.0 to 2.0	0.624 [0.600, 0.644]	0.151 [0.147, 0.155]
1.5 to 2.5	0.649 [0.604, 0.704]	0.127 [0.118, 0.134]
2.0 to 3.0	0.683 [0.629, 0.729]	0.134 [0.121, 0.143]
2.5 to 3.5	0.681 [0.604, 0.739]	0.115 [0.105, 0.128]
3.0 to 4.0	0.647 [0.600, 0.710]	0.079 [0.073, 0.087]
3.5 to 4.5	0.630 [0.583, 0.668]	0.095 [0.089, 0.101]
4.0 to 5.0	0.614 [0.557, 0.659]	0.104 [0.098, 0.111]
4.5 to 5.5	0.615 [0.541, 0.702]	0.101 [0.088, 0.114]
5.0 to 6.0	0.617 [0.550, 0.713]	0.102 [0.086, 0.121]

C Simulation Study

In the simulation study, we evaluated the following biopsy schedules: biopsy every year (annual), biopsy according to the PRIAS schedule (PRIAS), personalized biopsy schedules based on two fixed risk thresholds, namely, $\kappa = 10\%$, and automatically chosen $\kappa^*(v)$ (Section 3 of main manuscript), and automatically chosen $\kappa^*\{v \mid E(D) \leq 0.75\}$ with a constraint of 0.75 years (9 months) on expected delay in detecting progression. The choice of 0.75 years delay constraint is arbitrary and is only used to illustrate that applying the constraint limits the average delay at 0.75 years. We compare all the aforementioned schedules on two criteria, namely the number of biopsies they schedule and the corresponding time delay in detection of cancer progression, in years (time of positive biopsy - true time of cancer progression). The corresponding results, using 500 × 250 test patients are presented in 9. Since the simulated cohorts are based on PRIAS, roughly only 50% of the patients progress in the ten year study period. While, we are able to calculate total number of biopsies scheduled in all 500 × 250 test patients, but the time delay in detection of progression is available only for those patients who progress in ten years (*progressing*). Hence, we show the simulation results separately for *progressing* and *non-progressing* patients.

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

Progressing patients (50%)								
Schedule	$\mathrm{Q}_1^{\mathrm{nb}}$	μ^{nb}	$\mathrm{Med}^{\mathrm{nb}}$	$\mathrm{Q}_3^{\mathrm{nb}}$	$\mathrm{Q}_1^{\mathrm{d}}$	μ^{d}	$\mathrm{Med}^{\mathrm{d}}$	$\mathrm{Q}_3^{\mathrm{d}}$
Annual	1	3.71	3	6	0.29	0.55	0.57	0.82
PRIAS	1	2.88	2	4	0.38	0.92	0.74	1.00
$\kappa = 10\%$	1	2.55	2	4	0.45	1.00	0.85	1.33
$\kappa^*(v)$	1	2.46	2	3	0.45	0.89	0.86	1.26
$\kappa^*\{v \mid E(D) \leq 0.75\}$	1	3.39	3	5	0.32	0.61	0.63	0.88
Non-progressing patients (50%)								
Annual	10	10.00	10	10	-	-	-	-
PRIAS	4	6.40	6	8	-	-	-	-
$\kappa = 10\%$	4	4.91	5	6	-	-	-	-
$\kappa^*(v)$	6	6.22	6	7	-	-	-	-
$\kappa^*\{v \mid E(D) \leq 0.75\}$	8	8.68	9	9	-	-	-	-

D Source Code

Source code URL: `https://anonymous.4open.science/r/d862487e-9a1a-4472-9564-ff2be4c625fd/`.

It has also been uploaded as a zip file with the manuscript.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Journal of the American Statistical Association

Author Contributions Checklist Form

This form documents the artifacts associated with the article (i.e., the data and code supporting the computational findings) and describes how to reproduce the findings.

Commented [A1]: This document will be read by readers after an accepted manuscript is published so please phrase your responses for future readers. Please see https://jasa-acs.github.io/repro-guide/pages/acc.html for details, links to examples, and information on how to save your completed form as a PDF. In particular make sure to remove these comments when saving as a PDF.

Part 1: Data

☐ This paper **does not** involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

Commented [A2]: If this box is checked, please skip directly to the Code section. Otherwise, continue.

☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

Commented [A3]: A short (< 100 words) description of the data. More details can be provided in files accompanying the data.

In this work, we developed personalized test schedules for invasive diagnostic tests in the surveillance of chronic non-communicable diseases. We demonstrate our methodology for the prostate cancer surveillance scenario, using a model fitted to the world's largest prostate cancer active surveillance dataset abbreviated as PRIAS, dated April 2019. More than 100 medical centers from 17 countries contribute to PRIAS using a common study protocol, and web-based tool, both available at www.prias-project.org. The PRIAS dataset consists of patient age at inclusion, and repeated measurements data on biopsy Gleason grade groups (invasive diagnostic test), prostate-specific antigen or PSA (continuous: biomarker ng/mL), digital rectal examination or DRE (binary: tumor palpable or not), and magnetic resonance imaging or MRI (PI-RADS score).

Availability

☐ Data **are** publicly available
☒ Data **cannot be made** publicly available

If the data are publicly available, see the *Publicly available data* section. Otherwise, see the *Non-publicly available data* section, below.

Publicly available data

☐ Data are available online at:
☐ Data are available as part of the paper's supplementary material.

☐ Data are publicly available by request, following the process described here:

Commented [CP4]: If the data are only available by request to the authors or some other data owner, please make sure to explain the process of requesting access to the data.

☐ Data are or will be made available through some other mechanism, described here:

Non-publicly available data

Commented [A5]: If the data for this manuscript are publicly available, skip to the Description section below. Otherwise, continue.

The Journal of the American Statistical Association requires authors to make data accompanying their papers available to the scientific community except in cases where: 1) public sharing of data would be impossible, 2) suitable synthetic data are provided which allow the main analyses to be replicated (recognizing that results may differ from the "real" data analyses), and 3) the scientific value of the results and methods outweigh the lack of reproducibility.

Discussion of lack of publicly available data:

Commented [A6]: For example:

- why data sharing is not possible,
- what synthetic data are provided, and
- why the value of the paper's scientific contribution outweighs the lack of reproducibility.

The PRIAS dataset is not openly accessible, but access to the dataset can be requested based on a study proposal approved by the PRIAS steering committee. The website of the PRIAS program is www.prias-project.org. We do provide a synthetic dataset instead of the original dataset. Readers can use this dataset along with the provided code to understand and recreate the workflow required to make personalized schedules in their datasets.

The **main contribution of our work** is not the model fitted to the PRIAS dataset, but rather the methodology proposed to schedule invasive diagnostic tests. The methodology is generic for use in other surveillance scenarios (e.g., scheduling endoscopies in Barett's esophagus). Hence, the lack of direct public access to the prostate cancer dataset PRIAS, which we only used for demonstration purposes, is not a major limitation. The simulation study we conducted utilizes the parameter estimates of the model fitted to the PRIAS dataset and can be reproduced without access to the PRIAS dataset itself.

Description

File format(s)

Commented [CP7]: Check all that apply.

☐ CSV or other plain text:

☒ Software-specific binary format (.Rda, Python pickle, etc.):

☐ Standardized binary format (e.g., netCDF, HDF5, etc.):

☐ Other (described here):

SPSS .sav file

Data dictionary

Commented [CP8]: A data dictionary provides information that allows users to understand the meaning, format, and use of the data.

☐ Provided by the authors in the following file(s):

☐ Data file(s) is (are) self-describiing (e.g., netCDF files)

☒ Available at the following URL:

https://anonymous.4open.science/r/d862487e-9a1a-4472-9564-ff2be4c625fd/

Additional information (optional)

Commented [CP9]: Provide any details that would be helpful in understanding the data. If relevant, provide unique identifier/DOI/version information and/or license/terms of use.

Part 2: Code

Abstract

Commented [A10]: A short (< 100 words) description of the code. If necessary, more details can be provided in files that accompany the code.

We provide R code for the following purposes:

1. To generate synthetic PRIAS like datasets.
2. For cleaning the original PRIAS dataset for readers who obtain a copy of the raw dataset from the PRIAS consortium.
3. To fit a joint model to the synthetic or real PRIAS dataset using the R package **JMBayes**.
4. For reproducing the simulation study results comparing different biopsy schedules in PRIAS.
5. A generic API that can be used to schedule invasive tests in a personalized manner for any surveillance scenario. This API is compatible with R joint model objects fitted using the JMbayes package.
6. Code to reproduce figures.

Description

Code format(s)

Commented [CP11]: Check all that apply.

☒ Script files

☒ R ☐ Python ☐ Matlab

☐ Other:

☐ Package

☐ R ☐ Python ☐ MATLAB toolbox

☐ Other:

☐ Reproducible report

☐ R Markdown ☐ Jupyter notebook

☐ Other:

☒ Shell script

☐ Other (described here):

https://anonymous.4open.science/r/d862487e-9a1a-4472-9564-ff2be4c625fd/

Supporting software requirements

Version of primary software used

Commented [CP12]: For example, R version 3.6.2.

R 3.6.1

Libraries and dependencies used by the code

Commented [CP13]: Include version numbers (e.g., version numbers for any R or Python packages used)

In total the provided source code depends on four R packages:

- JMBayes
- survival
- MASS
- splines: part of the R core now.

Supporting system/hardware requirements (optional)

Commented [A14]: System/hardware requirements including operating system with version number, access to cluster, GPUs, etc.

The code has been tested to work with 64 bit R-3.6.1 installed on a laptop with 4th generation Intel Core-i7 processor, and 8GB RAM.

Parallelization used

☐ No parallel code used
☒ Multi-core parallelization on a single machine/node
Number of cores used: 2 to 10, users can customize it in our shell script
☐ Multi-machine/multi-node parallelization
Number of nodes and cores used:

License

☐ MIT License (default)
☐ BSD
☐ GPL v3.0

☒ Creative Commons

☐ Other (described here):

Additional information (optional)

Commented [CP15]: By default, submitted code will be published on the JASA Github repository (https://github.com/jasa-acs) as well as in the supplementary material. Authors are encouraged to also make their code available in a public repository. If relevant, please provide unique identifier/DOI/version information here.

Part 3: Reproducibility workflow

Commented [CP16]: The materials provided should provide a straightforward way for reviewers and readers to reproduce analyses with as few steps as possible.

Scope

The provided workflow reproduces:
☐ Any numbers proviided in text in the paper
☒ All tables and figures in the paper
☐ Selected tables and figures in the paper, as explained and justified here:

Code for Figure 1 to 4 is not provided because they are only illustrative. Code for Figure 5 and Figure 6 has been provided. Supplementary tables showing parameter estimates are based on the MCMC samples of the posterior distribution of the parameters of our model. The fitted model object has been provided with the manuscript.

Workflow details

Format(s)

Commented [CP17]: Check all that apply

☐ Single master code file
☒ Wrapper (shell) script(s)
☐ Self-contained R Markdown file, Jupyter notebook, or other literate programming approach
☒ Text file (e.g., a readme-style file) that documents workflow
☐ Makefile
☐ Other (more detail in 'Instructions' below)

Instructions

Commented [A18]: Describe how to use the materials provided to reproduce analyses in the manuscript. Additional details can be provided in file(s) accompanying the reproducibility materials.

https://anonymous.4open.science/r/d862487e-9a1a-4472-9564-ff2be4c625fd/
Contains the README file and source code.

Expected run-time

Approximate time needed to reproduce the analyses on a standard desktop machine:
☐ <1 minute
☐ 1-10 minutes
☐ 10-60 minutes

☐ 1-8 hours

☐ >8 hours

☒ Not feasible to run on a desktop machine, as described here:

Data analysis takes 1 to 8 hours, but simulation study requires a server computer. Computational times are provided in the README file.
https://anonymous.4open.science/r/d862487e-9a1a-4472-9564-ff2be4c625fd/

Additional documentation (optional)

Commented [A19]: Additional documentation provided (e.g., R package vignettes, demos or other examples) that show how to use the provided code/software in other settings.

Notes (optional)

Commented [A20]: Any other relevant information not covered on this form. If reproducibility materials are not publicly available at the time of submission, please provide information here on how the reviewers can view the materials (and make sure to remove this information when submitting the final version of this form for an accepted manuscript).