Personalized Schedules for Burdensome Surveillance Tests

Anirudh Tomer[*1], Daan Nieboer[2], Monique J. Roobol[3], Ewout W. Steyerberg[2,4], and Dimitris Rizopoulos[1]

[1]Department of Biostatistics, Erasmus University Medical Center, the Netherlands
[2]Department of Public Health, Erasmus University Medical Center, the Netherlands
[3]Department of Urology, Erasmus University Medical Center, the Netherlands
[4]Department of Biomedical Data Sciences, Leiden University Medical Center, the Netherlands

June 5, 2020

Abstract

Benchmark surveillance *tests* for diagnosing disease *progression* (biopsies, endoscopies, etc.) in early-stage chronic non-communicable disease patients (e.g., cancer, lung diseases) are usually invasive. For detecting progression timely, over their lifetime, patients undergo numerous invasive tests planned in a fixed one-size-fits-all manner (e.g., biannually). 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 studies.

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 particular *threshold* (e.g., 5% risk). This schedule is updated on each follow-up with newly gathered data. 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 quantities in a patient-specific manner for following any schedule by utilizing the predicted risk profile of the patient. Patients/doctors can employ these quantities to compare various personalized and fixed schedules objectively. Last, we implement our methodology in a web-application for prostate cancer patients.

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. Typically the benchmark tests used for confirming progression are *invasive*. Some examples of such tests are, 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 lung transplant (McWilliams et al. 2008) patients.

Invasive tests are repeated until progression is observed, typically as per a one-size-fits-all *fixed schedule*, e.g., biannually, (McWilliams et al. 2008, Bokhorst et al. 2015, Krist et al. 2007). With recurrent tests, progression is always detected with a time delay (Figure 1). A shorter delay in detecting progression (*benefit*) can provide a larger window of opportunity for curative treatment. However, with a fixed schedule, this means conducting tests frequently. Frequent tests are *burdensome* as they may cause pain and/or severe medical complications (Loeb et al. 2013, Krist et al. 2007). Consequently, patients may not always comply with frequent tests (Bokhorst et al. 2015, Le Clercq et al. 2015). In this regard, since one-size-fits-all fixed schedules do not differentiate between fast and slow/non-progressing patients (large proportion in some diseases), they often have a skewed burden-benefit ratio.

The goal of this work (Figure 1) is to optimize the number of invasive tests and the time delay in detecting progression 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 invasive test results; and longitudinal
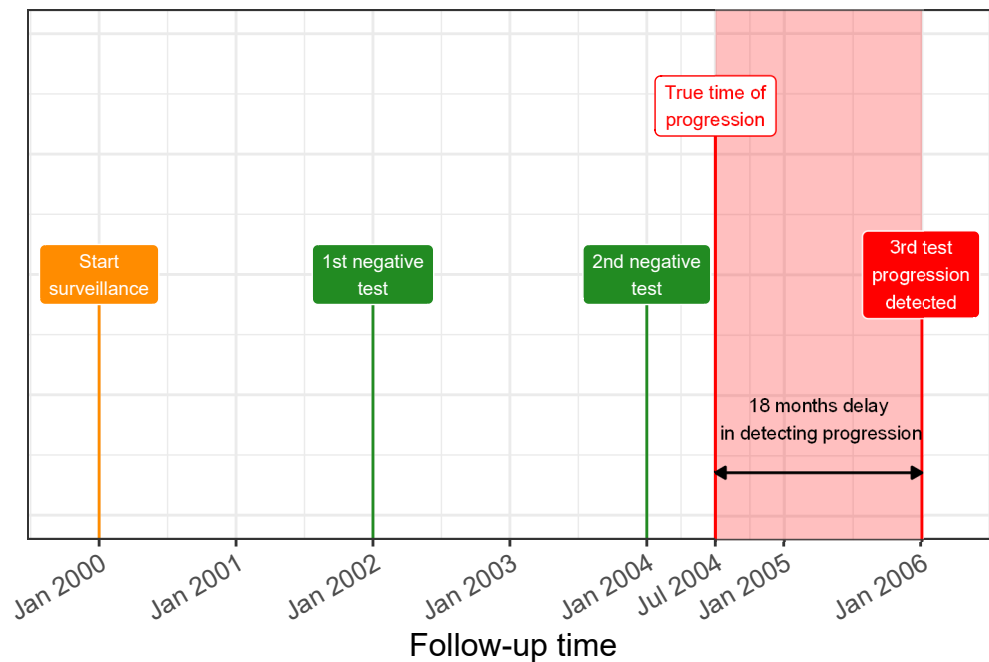
3
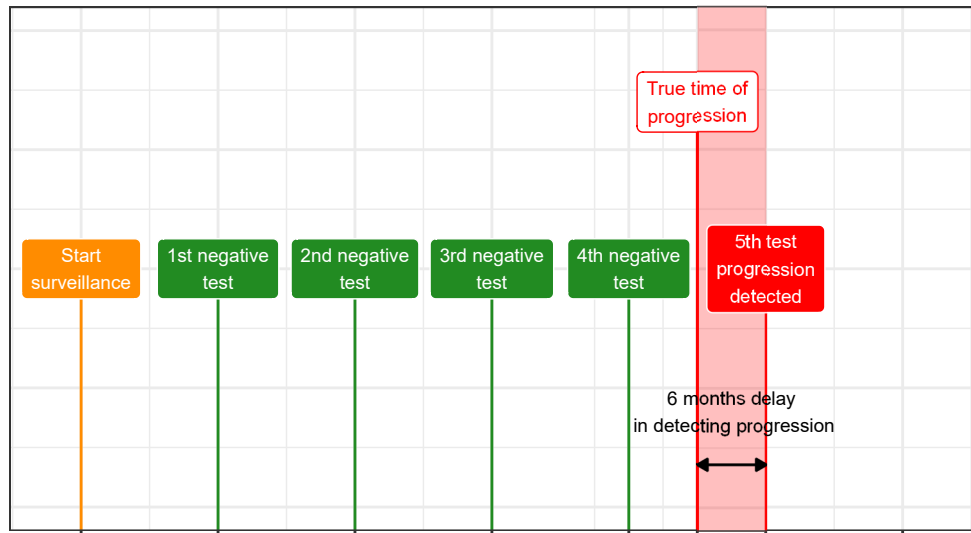
Figure 1: **Goal: Finding the optimal tradeoff between the number of invasive tests (burden) and time delay in detecting progression (shorter is beneficial)**. A progression is a non-terminal event in the surveillance of early-stage chronic non-communicable diseases. The true time of progression for 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**.

outcomes such as biomarkers, physical examination, and medical imaging measurements. Previous attempts at personalized scheduling can be divided into three categories. First, heuristic approaches such as decision making flowcharts (Bokhorst et al. 2015, Weusten et al. 2017). However, flowcharts discretize continuous clinical outcomes, often exploit only the last measurement, and ignore the measurement error in observed data. Second, partially observable Markov decision processes (Alagoz et al. 2010, Steimle & Denton 2017) for personalizing test decisions. Although the curse of dimensionality limits their application with continuous longitudinal outcomes. Third, personalized schedules may be obtained by optimizing an explicit utility function of the clinical parameters of interest (Bebu & Lachin 2017, Rizopoulos et al. 2015), including our previous work on scheduling biopsies in prostate cancer (Tomer et al. 2019, 2020). In this work, we will employ the third approach.

Our solution 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 & Davidian 2004, Rizopoulos 2012) because they are inherently personalized. Specifically, joint models utilize patient-specific random effects (McCulloch & Neuhaus 2005) to model longitudinal outcomes without discretizing them. Subsequently, we input the accumulated clinical data of a new patient into the fitted model to obtain their patient-specific cumulative-risk of progression at their current and future follow-up visits. We then create personalized schedules by planning tests on future visits where the predicted conditional 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 based on different risk thresholds. We estimate these two quantities for any

given schedule in a patient-specific manner using the predicted risk profile of the patient. 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, called prostate cancer research international active surveillance (Bokhorst et al. 2015), or PRIAS. It has 7813 patients (1134 progressions) with 104904 longitudinal measurements (Tomer et al. 2020). These patients have low/very-low grade cancer, often over-diagnosed due to prostate-specific antigen (PSA) based screening (Loeb, Bjurlin, Nicholson, Tammela, Penson, Carter, Carroll & Etzioni 2014). Surveillance aims to delay serious treatments (e.g., surgery, radiotherapy) until progression is observed. Consequently, patients are monitored routinely via 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, Carter, Schwartz, Fagerlin, Braithwaite & Lepor 2014). 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. We describe the personalized scheduling methodology in Section 3, and demonstrate them for prostate cancer surveillance patients in Section 4. In Section 5, we compare personalized and fixed schedules via a simulation study based on a joint model fitted to the PRIAS dataset.

6

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 different longitudinal outcomes in a unified framework, a joint model employs individual generalized linear mixed sub-models (McCulloch & 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, the complete vector of random effects $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}^\top, \ldots, \boldsymbol{b}_{Ki}^\top)^\top$ is assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix W.

2.2 Survival Sub-process

In the survival sub-process, 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:

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

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:

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

These formulations of $f(\cdot)$ postulate that the hazard of progression at time t may depend on underlying level $m_i(t)$ of the longitudinal outcome at t, or on both the level and velocity $m_i'(t)$ (e.g., PSA value and velocity in prostate cancer) of the outcome at t. Lastly, the baseline hazard $h_0(t)$ is modeled flexibly using P-splines (Eilers & 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 Web-Appendix 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^*. For this purpose, our calculations are based on the dynamic *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 defined as:

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

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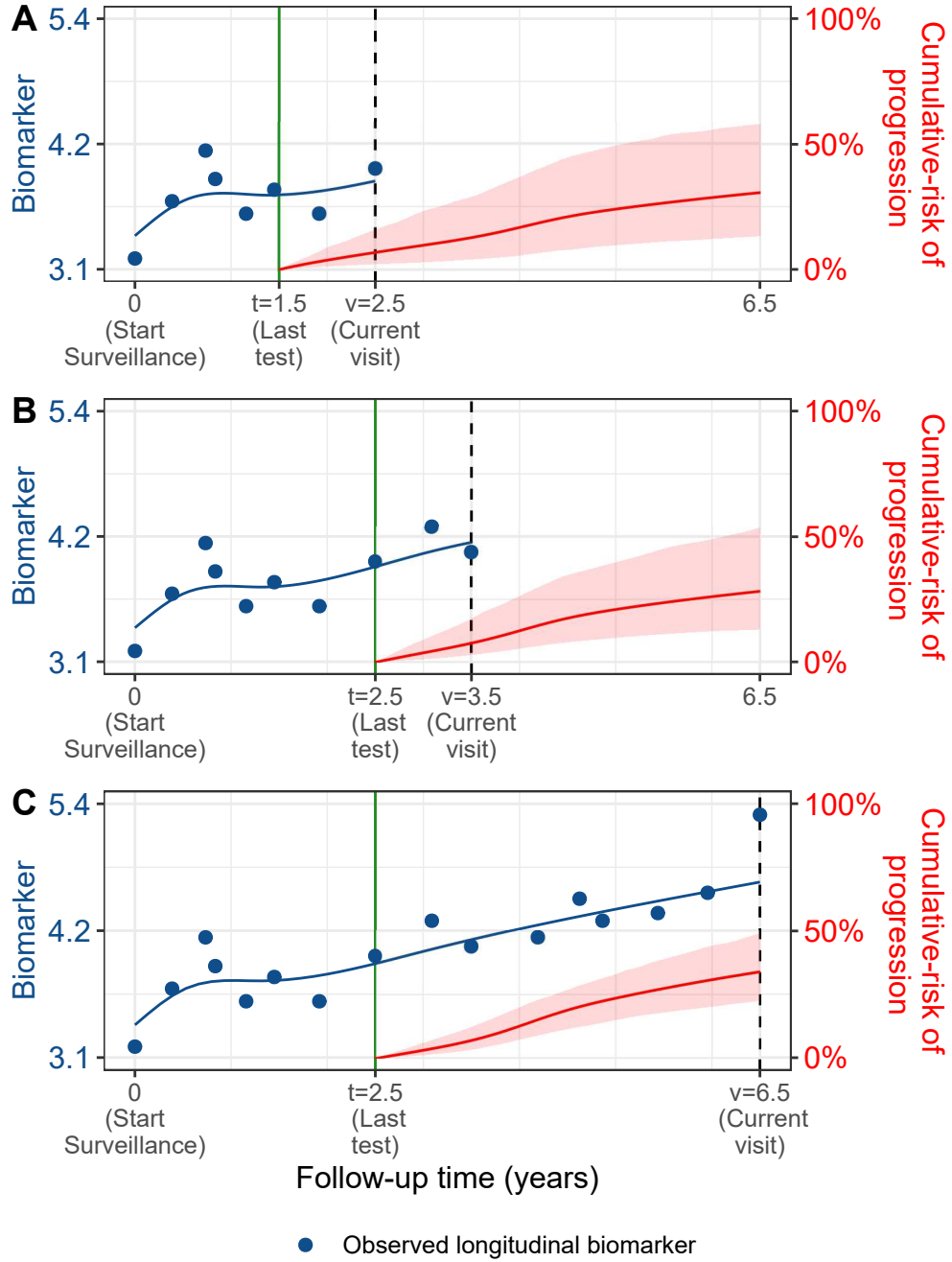
Figure 2: **Cumulative-risk of progression updated dynamically over follow-up as more patient data is gathered.** A single longitudinal outcome, namely, a continuous biomarker of disease progression, is used for illustration. **Panels A, B and C:** are ordered by the time of the current visit v (dashed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

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

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. In this regard, 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:

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

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^*
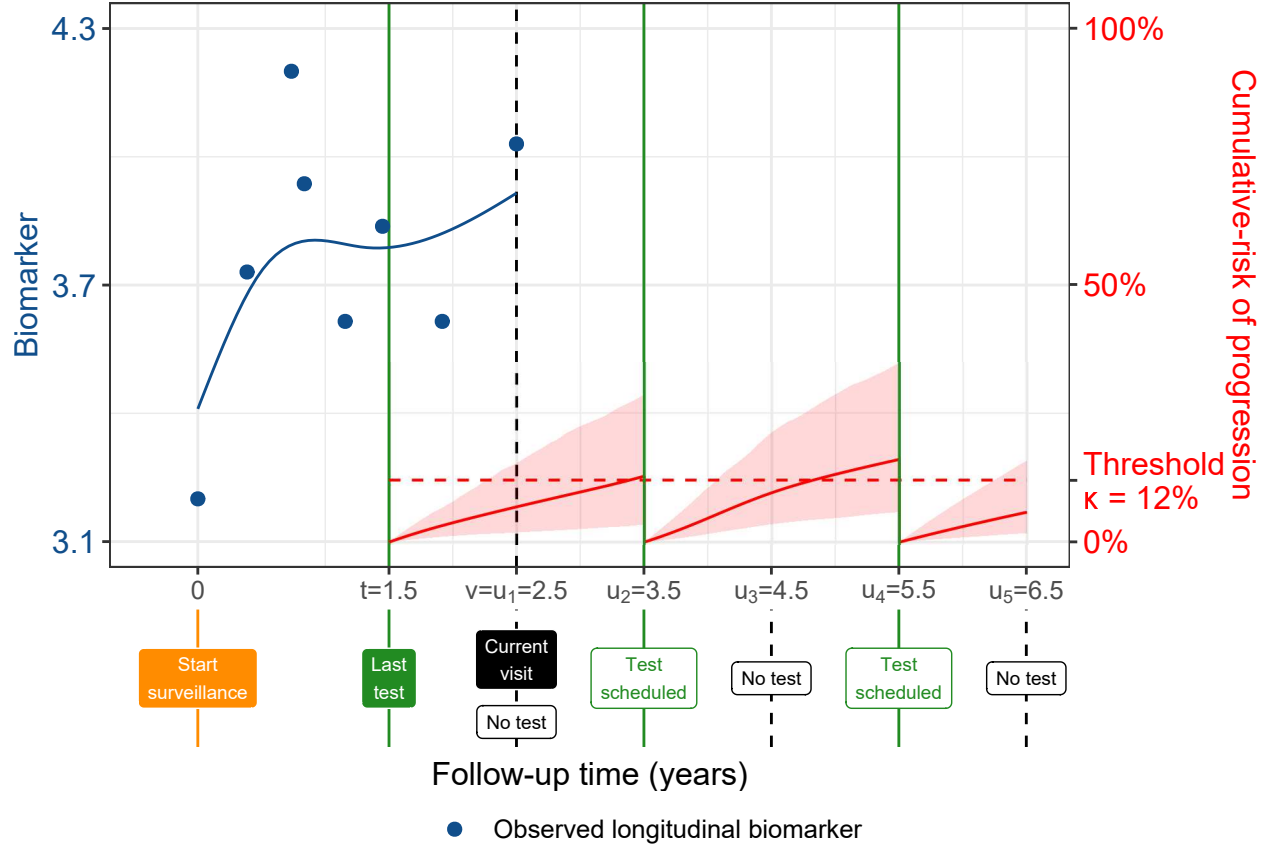
12

Figure 3: **Successive personalized test decisions (2) based on patient-specific cumulative-risk of progression**. Time of current visit: $v = 2.5$ years (black dashed 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 (observed: blue dots, fitted: blue line). 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 red line (95% credible interval shaded), and are updated each time a test is planned. 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.

13

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 $E\{\mathcal{N}_j(S_j^\kappa)\}$, given by the expression:

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

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:

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

14

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 time delay (Figure 1) that may occur 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

15

personalized test schedules obtained by varying κ in $[0, 1]$. An example of such a space is given in Figure 4.

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

$$\kappa^*(v) = \arg\min_{0 \leq \kappa \leq 1} \sqrt{\left[E\{\mathcal{N}_j(S_j^\kappa)\} - 1\right]^2 + \left[E\{\mathcal{D}_j(S^\kappa)\} - 0\right]^2}. \qquad (4)$$

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 & Wong 1994).

We considered shorter delays in detecting progression as the benefit of repeated tests. However, it is also common to describe the benefit of testing in terms of decision-theoretic measures such as quality-adjusted life-years/expectancy (QALY/QALE) gained (Sassi 2006). Optimizing (4) 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.
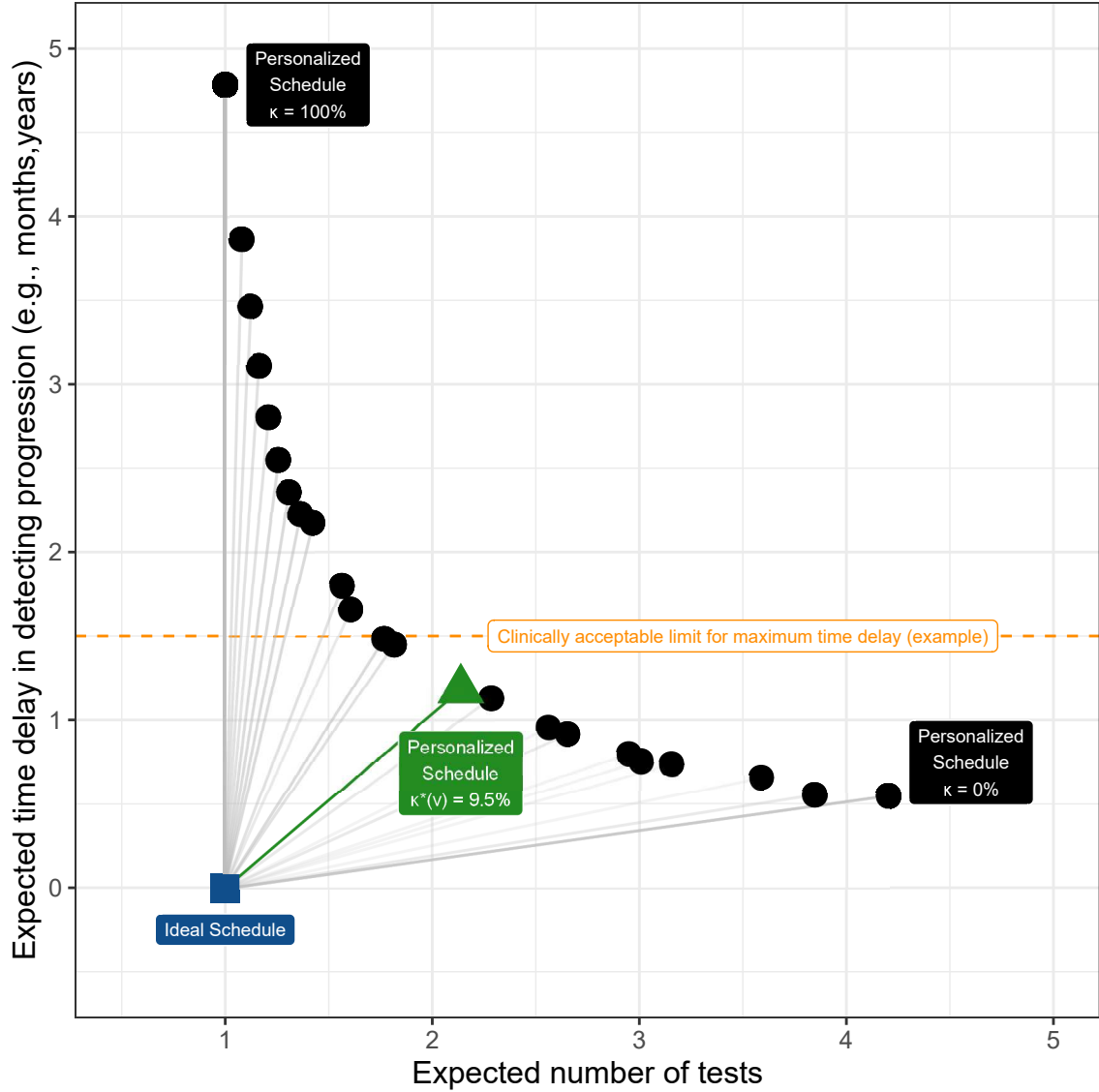
Figure 4: **Optimal current-visit time v specific risk threshold $\kappa^*(v)$ obtained using (4)** for the patient shown in Figure 3. Ideal schedule of tests: point $(1,0)$ shown as a blue square. It plans exactly one invasive test at the true time of progression T_j^* of a patient. Hence, the time 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 (shown with a 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 (orange dashed line).

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 reuse results from a joint model we previously fitted (Tomer et al. 2019) to the PRIAS dataset introduced in Section 1. This model 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). The model employed B-splines (De Boor 1978) to accommodate non-linear PSA evolution over follow-up. 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 (Tomer et al. 2019) using the R package **JM-bayes** (Rizopoulos 2016). While the complete model definition and parameter estimates are provided in Tomer et al. (2019), 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% (Web-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 area under the receiver operating characteristic curve (AUC) over the follow-up period (between 0.61 and 0.68). The mean absolute prediction error 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.

4.1 Personalized Biopsy Schedules for a Demonstration Prostate Cancer Patient

We utilized the joint model fitted to the PRIAS dataset to schedule biopsies in a demonstration prostate cancer patient shown in Figure 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.
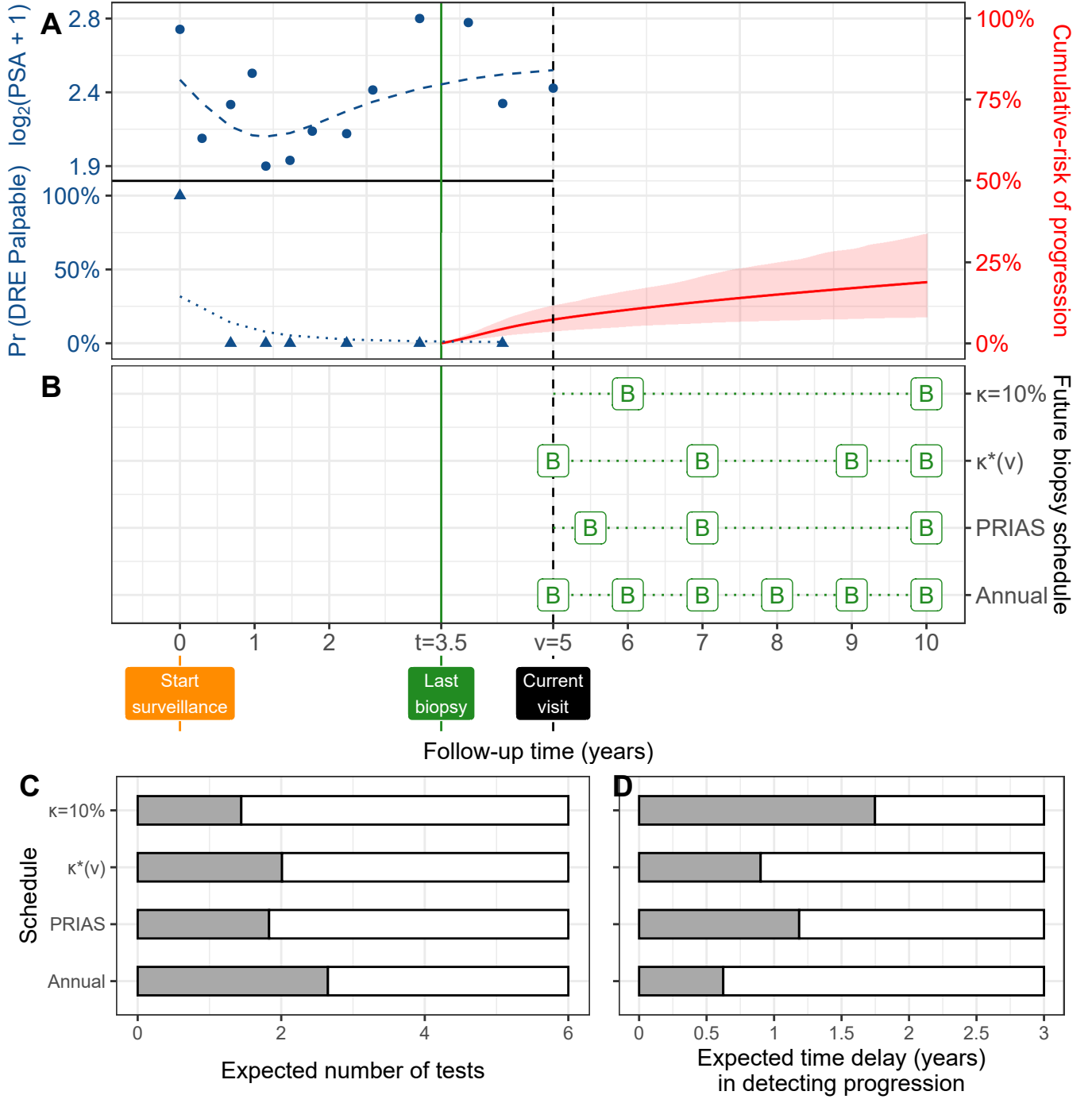
19

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: $\log_2(\text{PSA} + 1)$ trans-
formed (Tomer et al. 2019) PSA (observed: blue dots, fitted: dashed blue line), and binary
DRE (observed: blue triangles, fitted probability: dotted blue line). Cumulative-risk pro-
file: solid red line (95% credible interval shaded). **Panel B**: Different biopsy schedules are
shown with a 'B' indicating a biopsy. $\kappa = 10\%$ 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 an

20

4.2 Web-application for Personalized Biopsy Schedules Prostate Cancer Surveillance

We implemented the methodology for scheduling risk-based biopsies in prostate cancer active surveillance in a web-application (`https://emcbiostatistics.shinyapps.io/prias\_biopsy\_recommender/`). In this web-application, four types of risk-based personalized schedules are provided, namely, using 5%, 10%, and 15% fixed thresholds, and a current-visit time specific optimal threshold obtained via (4). Also, a general purpose source code for creating personalized test schedules is provided in Web-Appendix D. The source code is compatible with joint models fitted using the R package **JMbayes** (Rizopoulos 2016).

5 Simulation Study

Although we evaluated personalized schedules for a demonstration patient, we also intend to analyze and compare personalized and fixed schedules in a full cohort. Our criteria for comparison of schedules are the total number of invasive tests planned (burden), and the actual time delay in detecting progression (shorter is beneficial) for each schedule. Due to the periodical nature of schedules, the actual time delay in detecting progression cannot be observed in real-world surveillance. Hence, instead, we compare personalized versus fixed schedules via an extensive simulated randomized clinical trial in which each hypothetical patient undergoes each schedule. To keep our simulation study realistic, we employ the prostate cancer active surveillance scenario. Specifically, our simulated population is generated using the joint model fitted to the PRIAS cohort (Tomer et al. 2019).

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 (Tomer et al. 2019), 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 9 months (0.75 years), denoted $\kappa^*\{v \mid E(\mathcal{D}) \leq 0.75\}$. 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 in 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, but 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. 2017*a*), 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* utilized 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 certain threshold. We select the risk threshold automatically in a personalized manner, by optimizing a utility function of the patient-specific consequences of choosing a particular risk threshold based schedule. These consequences are, namely, the number of invasive tests (burden) planned in a schedule, and the expected time delay in detection of progression (shorter is beneficial) if the patient progresses. Last, we calculate this expected time delay in a personalized manner for both personalized and fixed schedules to assist patients/doctors in making a more informed decision of choosing a test schedule.

Using joint models gives us certain advantages. First, since joint models employ random-effects, the corresponding risk-based schedules are inherently personalized. Second, to predict this patient-specific risk of progression, joint models utilize all observed longitudinal measurements of a patient. Also, the continuous longitudinal outcomes are not discretized,
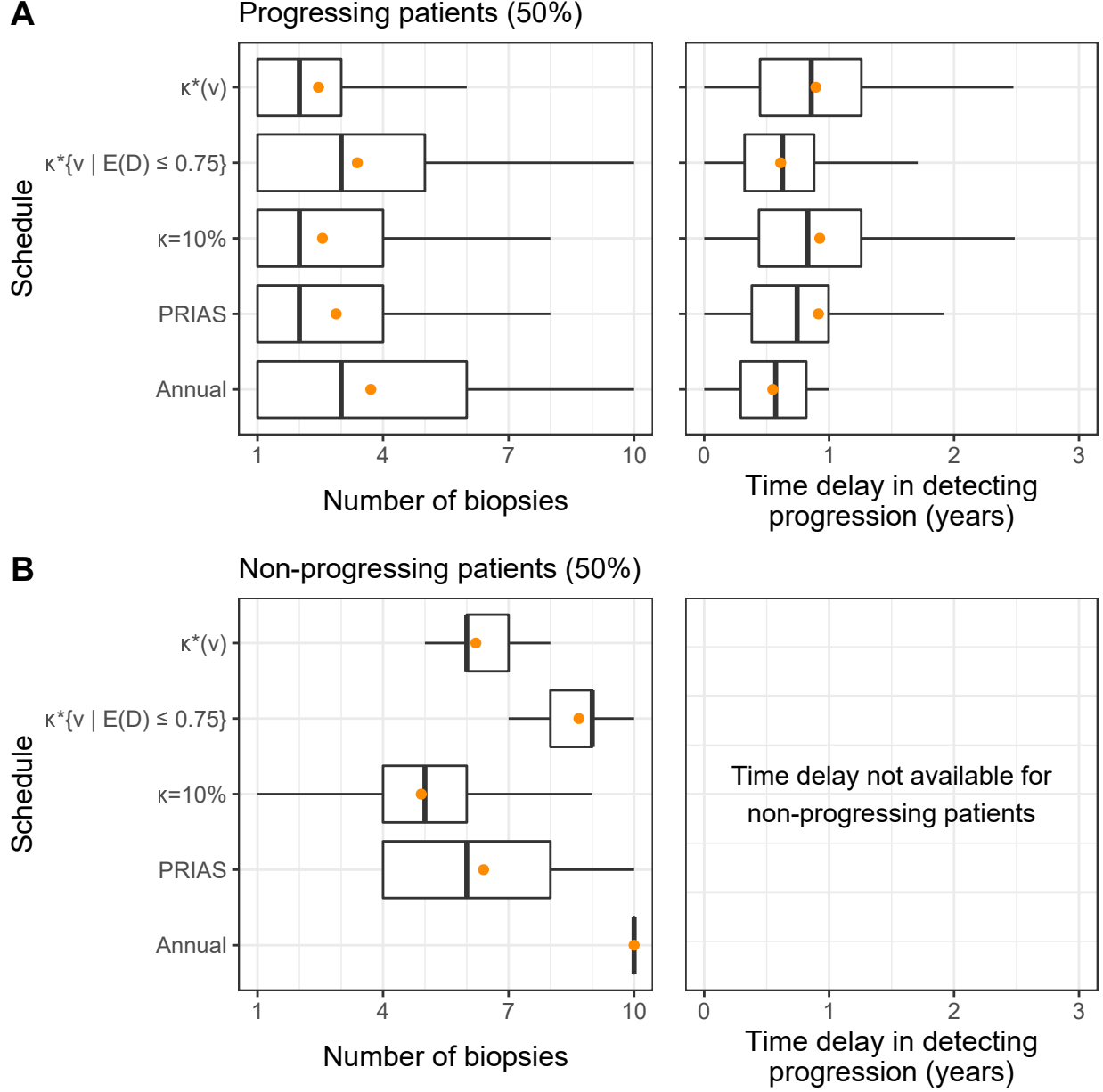
Figure 6: **Number of biopsies and the time delay in detecting cancer progression for various biopsy schedules** obtained via a simulation study. **Mean** is indicated by the orange circle. Time delay (years) is calculated as (time of positive biopsy - the actual simulated time of cancer progression). 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

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 should manifest the window of opportunity for curative treatment and additional benefits of observing progression early. Practitioners may extend/modify this 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 more 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

26

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.

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. Upgrading is susceptible to inter-observer variation and sampling error. Although models that account for these two issues (Balasubramanian & Lagakos 2003, Coley et al. 2017) will provide better risk estimates, the methodology for obtained personalized schedules can remain the same.

Acknowledgments

Data Availability

This work utilized results from a statistical model we previously fitted to the PRIAS dataset (Tomer et al. 2019). The PRIAS database is not openly accessible. However, access to the database can be requested on the basis of a study proposal approved by the PRIAS steering committee. The website of the PRIAS program is `www.prias-project.org`. For sake of completeness and reproducibility of results, we have presented the PRIAS based model's definition and parameter estimates in Web-Appendix B. The data of the demonstration patient in Figure 5 is provided in Web-Appendix B.

Supporting Information

Web Appendices referenced in this paper are available in the file titled 'supplementary.pdf'.

References

Alagoz, O., Ayer, T. & Erenay, F. S. (2010), 'Operations research models for cancer screening', *Wiley encyclopedia of operations research and management science* .

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

Bebu, I. & 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. & 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. & 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. & Wong, W. K. (1994), 'On the equivalence of constrained and compound optimal designs', *Journal of the American Statistical Association* **89**(426), 687–692.

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

de Carvalho, T. M., Heijnsdijk, E. A. & de Koning, H. J. (2017*a*), '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. & de Koning, H. J. (2017*b*), 'When should active surveillance for prostate cancer stop if no progression is detected?', *The Prostate* **77**(9), 962–969.

Eilers, P. H. & 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. & Humphrey, P. A. (2016), 'The 2014 international society of urological pathology (ISUP) consensus con-

ference 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. & 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. & 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. & Etzioni, R. (2014), 'Overdiagnosis and overtreatment of prostate cancer', *European urology* **65**(6), 1046–1055.

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

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

McCulloch, C. E. & Neuhaus, J. M. (2005), 'Generalized linear mixed models', *Encyclopedia of biostatistics* **4**.

McWilliams, T. J., Williams, T. J., Whitford, H. M. & 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., Taylor, J. M., Van Rosmalen, J., Steyerberg, E. W. & 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. & Denton, B. T. (2017), Markov decision processes for screening and treatment of chronic diseases, *in* 'Markov Decision Processes in Practice', Springer, pp. 189–222.

Taylor, J. M., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T. & Sandler, H. (2013), 'Real-time individual predictions of prostate cancer recurrence using joint models', *Biometrics* **69**(1), 206–213.

Tomer, A., Nieboer, D., Roobol, M. J., Bjartell, A., Steyerberg, E. W., Rizopoulos, D. & et al. (2020), 'A ready to use web-application providing a personalized biopsy schedule for men with low-risk PCa under active surveillance', - **manuscript submitted to European Urology**(-), –.

Tomer, A., Rizopoulos, D., Nieboer, D., Drost, F.-J., Roobol, M. J. & Steyerberg, E. W. (2019), '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. & Jemal, A. (2015), 'Global cancer statistics, 2012', *CA: A Cancer Journal for Clinicians* **65**(2), 87–108.

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

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.