Biometrics 00, 1–?? DOI: 10.1111/j.1541-0420.2005.00454.x
December 0000

Personalized Schedules for Burdensome Surveillance Tests

Anirudh Tomer[1,*], Daan Nieboer[2,3], 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

[*]*email:* a.tomer@erasmusmc.nl

Summary: 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 progression-risk based personalized test schedules that aim to balance better the number of tests (burden) and time delay in detecting progression (shorter is beneficial) 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 auxiliary 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 progression-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 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.

Key words: 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* for timely detecting disease *progression*. 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 invasive 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 periodical 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.

[Figure 1 about here.]

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 patient-specific clinical data accumulated over surveillance follow-up. This data includes baseline characteristics; previous invasive test results; and longitudinal 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 and 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 and 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 use 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 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. Specifically, we optimize a utility function of the expected number of tests (burden) and time delay in detecting progression (shorter is beneficial) for different risk threshold based personalized schedules. 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 objectively compare various personalized and fixed schedules.

1.1 *Motivational Study*

Our motivation comes from the problem of scheduling biopsies in the world's largest prostate cancer surveillance study, prostate cancer research international active surveillance (Bokhorst et al., 2015) or PRIAS. It has 7813 patients (1134 cancer 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 et al., 2014). Active surveillance aims to delay serious treatments (e.g., surgery, radiotherapy) until progression is observed. Hence, 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 $\geqslant 2$ is the reference test for confirming progression. Most often biopsies are scheduled annually (Loeb et al., 2014). However, such a frequent schedule can put unnecessary burden on patients with slow/non-progressing cancers and consequent 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.

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 $l_i < T_i^* \leqslant r_i$ (Figure 1), with r_i and l_i denoting 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 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, 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:

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

where $h_0(\cdot)$ denotes the baseline hazard, $\mathcal{M}_{ki}(t) = \{m_{ki}(s) \mid 0 \leqslant s < t\}$ 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 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 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 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 on 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., $v \geqslant t$ (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:

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

The cumulative-risk function $R_j(\cdot)$, illustrated in Figure 2, is dynamic in the sense that it automatically updates over time as more longitudinal data becomes available.

[Figure 2 about here.]

3.2 *Personalized 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 $U = \{u_1, \dots, u_L\}$ 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 last time u_L can be chosen based on the available information in the original 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. In particular, the test decision at time u_l denoted as $Q_j^\kappa(u_l \mid \cdot)$ is given by (e.g., caption of Figure 3),

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

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, on which all future time points $u_l \in U$ 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 $t_{l+1} = u_l$ to account for the possibility that progression may occur after time $u_l < T_j^*$. Hence, the time of last test t_l is defined as,

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

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

[Figure 3 about here.]

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

To facilitate shared-decision making, we translate our proposed decision rule, i.e., the choice of a specific κ, into two clinically relevant quantities. First, the number of tests (burden) we expect to perform for patient j if threshold κ is used. 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 never progressed in the period $[t, u_L]$. Under this assumption, the subset of future time points in U at which a test is to be conducted results into a personalized schedule of planned future tests (e.g., Figure 3 with κ=12%), given by:

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

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 denoting the number of performed tests in conjunction with the true progression time T_j^* as:

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

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

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

where

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

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:

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

The expected delay will be the expected value of $\mathcal{D}_j(S_j^\kappa)$ given by the expression:

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

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

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

The personalized schedule in (3), and the corresponding personalized expected number of tests and time delay, all have the advantage of getting updated with newly collected data over follow-up. Also, 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. Also, through the timing and 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, and in accordance to our developments in the previous section, we translate different choices for this parameter into the expected number of test and time delay. More specifically, for a specific patient j and at the current visit time v, we can construct the bi-dimensional Euclidean space of the expected total number of tests (x-axis) and time delay in detecting progression (y-axis) for test schedules planned by varying κ in $[0, 1]$. An example of such a space is given in Figure 4.

[Figure 4 about here.]

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 a current visit time specific risk threshold $\kappa^*(v)$ can be chosen as the threshold that minimizes 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.,

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

Additional clinical consequences of following a particular schedule, such as (quality-adjusted) life-years saved, can also be accommodated in (4). This requires first setting a point of optimality in a higher dimensional Euclidean space of such consequences, and then minimizing the Euclidean distance relative to this point of optimality.

An alternative approach is to constrain one of the two dimensions. For example, patients/doctors may not agree to more than a maximum number of planned future tests. They may also be apprehensive about having an expected time delay higher than a certain number of months. In such situations, the Euclidean distance in (4) can be minimized under constraints on the expected number of tests and/or expected time delay (Figure 4). An additional benefit of this approach is that it alleviates the issue of time delay and number of tests having different units of measurement Cook and Wong (1994).

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 1.1. 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 **JMbayes** (Rizopoulos, 2016). While the full 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 cumulative-risk of progression at the maximum study period of ten years was 50% (Web-Figure 1). This indicates that many patients may not require all the annual biopsies they are currently 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 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 PRIAS patient (Figure 5), starting from his current visit at year five, until year ten of follow-up. This patient has not progressed until year 3.5, and hence even if he incurs a delay in detecting progression of up to three years, it may not lead to adverse outcomes (de Carvalho et al., 2017). Also, since his cumulative-risk of progression at year ten is only 18.8%, he is likely to progress slowly. Consequently, risk-based fewer biopsies are planned in risk-based personalized schedules than the widely used annual schedule (Panel B, Figure 5). In addition, in both personalized schedule based on a fixed risk threshold of 10% and automatically chosen risk threshold $\kappa^*(v)$, the expected delay in detecting progression is much less the aforementioned limit of three years (Panel D, Figure 5).

The current PRIAS protocol for biopsies is fixed biopsies at year one, four, seven, and ten of follow-up, and every five years after that. Additional annual biopsies are scheduled if a patient's PSA doubling-time (Bokhorst et al., 2015) is high. Our event of interest is cancer progression.

[Figure 5 about here.]

5. Simulation Study

Although we demonstrated personalized schedules for a real 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. However, 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. More specifically, our simulated population is manifested by the joint model fitted to the PRIAS cohort (Web-Appendix B.3).

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 a set of longitudinal DRE and PSA measurements at the same follow-up visit times as given in the PRIAS protocol. We then split each dataset into training (750 patients) and test (250 patients) parts, and generate a random and noninformative censoring time for the training patients. All test and training 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 (Web-Appendix B.3), 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 three personalized biopsy schedules. First using a fixed risk threshold of $\kappa = 10\%$. Second, an automatically chosen visit-specific threshold $\kappa^*(v)$. Third, an automatic threshold under the constraint that expected delay is less than 9 months (0.75 years) $\kappa^*\{v \mid E(D) \leqslant 0.75\}$. We also conduct biopsies according to the currently practiced PRIAS and annual schedules. Successive personalized biopsy decisions 2 are made only on the standard PSA follow-up visits, utilizing clinical data accumulated only until the corresponding current visit time. 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 *Results*

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 the total number of biopsies scheduled in all 500×250 test patients, but the time delay in detecting 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 in Panel A, and Panel B of Figure 6, respectively.

For *progressing* patients (Panel A, Figure 6), the annual schedule leads to the maximum number of biopsies (Median 3, Inter-quartile range or IQR: 1–6). However, it also guarantees a maximum time delay of one year for all patients. The PRIAS protocol schedules much fewer biopsies (Median 1, IQR: 2–4), but also has a higher time delay (Median 0.74, IQR: 0.38–1.00 years). The personalized schedule based on automatically chosen 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 mean delay for schedule $\kappa^*\{v \mid E(D) \leqslant 0.75\}$ is equal to 0.61 years (Median 0.63, IQR: 0.32–0.88 years) and hence works as expected. Unless the patient progresses within the first year of prostate cancer active surveillance, a delay of up to three years may not increase the risk of adverse downstream outcomes in (Inoue et al., 2018; de Carvalho et al., 2017).

The patients who are at the most advantage with the personalized schedules are the *non-progressing* patients (Panel B, Figure 6). For all of these patients, the annual schedule leads to 10 (unnecessary) biopsies. The schedule of the PRIAS program schedules a median of 6 (IQR: 4–8) biopsies. In comparison, the schedule based on automatically chosen risk threshold $\kappa^*(v)$ schedules a median of 6 (IQR: 6–7) biopsies, and schedule based on a fixed risk threshold of 10% schedules only median of 5 (IQR: 4–6) biopsies.

[Figure 6 about here.]

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 the framework of joint models for time-to-event and longitudinal data. Our approach first combines a patient's auxiliary longitudinal data (e.g., biomarkers) and results from previous invasive tests to estimate the patient-specific cumulative-risk of disease progression over his current and future follow-up time period. Then, using this risk profile, we schedule future invasive tests whenever the patient's cumulative-risk of progression is predicted to be above a certain threshold. We select this 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) for a particular schedule, and the expected time delay in detection of progression (shorter is beneficial) if that schedule is followed. 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.

The use of joint models gives our schedules certain advantages. First, joint models utilize individualized random-effects, making our schedules inherently personalized. Second, the patient-specific risk of progression employed by the proposed personalized schedules is estimated by utilizing all observed longitudinal and clinical data of a patient. In addition, the continuous longitudinal outcomes are not discretized, which is commonly a case in Markov Decision Process based (Alagoz et al., 2010; Steimle and Denton, 2017), and flowchart-based test schedules. Third, our schedules update automatically with more patient data over follow-up. Last, although this work concerns with the use of personalized schedules in disease surveillance, the methodology is generic for use under a screening setting as well.

Since our schedules are risk-based, we proposed a utility function to automate the choice of a risk threshold based schedule. The utility function that we proposed focused only on two aspects of a schedule, namely the burden and the benefit. In this regard, we chose the expected number of invasive tests in a schedule (burden) and time delay in detection of progression (less is beneficial) because they are easy to interpret and are critical in making the decision of an invasive test. We chose these two criteria because they also manifest financial and medical burden of tests, window of opportunity for curative treatment, and additional benefits of detecting disease early. Since we calculated both expected number of tests and time delay in a patient-specific manner for both personalized and fixed schedules, patients/doctors can compare and choose various risk-based and fixed schedules according to their preferences for the expected burden-benefit ratio. Additional measures such as (quality-adjusted) life-years saved can also be easily added in our utility function.

We evaluated the efficacy of personalized schedules in a full cohort via a realistic simulation randomized clinical trial for prostate cancer active surveillance patients. We observed that the personalized schedule that used an automatically chosen risk threshold using (4) reduced unnecessary biopsies for patients who did not observe progression in the study period, compared to annual schedule. In contrast, in patients who observed progression, the personalized schedule with automatically chosen risk threshold scheduled fewer biopsies at the cost of having a slightly more time delay in detecting progression than the fixed schedules. However, this by no means is the limit of the performance of the personalized schedules. In general, personalized schedules employing 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 of our work. First, in practice, most cohorts observe Type-I right censoring. Hence, the cumulative-risk profiles of patients and the calculation of expected time delay in detection of progression is only possible up to the time of Type-I censoring. This problem can only be resolved as more follow-up data become available over time. We proposed a joint model which assumes 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 too. Models which account for this variation (Coley et al., 2017; Balasubramanian and Lagakos, 2003) will be interesting to investigate further.

Acknowledgments

The first and last authors would like to acknowledge support by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (the national research council of the Netherlands) VIDI grant nr. 016.146.301, and Erasmus University Medical Center funding. Part of this work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. The authors also thank the Erasmus University Medical Center's Cancer Computational Biology Center for giving access to their IT-infrastructure and software that was used for the computations and data analysis in this study. Last, we would like to thank the PRIAS consortium for enabling this research project.

Supporting Information

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

References

Alagoz, O., Ayer, T., and Erenay, F. S. (2010). Operations research models for cancer screening. *Wiley encyclopedia of operations research and management science* .

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

Bebu, I. and Lachin, J. M. (2017). Optimal screening schedules for disease progression with application to diabetic retinopathy. *Biostatistics* **19,** 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,** 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,** 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,** 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,** 687–692.

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

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

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11,** 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,** 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,** 1–9.

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,** 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,** 325–333.e2.

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

Loeb, S., Carter, H. B., Schwartz, M., Fagerlin, A., Braithwaite, R. S., and Lepor, H. (2014). Heterogeneity in active surveillance protocols worldwide. *Reviews in urology* **16,** 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,** 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,** 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,** 1–46.

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,** 149–164.

Steimle, L. N. and Denton, B. T. (2017). Markov decision processes for screening and treatment of chronic diseases. In *Markov Decision Processes in Practice*, pages 189–222. Springer.

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,** 206–213.

Tomer, A., Nieboer, D., Roobol, M. J., Bjartell, A., Steyerberg, E. W., Rizopoulos, D., and 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,** –.

Tomer, A., Rizopoulos, D., Nieboer, D., Drost, F.-J., Roobol, M. J., and Steyerberg, E. W. (2019). Personalized decision making for biopsies in prostate cancer active surveillance programs. *Medical Decision Making* **39,** 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,** 87–108.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14,** 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 barretts esophagus: European society of gastrointestinal endoscopy (esge) position statement. *Endoscopy* **49,** 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.

Received October 0000. *Revised February* 0000. *Accepted March* 0000.

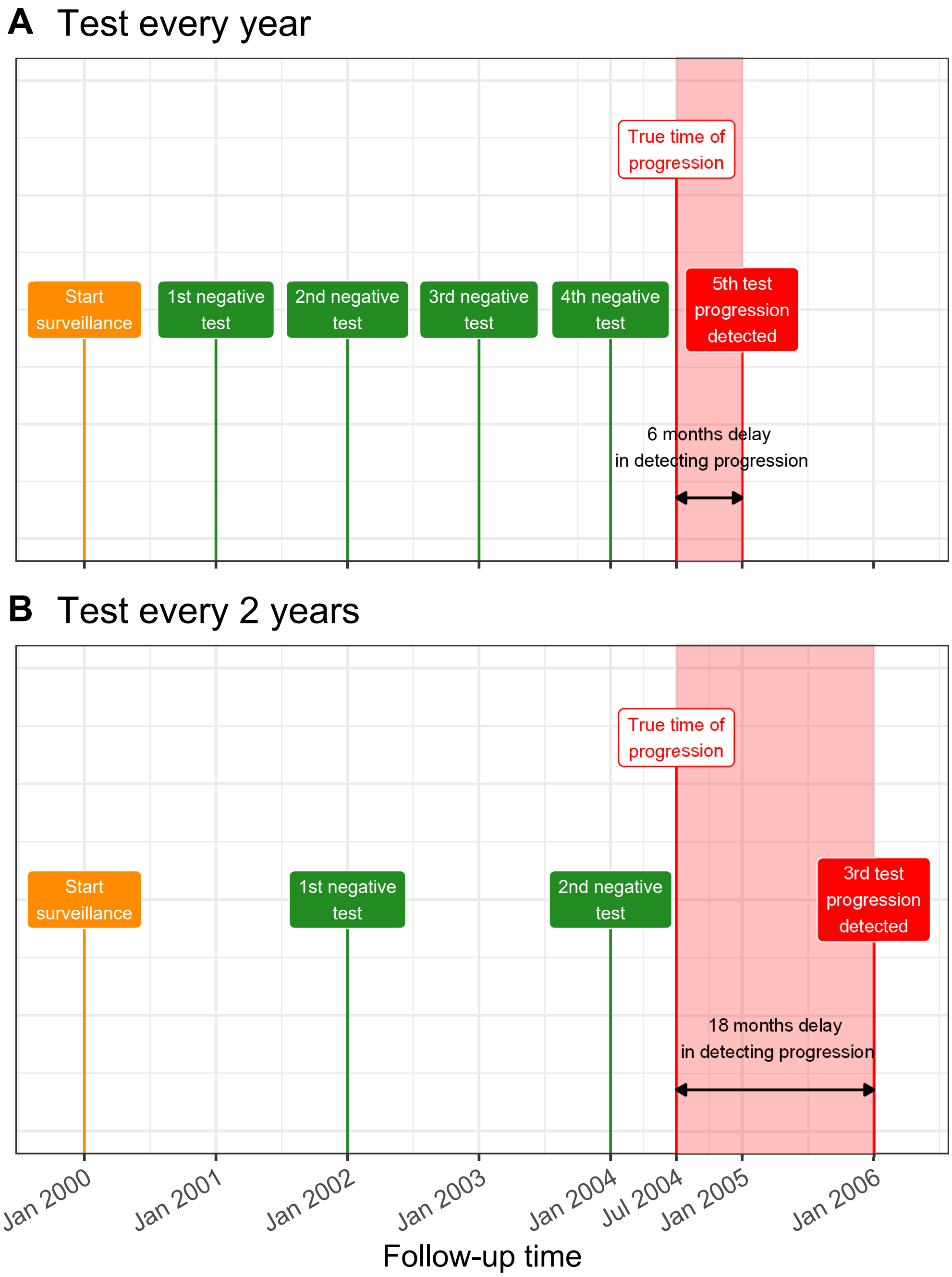

Figure 1. Goal: Finding the optimal tradeoff between the number of invasive tests (burden) and time delay in detecting progression (shorter is beneficial). Progression is a non-terminal event of interest. The true time of progression for this patient July 2004. Since invasive tests are conducted periodically, progression is interval-censored and always observed with a delay. Specifically, more frequent invasive tests in **Panel A** lead to a shorter time delay in detecting progression than less frequent 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**.

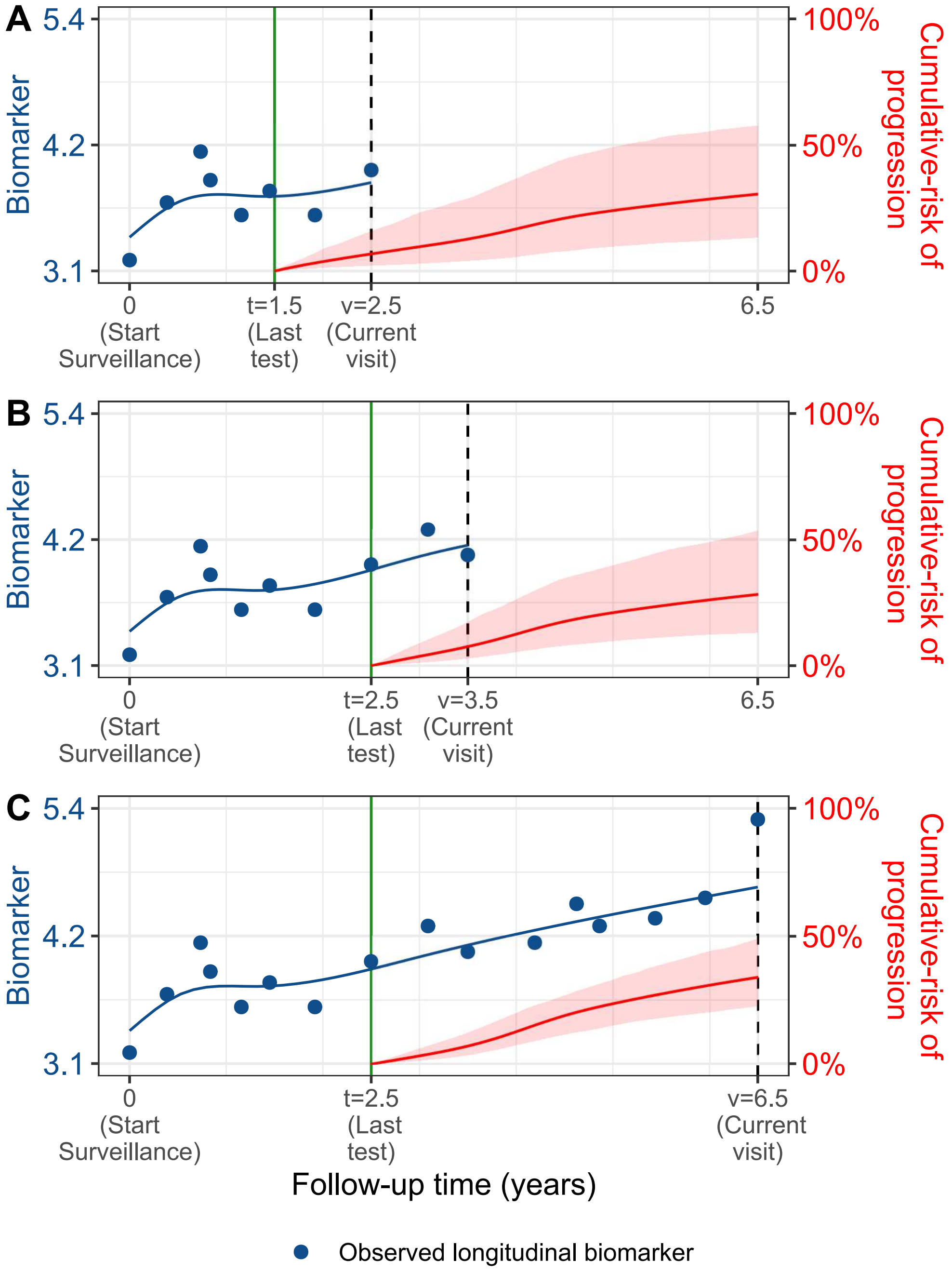

Figure 2. Cumulative-risk of progression updated dynamically over follow-up as more patient data is gathered. A single longitudinal outcome, namely, a continuous biomarker of disease progression, is used for illustration. **Panels A, B and C:** are ordered by the time of the current visit v (dashed vertical 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.

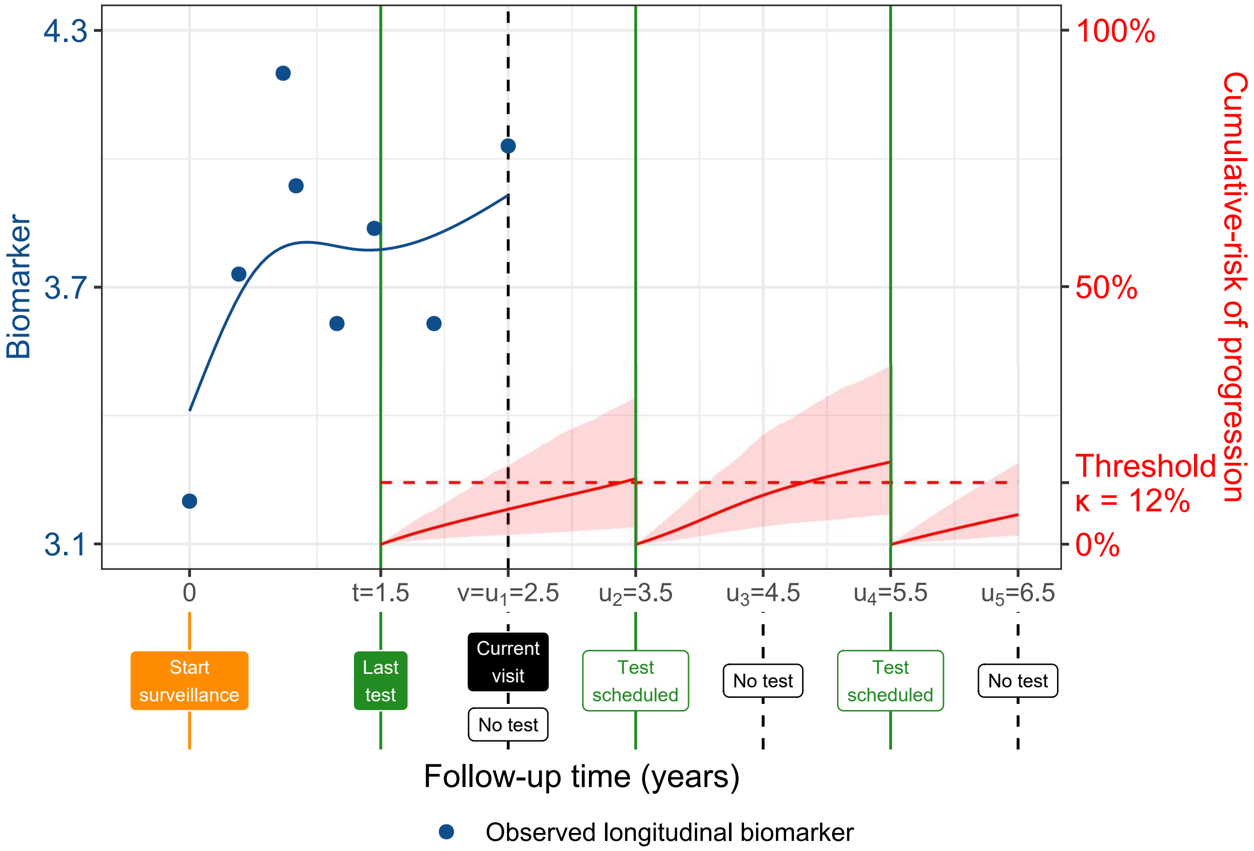

Figure 3. Illustration of Personalized Decision Rule (2) Based on Patient-specific Cumulative-risk of Progression. Time of current visit: $v = 2.5$ years (black dashed line). Time of the last test on which progression was not observed: $t = 1.5$ years. Longitudinal data up to current visit: $\mathcal{Y}_j(v)$ 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 (confidence 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: $Q^\kappa_j(u_1 = 2.5 \mid t_1 = 1.5, v) = 0$, $Q^\kappa_j(u_2 = 3.5 \mid t_2 = 1.5, v) = 1$, $Q^\kappa_j(u_3 = 4.5 \mid t_2 = 3.5, v) = 0$, $Q^\kappa_j(u_4 = 5.5 \mid t_2 = 3.5, v) = 1$, and $Q^\kappa_j(u_5 = 6.5 \mid t_5 = 4.5, v) = 0$. All values are illustrative.

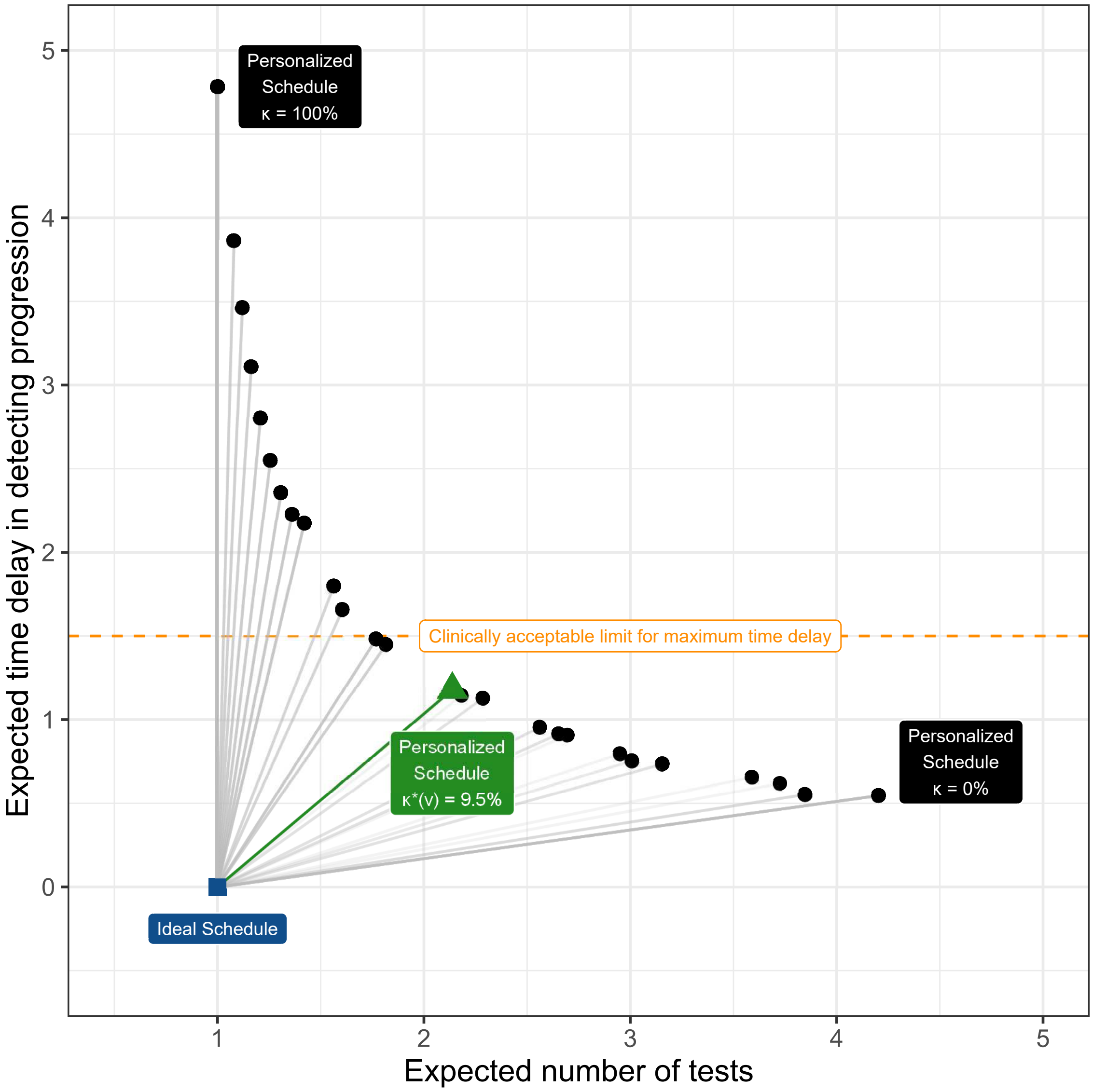

Figure 4. Automatic choice of risk threshold $0 \leqslant \kappa \leqslant 1$ using (4). The ideal schedule of tests at point (1,0) is shown as a blue square. It plans exactly one invasive test at the true time of progression T_j^* of a patient and hence leads to a zero time delay in detecting progression. Personalized schedules based on a grid of thresholds chosen between $0 \leqslant \kappa \leqslant 1$ are shown with black circles. Higher thresholds lead to fewer tests, but also higher expected time delay. We propose to choose the personalized schedule based on $\kappa^*(v) = 9.5\%$ threshold (green triangle). This is because it has the least Euclidean distance (shown with a green line) to the ideal schedule. It is also possible to find the least distance under a certain clinically acceptable limit on time delay (orange dashed line), or number of tests.

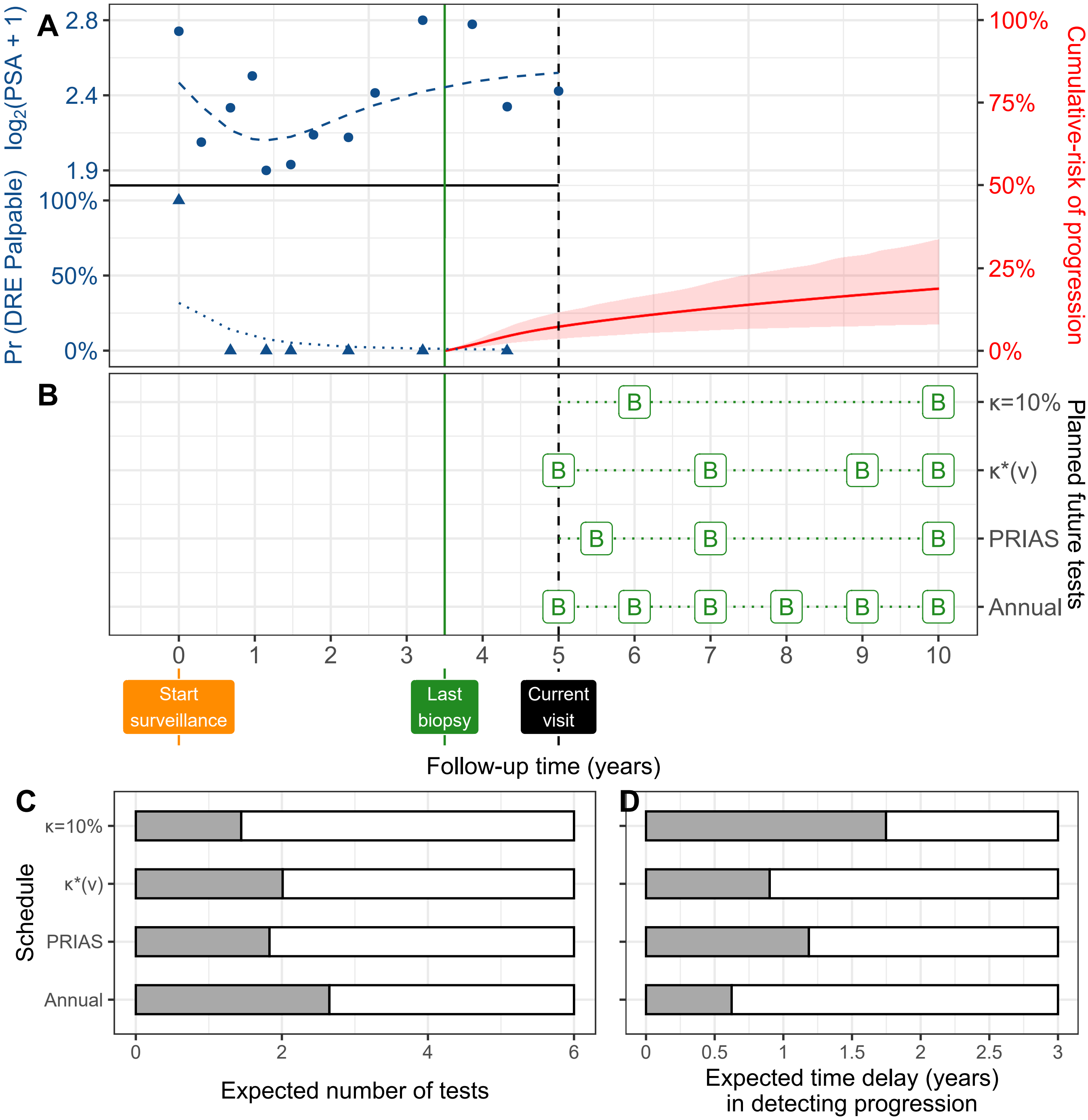

Figure 5. Illustration of personalized schedules for a demonstration PRIAS patient: In **Panel A**: Time of last negative biopsy is year 3.5 (vertical green solid line). Longitudinal data: DRE (blue triangles) and PSA (blue circles). The current visit is year five (vertical black dashed line). The estimated cumulative-risk profile is shown with a solid red line (95% credible interval is shaded). It is 18.8% at year ten (horizon). In **Panel B**, we visualize different biopsy schedules, with a 'B' indicating a biopsy. $\kappa = 10\%$ and $\kappa^*(v)$ are personalized biopsy schedules using a fixed risk threshold of 10%, and automatically chosen threshold (4), respectively. PRIAS and Annual denote the PRIAS biopsy schedule (paragraph 2 of Section 4) 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., 2017).

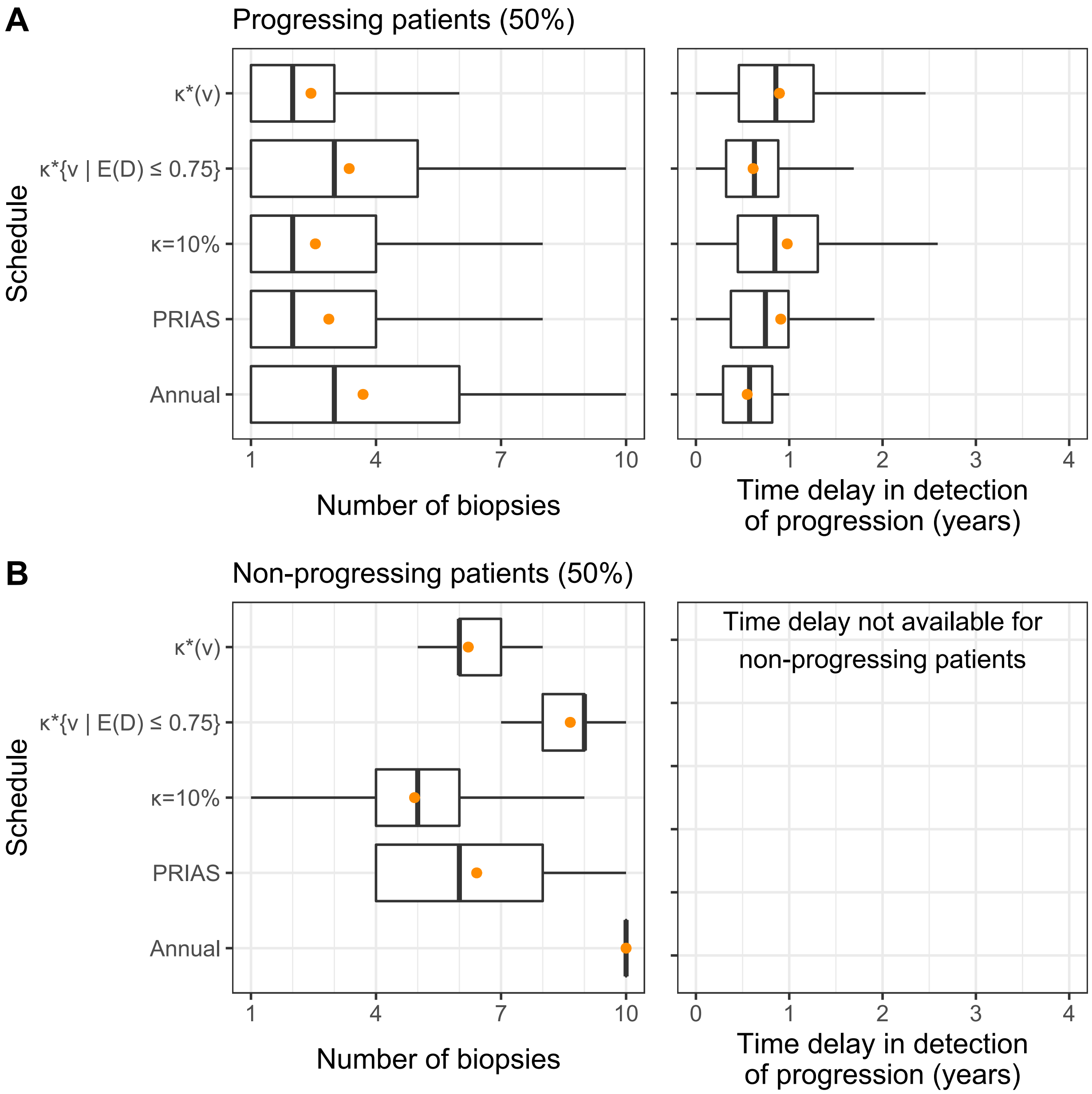

Figure 6. Boxplot showing variation in the number of biopsies, and the time delay in the detection of cancer progression for various biopsy schedules. **Mean** is indicated by the orange circle. Time delay (years) is calculated as (time of positive biopsy - the true time of cancer progression). Biopsies are conducted until cancer progression is detected. **Panel A:** results for simulated patients who obtained cancer progression in the ten year study period (*progressing*). **Panel B:** results for simulated patients who did not obtain cancer progression in the ten year study period (*non-progressing*). Types of schedules: $\kappa = 10\%$ and $\kappa^*(v)$ schedules a biopsy if the cumulative-risk of cancer progression at a visit is more than 10%, and an automatically chosen threshold (4), respectively. Schedule $\kappa^*\{v \mid E(D) \leqslant 0.75\}$ 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).