PRIAS Personalized Biopsy Schedules

Firstname1 Lastname1

Abstract

Low risk prostate cancer patients are often encouraged to join active surveillance (AS) programs rather than taking immediate treatment. In most AS programs repeat biopsies are conducted annually or as per a common fixed schedule for these patients. When the Gleason score based on biopsies is found to be upgraded, the patients are given curative treatment. It has been found that such fixed and frequent schedules for biopsies discourage patients to receive repeat biopsies, and also bring a financial burden. Motivated by the world's largest AS program PRIAS, in this work we present personalized schedules for biopsies to counter these problems. Our methods create separate schedules for every person, on the basis of the evolution of his prostate-specific antigen (PSA) levels as well as results from previous repeat biopsies. We discuss criteria for evaluation of biopsy schedules, and then use them to compare the efficacy of personalized schedules with that of existing biopsy schedules.

1 Introduction

In this decade prostate cancer is the second most frequently diagnosed cancer (14% of all cancers) in males worldwide, and the most frequent (19% of all cancers in USA alone) in economically developed countries (Siegel, Miller, and Jemal, 2017; Torre et al., 2015). With increase in life expectancy and increase in number of screening programs, an increase in diagnosis of low grade prostate cancers has been observed (Potosky et al., 1995). A major issue of screening programs that also has been established in other types of cancers (e.g. breast cancer) is over-diagnosis. To avoid overtreatment, patients diagnosed with low grade prostate cancer are often motivated to join active surveillance (AS) programs. The goal of AS is to routinely check the progression of prostate cancer and avoid serious treatments such as surgery or chemotherapy as long as they are not needed.

Currently the largest AS program worldwide is PRIAS (www.prias-project.org) (Bokhorst et al., 2015). Patients enrolled in PRIAS are closely monitored using serum prostate-specific antigen (PSA) levels, digital rectal examination (DRE) and repeat prostate biopsies. Biopsies are evaluated using the Gleason grading system. Gleason scores range between 2 and 10, where a score 10 corresponds to a very serious state of prostate cancer. Patients who join PRIAS have a Gleason score of 6 or less, DRE score of cT2c or less and a PSA of 10 ng/mL or less at the time of induction. Although a PSA doubling time, also called PSA-DT (measured as the inverse of the slope of regression line through the base 2 logarithm of PSA values) of less than 3 years, DRE of cT3 or more, and a Gleason score more than 6 are indicators of prostate cancer progression, only DRE and Gleason scores are considered to be conclusive in this regard (Bokhorst et al., 2016). If

either the DRE or the Gleason score are found to be higher than the aforementioned thresholds, then the patient is removed from AS for further curative treatment. When the Gleason score becomes greater than 6, it is also known as Gleason reclassification (referred to as GR hereafter).

Gleason scores are reliable, however the associated biopsies are difficult to conduct, cause pain and have serious side effects such as hematuria and sepsis for prostate cancer patients (Loeb et al., 2013). Due to these reasons PRIAS as well as the majority of the AS programs worldwide strongly adhere to the rule of not having more than 1 biopsy per year. Performing a biopsy every year (we refer to it as annual schedule hereafter) has the advantage that it is possible to detect GR within 1 year of its occurrence. The drawbacks of this schedule though, are not only medical but also financial. Keegan et al., 2012 have shown that if a biopsy is performed every year then the costs of AS per head, at 10 years of follow-up exceed the costs of treatment (brachytherapy or prostatectomy) at 6 and 8 years of follow-up, respectively. They also found that performing biopsy every other year led to 99% increase in savings (AS vs. primary treatment) per head over a period of 10 years compared to annual schedule. Despite this, several AS programs employ the annual schedule (Tosoian et al., 2011; Welty et al., 2015). For patients enrolled in PRIAS the schedule is comparatively less rigorous. In PRIAS schedule one biopsy is performed at the time of induction, and the rest at year 1, 4, 7, 10 and every 5 years thereafter. An exception is made, if at any time a patient has PSA-DT less than 10 years, wherein annual schedule for biopsy is prescribed.

PRIAS schedule is less rigorous than annual schedule, yet it has a high non-compliance rate for repeat biopsies. Bokhorst et al., 2015 reported that the percentage of men receiving repeat biopsies decreased from 81% at year 1 to 60% in year 4, 53% in year 7 and 33% in year 10 of follow up. Non-compliance of biopsy schedule reduces the effectiveness of AS programs, as progression is detected late. When compliance is high, patients whose cancer progress slowly often end up having biopsies when they are not needed. For a patient with faster progressing cancer, crude measures such as PSA-DT are employed to decide timing of biopsies. The fact that existing schedules require improvement is also evident in some of the reasons given by patients for non-compliance: 'patient does not want biopsy', 'PSA stable', 'complications on last biopsy'and 'no signs of disease progression on previous biopsy'.

This paper is motivated by the need to reduce the burden of biopsies and most optimally find the onset of Gleason reclassification. To this end, we intend to create schedules for biopsies which improve upon the existing PRIAS and annual schedule. For this purpose, one approach which stands out in particular in literature is personalized scheduling. That is, a different schedule for

every patient and/or scenario. For e.g. Bebu and Lachin, 2017 have proposed cost optimized personalized schedules based on Markov models. O'Mahony et al., 2015 have proposed cost optimized personalized equi-spaced screening intervals, using Microsimulation Screening Analysis (MISCAN) models. Parmigiani, 1998 have used information theory to come up with schedules for detecting time to event in the smallest possible time interval. Most of these methods however create an entire schedule in advance. In contrast Rizopoulos et al., 2016 have proposed dynamic personalized schedules for longitudinal biomarkers using the framework of joint models for time to event and longitudinal data(Rizopoulos, 2012; Tsiatis and Davidian, 2004).

The schedules we propose in this paper are tailored separately for every patient, and are dynamic. That is, biopsies are scheduled at different time points per patient utilizing his available PSA measurements and previous biopsy results up to that point in time. We achieve this using joint models. Based on these models we obtain a full specification of the joint distribution of the PSA levels and time of GR. We further use it to define a patient-specific posterior predictive distribution of the time of GR given the observed PSA measurements and previous biopsies. Using the general framework of Bayesian decision theory, we propose a set of loss functions which are minimized to find the optimal time of performing a biopsy. These loss functions yield us two categories of personalized schedules, those based on expected time of GR and those based on the risk of GR. We also analyze an approach where the two types of schedules are combined. To compare the proposed personalized schedules with the PRIAS and annual schedule we conduct a simulation study, and then discuss various metrics for evaluating efficacy of each schedule, and choosing the most suitable one. It is important to note that a schedule for measurement of DRE is not of interest since it is a non invasive procedure and has no serious medical implications. Thus the only event of interest is GR and not DRE crossing the threshold of cT2c.

As mentioned earlier, to achieve personalization the proposed schedules utilize information from PSA measurements. This is important because PSA measurements are easy to obtain, and thus they are consequently cost effective. PSA measurement process does not lead to any side effects and thus compliance rate for measurement of PSA levels is high(91% in PRIAS). Most importantly, it was found in PRIAS that PSA-DT was indicative of GR Bokhorst et al., 2015. The information from PSA was however not fully utilized. More specifically, when PSA was observed to be stable, patients/doctors in PRIAS not always complied with the biopsy schedule. In this regard, the usage of joint models allows more sophisticated modeling of PSA levels and information from repeat biopsies than PRIAS, and thus offers a more informative decision making process.

The rest of the paper is organized as follows. Section 2 covers briefly the joint modeling framework. Section 3 details the personalized scheduling approaches we have proposed in this paper. In Section 4 we demonstrate personalized schedules in a real world scenario by employing them for the patients from the PRIAS program. Lastly, in Section 5, we present the results from a simulation study we conducted, to compare personalized schedules with PRIAS schedule and annual schedule.

2   Joint model for time to event and longitudinal outcomes

We start with the definition of the joint modeling framework that will be used to fit a model to the available dataset, and then to plan biopsies for future patients.

2.1   Joint model specification

Let T_i^* denote the true GR time for the i-th patient enrolled in an AS program. Let the vector of times at which biopsies are conducted for this patient be denoted by $T_i^b = \{T_{i0}^b, T_{i1}^b, \ldots, T_{iN_i^b}^b; T_{ij}^b < T_{ik}^b, \forall j < k\}$, where N_i^b are the total number of biopsies conducted. T_i^* cannot be observed directly and it is only known that it falls in an interval $(l_i, r_i]$, where $l_i = T_{iN_i^b - 1}^b, r_i = T_{iN_i^b}^b$ if the GR is observed, and $l_i = T_{iN_i^b}^b, r_i = \infty$ if patient drops out of AS before GR is observed. Further let $\boldsymbol{y}_i$ denote the $n_i \times 1$ vector of PSA levels for the i-th patient. For a sample of n patients the observed data is denoted by $\mathcal{D}_n = \{l_i, r_i, \boldsymbol{y}_i; i = 1, \ldots, n\}$.

The longitudinal outcome of interest, namely PSA level is continuous in nature and thus to model it the joint model utilizes a linear mixed effects model (LMM) of the form:

$$y_i(t) = m_i(t) + \varepsilon_i(t)$$
$$= \boldsymbol{x}_i^T(t)\boldsymbol{\beta} + \boldsymbol{z}_i^T(t)\boldsymbol{b}_i + \varepsilon_i(t)$$

where $\boldsymbol{x}_i(t)$ denotes the row vector of design matrix for fixed effects and $\boldsymbol{z}_i(t)$ denotes the row vector for random effects. Correspondingly the fixed effects are denoted by $\boldsymbol{\beta}$ and random effects by $\boldsymbol{b}_i$. The random effects are assumed to be normally distributed with mean zero and $q \times q$ covariance matrix $\boldsymbol{D}$. $m_i(t)$ denotes the true and unobserved value of longitudinal outcome at time t, i.e. unlike $y_i(t)$ it is not contaminated with measurement error $\varepsilon_i(t)$. The error is assumed to be normally distributed with mean zero and variance σ^2, and is independent of the random effects $\boldsymbol{b}_i$.

To model the effect of longitudinal outcome on hazard of GR, joint models utilize a relative risk sub-model. The hazard of GR for patient i at any time point t, denoted by $h_i(t)$, depends on

a function of subject specific linear predictor $m_i(t)$ and/or the random effects:

$$h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) = \lim_{\Delta t \to 0} Pr\{t \leq T_i^* < t + \Delta t \mid T_i^* \geq t, \mathcal{M}_i(t), \boldsymbol{w}_i\}/\Delta t$$

$$= h_0(t) \exp\left[\boldsymbol{\gamma}^T \boldsymbol{w}_i + f\{M_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\}\right]$$

where $\mathcal{M}_i(t) = \{m_i(v), 0 \leq v \leq t\}$ denotes the history of the underlying longitudinal process up to time t. $\boldsymbol{w}_i$ is a vector of baseline covariates and $\boldsymbol{\gamma}$ are the corresponding parameters. The function $f(\cdot)$ parametrized by vector $\boldsymbol{\alpha}$ specifies the functional form (Brown, 2009; Rizopoulos, 2012, 2016; Rizopoulos et al., 2014; Taylor et al., 2013) of longitudinal outcome that is used in the linear predictor of the relative risk model. Some functional forms relevant to the problem at hand, and their interpretation are the following:

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

These formulations of $f(\cdot)$ postulate that the hazard of GR at time t may be associated with the underlying level of the biomarker $m_i(t)$, or with both the level and slope of the longitudinal profile $m_i'(t)$ at time t. Lastly, $h_0(t)$ is the baseline hazard at time t, and is modeled flexibly using P-splines. More specifically:

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

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

2.2 Parameter estimation

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

parameters is given by:

$$p(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D}_n) \propto \prod_{i=1}^{n} p(l_i, r_i, \boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

$$\propto \prod_{i=1}^{n} p(l_i, r_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{y_i} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

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

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

where $\boldsymbol{X}_i = \{\boldsymbol{x}_i(t_{i1})^T, \ldots, \boldsymbol{x}_i(t_{in_i})^T\}^T$ and $\boldsymbol{Z}_i = \{\boldsymbol{z}_i(t_{i1})^T, \ldots, \boldsymbol{z}_i(t_{in_i})^T\}^T$. The likelihood contribution of the time to GR outcome is given by:

$$p\{l_i, r_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}\} = \exp\left\{ - \int_0^{l_i} h_i(s \mid \mathcal{M}_i(s), \boldsymbol{w}_i)\,\mathrm{d}s \right\} - \exp\left\{ - \int_0^{r_i} h_i(s \mid \mathcal{M}_i(s), \boldsymbol{w}_i)\,\mathrm{d}s \right\} \quad (1)$$

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

We use independent normal priors with zero mean and variance 100 for the fixed effects $\boldsymbol{\beta}$ and inverse Gamma priors for σ^2. For the variance–covariance matrix $\boldsymbol{D}$ of the random effects we take inverse Wishart prior with an identity scale matrix and degrees of freedom equal to the number q of the random effects. For the relative risk model's parameters $\boldsymbol{\gamma}$ and the association parameters $\boldsymbol{\alpha}$, we use independent normal priors with zero mean and variance 100. For the penalized version of the B-spline approximation to the baseline hazard, we use the following prior for parameters γ_{h_0} (Lang and Brezger, 2004):

$$p(\gamma_{h_0} \mid \tau_h) \propto \tau_h^{\rho(\boldsymbol{K})/2} \exp\left\{ - \frac{\tau_h}{2} \gamma_{h_0}^T \boldsymbol{K} \gamma_{h_0} \right\}$$

where τ_h is the smoothing parameter that takes a Gamma(1, 0.005) hyper-prior in order to ensure a proper posterior for γ_{h_0}, $\boldsymbol{K} = \Delta_r^T \Delta_r + 10^{-6}\boldsymbol{I}$, where Δ_r denotes the r-th difference penalty matrix, and $\rho(\boldsymbol{K})$ denotes the rank of $\boldsymbol{K}$.

3 Personalized schedules for repeat biopsies

Once a joint model for GR and PSA levels is obtained, the next step is to use it to create personalized schedules for biopsies. To elucidate the scheduling methods, let us assume that a personalized schedule is to be created for a new patient enumerated j, who is not present in the original sample

of patients $\mathcal{D}_n$. Further let us assume that this patient did not have a GR at his last biopsy performed at time t, and that the PSA levels are available up to a time point s. The goal is to find the optimal time $u \geq \max(t, s)$ of the next biopsy.

3.1 Posterior predictive distribution for time to GR

Let $\mathcal{Y}_j(s)$ denote the history of PSA levels taken up to time s for patient j. The information from PSA history and repeat biopsies is manifested by the posterior predictive distribution $g(T_j^*)$, given by (conditioning on baseline covariates $\boldsymbol{w}_i$ is dropped for notational simplicity hereafter):

$$
\begin{aligned}
g(T_j^*) &= p\big(T_j^* \mid T_j^* > t, \mathcal{Y}_j(s), \mathcal{D}_n\big) \\
&= \int p\big(T_j^* \mid T_j^* > t, \mathcal{Y}_j(s), \boldsymbol{\theta}\big) p\big(\boldsymbol{\theta} \mid \mathcal{D}_n\big) \, \mathrm{d}\boldsymbol{\theta} \\
&= \int \int p\big(T_j^* \mid T_j^* > t, \boldsymbol{b_j}, \boldsymbol{\theta}\big) p\big(\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(s), \boldsymbol{\theta}\big) p\big(\boldsymbol{\theta} \mid \mathcal{D}_n\big) \, \mathrm{d}\boldsymbol{b}_j \, \mathrm{d}\boldsymbol{\theta}
\end{aligned}
\tag{2}
$$

The posterior predictive distribution depends on the observed longitudinal history via the random effects $\boldsymbol{b_j}$. The posterior distribution of the parameters $\boldsymbol{\theta}$, denoted by $p(\boldsymbol{\theta} \mid \mathcal{D}_n)$ is obtained as in Section 2.2.

3.2 Loss functions

To find the time u of next biopsy, we use principles from statistical decision theory in a Bayesian setting (Berger, 1985; Robert, 2007). More specifically, we propose to choose future biopsy time u by minimizing the posterior expected loss $E_g\big[L\{T_j^*, u\}\big]$, where the expectation is taken w.r.t. the posterior predictive distribution $g(T_j^*)$.

$$
E_g\big[L\{T_j^*, u\}\big] = \int_t^\infty L\{T_j^*, u\} p\big(T_j^* \mid T_j^* > t, \mathcal{Y}_j(s), \mathcal{D}_n\big) \, \mathrm{d}T_j^*
$$

Various loss functions $L\{T_j^*, u\}$ have been proposed in literature (Robert, 2007). The ones we utilize, and the corresponding motivations are presented next.

3.2.1 Expected and median time of GR

One of the reasons, patients did not comply with the existing PRIAS schedule was 'complications on a previous biopsy '. Therefore, it makes sense to have as less biopsies as possible. In the ideal case only 1 biopsy, performed at the exact time of GR is sufficient. Hence, neither a time which overshoots the true GR time T_j^*, nor a time which undershoots is preferred. In this regard, the squared loss function $L\{T_j^*, u\} = (T_j^* - u)^2$ and absolute loss function $L\{T_j^*, u\} = |T_j^* - u|$ have the properties that the posterior expected loss is symmetric on both sides of T_j^*. Secondly, both

loss functions have well known solutions available. The posterior expected loss for the squared loss function is given by:

$$E_g\big[L\{T_j^*, u\}\big] = E_g\big[\{T_j^* - u\}^2\big]$$
$$= E_g\big[\{T_j^*\}^2\big] + u^2 - 2uE_g[T_j^*] \tag{3}$$

The posterior expected loss in (3) attains its minimum at $u = E_g[T_j^*]$, also known as expected time of GR. The posterior expected loss for the absolute loss function is given by:

$$E_g\big[L\{T_j^*, u\}\big] = E_g\big[|T_j^* - u|\big]$$
$$= \int_u^\infty (T_j^* - u)g(T_j^*)\,\mathrm{d}T_j^* + \int_t^u (u - T_j^*)g(T_j^*)\,\mathrm{d}T_j^* \tag{4}$$

The posterior expected loss in (4) attains its minimum at the median of $g(T_j^*)$, given by $u = \pi_j^{-1}(0.5 \mid t, s)$, where $\pi_j^{-1}(\cdot)$ is the inverse of dynamic survival probability $\pi_j(u \mid t, s)$ of patient j (Rizopoulos, 2011). It is given by:

$$\pi_j(u \mid t, s) = \Pr\big\{T_j^* \geq u \mid T_j^* > t, \mathcal{Y}_j(s), D_n\big\}, u \geq t \tag{5}$$

For ease of readability we denote $\pi_j^{-1}(0.5 \mid t, s)$ as median$[T_j^*]$ hereafter.

3.2.2 Dynamic risk of GR

In a practical scenario it is possible that a doctor or a patient may not want to exceed a certain risk $1 - \pi_j(u \mid t, s)$ of GR since the last biopsy. This may be also useful in the cases where variance of $g(T_j^*)$ is high, rendering expected time of GR, or any other measure of central tendency of $g(T_j^*)$ unsuitable. The personalized scheduling approach based on dynamic risk of GR, schedules the next biopsy at a time point u such that the dynamic risk of GR is higher than a certain threshold $1 - \kappa$, $\kappa\epsilon[0, 1]$ beyond u. To this end, the posterior expected loss for the following multilinear loss function can be minimized to find the optimal u:

$$L_{k_1, k_2}(T_j^*, u) = \begin{cases} k_2(T_j^* - u) & \text{if } T_j^* > u \\ k_1(u - T_j^*) & \text{otherwise} \end{cases} \tag{6}$$

where $k_1 > 0$, $k_2 > 0$ are constants parameterizing the loss function. The posterior expected loss function $E_g\big[L_{k_1, k_2}\{T_j^*, u\}\big]$ obtains its minimum at $u = \pi_j^{-1}\big\{k_1/(k_1 + k_2) \mid t, s\big\}$ (Robert, 2007). The choice of k_1, k_2 is equivalent to the choice of κ. More specifically, $\kappa = k_1/(k_1 + k_2)$.

3.3 A mixed approach between $E_g[T_j^*]$ and dynamic risk of GR

When the variance $\text{var}_g[T_j^*]$ of $g(T_j^*)$ is small, then $E_g[T_j^*]$ as well as $\text{median}[T_j^*]$ are practically very useful. However when the variance is large, there may not be a clear central tendency of the distribution. Thus a biopsy scheduled using $E_g[T_j^*]$ or $\text{median}[T_j^*]$ will overshoot/undershoot T_j^* by a big margin. In case it undershoots the target time, more biopsies will be required until GR is detected at some time point $T_{jN_j^b} > T_j^*$. In case of overshooting the target, the overshooting margin can be measured as an offset $O_j = T_{jN_j^b} - T_j^*$. The maximum acceptable O_j in PRIAS is 3 years, which corresponds to the time gap between biopsies of the PRIAS fixed schedule. When $\text{var}_g[T_j^*]$ is large, the proposals based on $E_g[T_j^*]$ or $\text{median}[T_j^*]$ can have a large O_j. Thus we propose that if the difference between the 0.025 quantile and $E_g[T_j^*]$ or $\text{median}[T_j^*]$ is more than 3 years then proposals based on dynamic risk of GR be used instead. We call this approach a mixed approach.

3.4 Estimation

3.4.1 Estimation of $E_g[T_j^*]$ and $\text{var}_g[T_j^*]$

Since there is no closed form solution available for $E_g[T_j^*]$, for its estimation we utilize the following relationship between expected time of GR and dynamic survival probability:

$$E_g[T_j^*] = t + \int_t^\infty \pi_j(u \mid t, s)\, \mathrm{d}u \tag{7}$$

There is no closed form solution available for the integral and hence we approximate it using Gauss-Kronrod quadrature. We preferred this approach over Monte Carlo methods to estimate $E_g[T_j^*]$ from the posterior predictive distribution $g(T_j^*)$. This was done because sampling directly from $g(T_j^*)$ involved an additional step of sampling from the distribution $p(T_j^* \mid T_j^* > t, \boldsymbol{b_j}, \boldsymbol{\theta})$, as compared to the estimation of $\pi_j(u \mid t, s)$ (Rizopoulos, 2011). Thus the latter approach was computationally faster. As mentioned earlier, a limitation of $E_g[T_j^*]$ is that it is practically useful only when the $\text{var}_g[T_j^*]$ is small. The variance is given by:

$$\text{var}_g[T_j^*] = 2 \int_t^\infty (u - t)\pi_j(u \mid t, s)\, \mathrm{d}u - \left\{ \int_t^\infty \pi_j(u \mid t, s)\, \mathrm{d}u \right\}^2 \tag{8}$$

Since a closed form solution is not available for the variance expression, it is estimated similar to the estimation of $E_g[T_j^*]$. The variance depends both on last biopsy time t and PSA history $\mathcal{Y}_j(s)$. The impact of the observed information on variance is demonstrated in Figure 4 and discussed in Section 4.2.

3.4.2 Estimation of κ

For schedules based on dynamic risk of GR, the value of κ dictates the biopsy schedule and thus its choice has important consequences. In certain cases it may be chosen on the basis of doctor's advice or the amount of risk that is acceptable to the patient. For e.g. if maximum acceptable risk is 75% then $\kappa = 0.25$.

In cases where the choice of k cannot be based on the input of the physician or the patients, we propose to automate the choice of this threshold parameter. More specifically, we propose to choose a κ for which a binary classification accuracy measure (López-Ratón et al., 2014; Sokolova and Lapalme, 2009), discriminating between cases and controls, is maximized. In PRIAS, cases are patients who experience GR and the rest are controls. However, a patient can be in control group at some time t and in the cases at some future time point $t + \Delta t$, and thus time dependent binary classification is more relevant. In joint models, a patient j is predicted to be a case if $\pi_j(t + \Delta t \mid t, s) \leq \kappa$ and a control if $\pi_j(t + \Delta t \mid t, s) > \kappa$ (Rizopoulos, 2016). The time window Δt can be either chosen on a clinical basis, or such that uncertainty in estimation of $\pi_j(t + \Delta t \mid t, s)$ is below a certain threshold or it can even be chosen such that $AUC(t, \Delta t, s)$ (Rizopoulos, 2016) is largest, i.e. Δt for which the model has the most discriminative capability at time t.

To automatically select the threshold κ, the first binary classification accuracy measure we use is Youden's J. It is defined as (the binary classification measures are functions of $t, \Delta t, s$, however the notation is dropped for readability):

$$J = \text{TPR} - \text{FPR}, J\epsilon[-1, 1],$$

$$\text{TPR} = \Pr\big\{\pi_j(t + \Delta t \mid t, s) \leq \kappa \mid T_j^* \epsilon(t, t + \Delta t]\big\},$$

$$\text{FPR} = \Pr\big\{\pi_j(t + \Delta t \mid t, s) > \kappa \mid T_j^* > t + \Delta t\big\}$$

where TPR and FPR denote time dependent true positive rate and false positive rate (Rizopoulos, 2016). The optimal value of κ is $\arg\max_{\kappa} J$. The second binary classification accuracy measure we use is F_1-Score. Unlike Youden's J, F_1-Score is a measure of classification accuracy for cases only. It is defined as:

$$F_1 = \frac{2}{\{TPR\}^{-1} + \{PPV\}^{-1}}, F_1\epsilon[0, 1],$$

$$\text{PPV} = \Pr\big\{T_j^* \epsilon(t, t + \Delta t] \mid \pi_j(t + \Delta t \mid t, s) \leq \kappa\big\}$$

where PPV denotes the time dependent positive predictive value. The optimal value of κ is

$$\arg\max_{\kappa} F_1.$$

3.5 Algorithm

Given the personalized scheduling methods, the next step is to iteratively create an entire schedule until GR is detected for the patient. To this end, the algorithm in Figure 1 elucidates the process of creating a personalized schedule for patient j. Since PRIAS and most AS programs strongly advise against conducting more than 1 biopsy per year, the algorithm adjusts the optimal time u of biopsy in case the last biopsy was performed less than an year ago.

4 Personalized schedules for patients in PRIAS

To demonstrate how the personalized schedules work, we apply them to the patients enrolled in PRIAS. To this end, we divide the PRIAS dataset into training(5264 patients) and demonstration datasets (3 patients). We fit a joint model to the training dataset and then use it to create personalized schedules for patients in demonstration dataset. We fit the joint model using the R package JMbayes (Rizopoulos, 2016), which uses the Bayesian methodology to estimate the model parameters.

4.1 Fitting the joint model to PRIAS dataset

The PRIAS dataset contains PSA levels and time intervals in which GR was detected, for 5264 prostate cancer patients. For every patient the age at the time of induction in AS was recorded. PSA was measured every 3 months for first 2 years and every 6 months thereafter. To detect GR, biopsies were conducted at different time points on the basis of a predetermined schedule as well as PSA-DT as described in Section 1. For the longitudinal analysis of PSA levels we used $\log_2 PSA$ measurements instead of the raw data. This because the PSA scores take very large values around the time of disease progression, indicating that the underlying distribution for PSA values is right skewed. The longitudinal sub-model of the joint model we fit is given by:

$$
\begin{aligned}
\log_2 PSA(t) &= m_i(t) + \varepsilon_i(t), \\
m_i(t) &= (\beta_0 + b_{i0}) + \beta_1(Age - 70) + \beta_2(Age - 70)^2 \\
&\quad + \sum_{k=1}^{4} \beta_{k+2} B_k(t, \mathcal{K}) + b_{i1} B_7(t, 0.1) + b_{i2} B_8(t, 0.1) \\
\varepsilon_i(t) &\sim N(0, \sigma^2),
\end{aligned}
\tag{9}
$$

The evolution of PSA over time is modeled flexibly using B-splines. For the fixed effects part the spline consists of 3 internal knots. The internal knots are at $\mathcal{K} = \{0.1, 0.5, 4\}$ years, and boundary
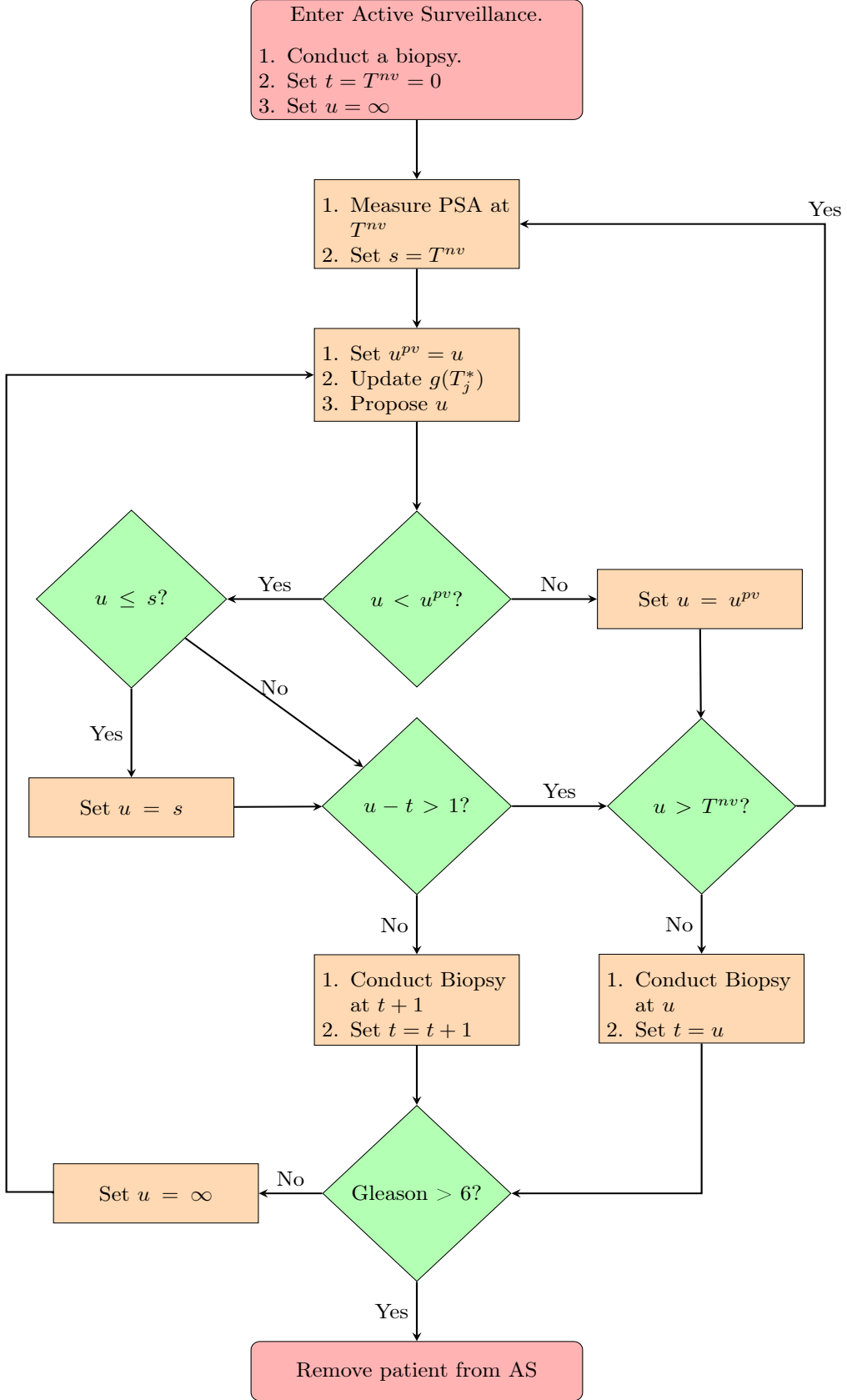
Figure 1: Algorithm for creating a personalized schedule for patient j. t denotes the time of the latest biopsy. s denotes the time of the latest available PSA measurement. u denotes the proposed time of personalized biopsy based on $g(T_j^*)$. u^{pv} denotes the time at which a repeat biopsy was proposed at the last visit to the hospital. T^{nv} denotes the time of the next visit for PSA measurement.

knots are at 0 and 7 years. For the random effects part there is only 1 internal knot at 0.1 years and the boundary knots are at 0 and 7 years. The choice of knots was based on exploratory analysis as well as on the basis of model selection criteria AIC and BIC. Age of patients was median centered to avoid numerical instabilities while estimating the parameters in the model. For the relative risk sub-model the hazard function we fit is given by:

$$h_i(t) = h_0(t) \exp \left[\gamma_1 \{Age - 70\} + \gamma_2 \{Age - 70\}^2 + \alpha_1 m_i(t) + \alpha_2 m'_i(t) \right] \tag{10}$$

where α_1 and α_2 are measures of strength of association between hazard of GR and PSA value $m_i(t)$ and PSA velocity $m'_i(t)$, respectively. As mentioned earlier, in PRIAS study PSA-DT is used to decide the schedule of biopsies. However PSA-DT is computed using observed PSA values, and thus interval censoring observed in PRIAS is independent and non informative of underlying health of the patient.

4.1.1 Parameter Estimates

The posterior parameter estimates for the joint model we fitted to the PRIAS dataset are shown in Table 1 and Table 2. Since the longitudinal evolution of $\log_2 PSA$ is modeled with non-linear terms, the interpretation of the coefficients corresponding to time is not straightforward. In lieu of the interpretation we present the fitted evolution of PSA (Figure 2) over a period of 10 years for a patient who is 70 years old. It can be seen that after the first 6 months the PSA levels steadily increase over the follow up period. Since the model for PSA has only additive terms, this evolution remains same for all patients. The effect of age only affects the baseline PSA score. However it is so small that it can be ignored for all practical purposes.

Table 1: Longitudinal sub-model estimates for joint model.

	Mean	Std. Dev	2.5%	97.5%	P
Intercept	2.455	0.012	2.433	2.480	<0.000
(Age - 70)	0.003	0.001	4.9×10^{-4}	0.006	0.032
(Age - 70) × (Age - 70)	-0.001	1.4×10^{-4}	-0.001	-3.5×10^{-4}	<0.000
Spline: visitTimeYears[0, 0.5]	-0.006	0.012	-0.031	0.017	0.674
Spline: visitTimeYears[0.5, 1.2]	0.228	0.019	0.192	0.265	<0.000
Spline: visitTimeYears[1.2, 2.5]	0.140	0.029	0.088	0.197	<0.000
Spline: visitTimeYears[2.5, 7]	0.303	0.039	0.227	0.379	<0.000
σ	0.324	0.001	0.321	0.326	

For the relative risk sub-model, the parameter estimates in Table 2 show that only $\log_2 PSA$ velocity is strongly associated with hazard of GR. For any patient, a unit increase in $\log_2 PSA$ velocity corresponds to a 11 time increase in hazard of GR. The effect of $\log_2 PSA$ value and effect of Age on hazard of GR are small enough to be safely ignored for all practical purposes.
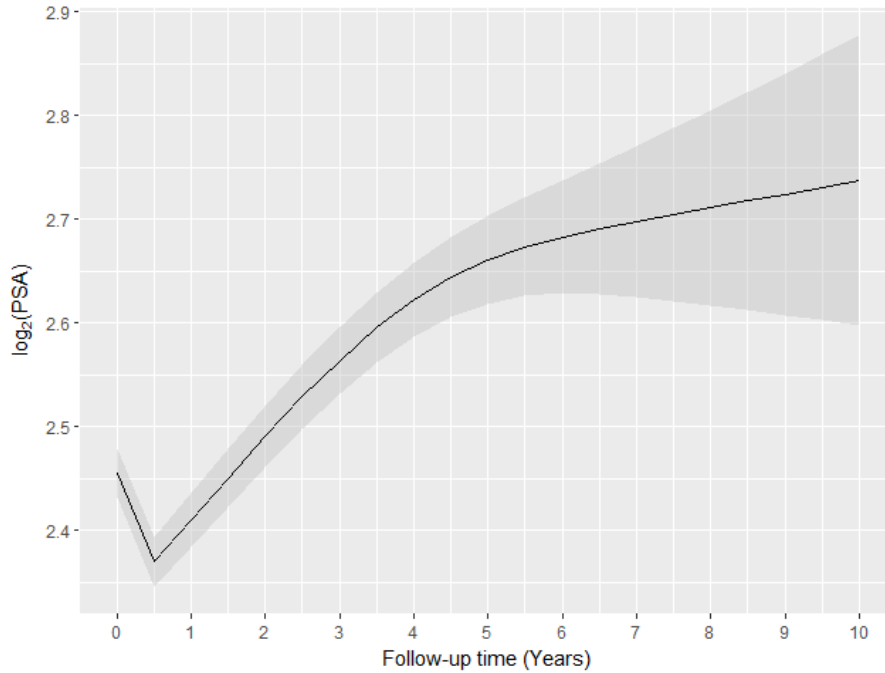
Figure 2: Fitted evolution of $\log_2 PSA$ over a period of 10 years, for a patient who was inducted in AS at the Age of 70 years.

Table 2: Survival sub-model estimates for joint model.

Variable	Mean	Std. Dev	2.5%	97.5%	P
Age - 70	0.037	0.006	0.025	0.0490	<0.000
(Age - 70) $\times$ (Age - 70)	-0.001	0.001	-0.003	1.8×10^{-4}	0.104
$\log_2 PSA$	-0.049	0.064	-0.172	0.078	0.414
Slope: $\log_2 PSA$	2.407	0.319	1.791	3.069	<0.000

4.2 Demonstration of personalized schedules

In this section, we demonstrate how personalized schedules adapt the time of performing a biopsy according to the PSA history and results from repeat biopsies. We demonstrate this using schedules based on expected time of GR and dynamic risk of GR. For the latter we select κ such that Youden's J is maximized (Section 3.4.2). The 3 patients we have chosen for the demonstration dataset are part of PRIAS program and never experienced GR. In addition, they have had their repeat biopsies already. Hence a full scale comparison between PRIAS biopsy schedule and personalized scheduling algorithm's biopsy schedule is not possible.

The first patient of interest is patient 3174 who was inducted in the PRIAS program at the age of 74 years. The posterior predictive distribution $g(T_j^*)$ for this patient depends only on the PSA levels since no repeat biopsies were conducted in the time period we considered. The evolution of PSA, time of last biopsy and proposed times of biopsies are shown in Figure 3. It can be seen that the PSA remains stable for the first 2 years of follow up, but increases rapidly after that for the

14

next 2 years. Since the hazard of GR depends on PSA velocity (Table2), the schedule of biopsy based on expected time of GR adjust the times of biopsy according to the steep rise in PSA profile. More specifically, at 2 years the proposed biopsy time is 12.5 years whereas at 4 years it decreases to 5.3 years. For schedules based on dynamic risk of GR, the κ value which maximized Youden's J was found to be between 1 and 0.9 at all time points. This survival probability corresponds to times very close to the first biopsy (time 0) due to the sharp rise in PSA values. Hence the biopsies are scheduled much earlier than those based on expected time of GR.
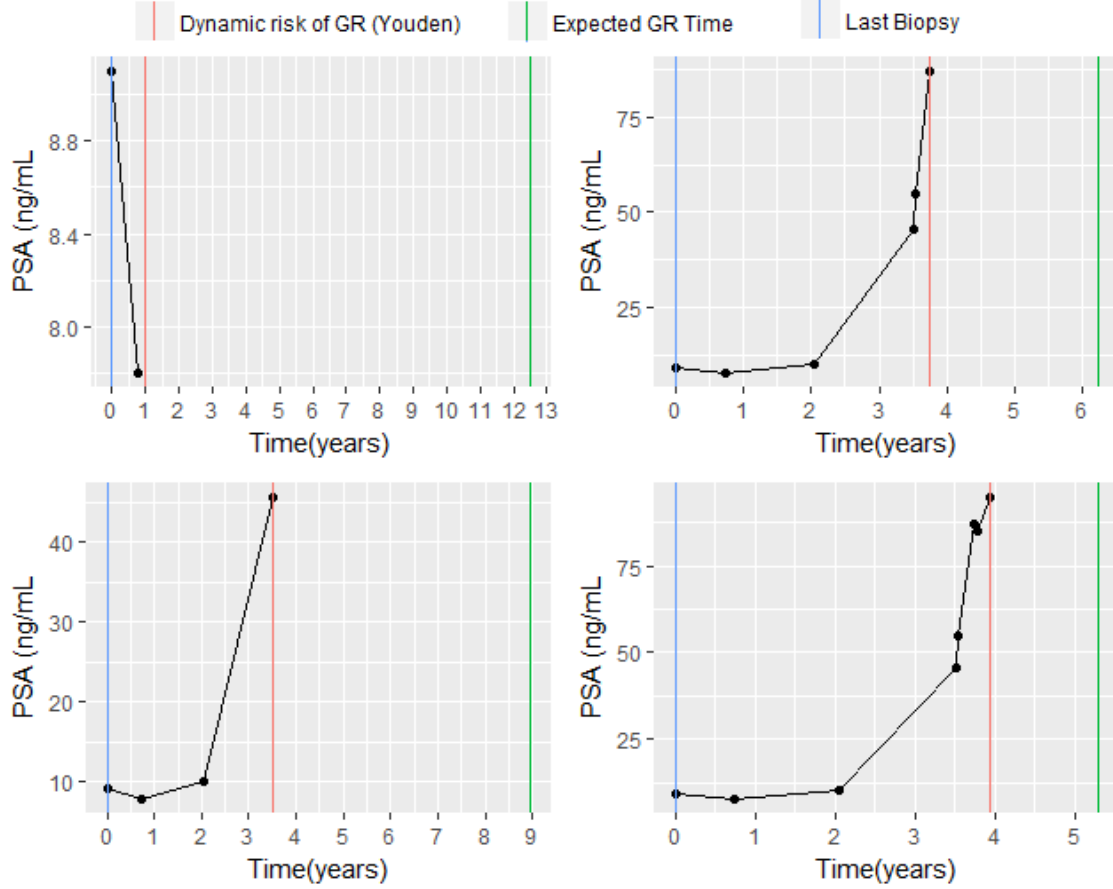
Figure 3: Proposed biopsy times for patient 3174 from PRIAS.

It is important to note that for patient 3174, a biopsy scheduled using expected time of GR at year 2 is not as useful as a biopsy scheduled using the same method at year 4. This because $\text{var}_g[T_j^*]$ is considerably lower at year 4 as shown in Figure 4a. The variance doesn't depend much on number of PSA measurements, but rather on the PSA profile. In the case at hand the variance drops quickly when PSA levels increase sharply, which is in line with PSA velocity $m_i'(t)$ being a strong predictor of the hazard of GR (Table 2).

The second patient of interest is patient 911. Figure 5 shows the evolution of PSA, time of last
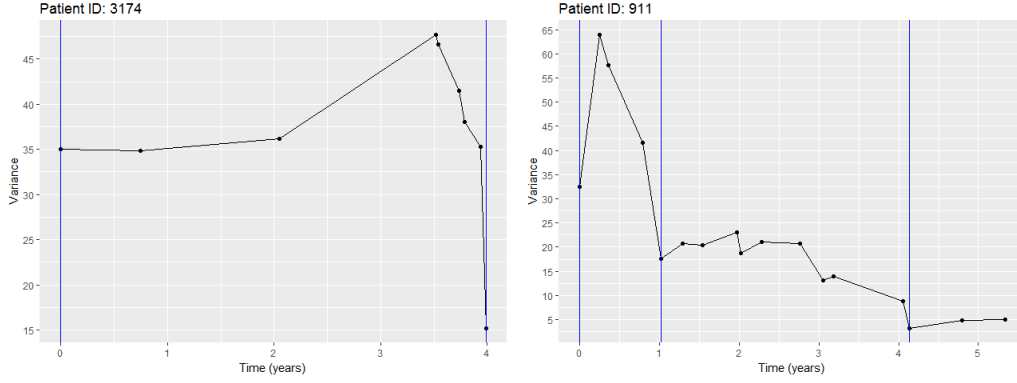
(a) Variance of the predictive distribution $g(T_j^*)$ over a period of 4 years for patient 3174. Blue vertical lines indicate biopsies.

(b) Variance of the predictive distribution $g(T_j^*)$ over a period of 5 years for patient 911. Blue vertical lines indicate biopsies.

Figure 4: Variance of the predictive distribution $g(T_j^*)$

biopsy and proposed biopsy times for this patient. Between year 1.5 and year 2, the PSA rises sharply, and accordingly the personalized schedules based on expected time of GR prepone the proposed biopsy time from 14.2 years to 13.8 years. Between year 2 and year 3 the PSA decreases sharply and accordingly the proposed biopsy times are postponed from 13.8 years to 16.6 years. It can also be seen that PSA remains stable up to to year 4. Lastly, because no GR is found at the repeat biopsy performed at year 4.1, this further leads to postponing of the biopsy times to 18.7 years. As for the schedule based on dynamic risk of GR, the optimal κ values are always between 1 and 0.9 at all time points. Correspondingly, the biopsies are conducted very early until a negative biopsy is found at year 4.1, at which we see that the proposed biopsy time is postponed from 3.21 years to 14.8 years, despite the κ being equal to 0.98 for both. For patient 911 the biopsies scheduled using $E_g[T_j^*]$ at year 4.1 are expected to be very close to the true GR time of the patient. This because the variance $\text{var}_g[T_j^*]$ is quite low at year 4 as shown in Figure 4b. From the figure it is also evident that the variance decreases considerably each time information about T_j^* from a repeat biopsy is available.

From the previous 2 cases we saw that proposed times of biopsy depend on PSA velocity as well as repeat biopsies. Using the profile of patient 2340 we discuss a case where information from PSA levels and repeat biopsies is conflicting. In Figure 6a we can see that the PSA levels for this patient become 4 times between year 1 and year 3, however during the same period a repeat biopsy (year 2.5) was found to be negative. Correspondingly, the personalized schedule based on expected time of GR postpone the time of next biopsy from 14.5 to 15 years. However if there were no repeat biopsy conducted, then only the information from rising PSA levels would've been considered. This is shown in Figure 6b, where we can see that personalized schedule based on expected time of GR prepone the time of next biopsy from 12.5 to 11.5 years. As for dynamic risk
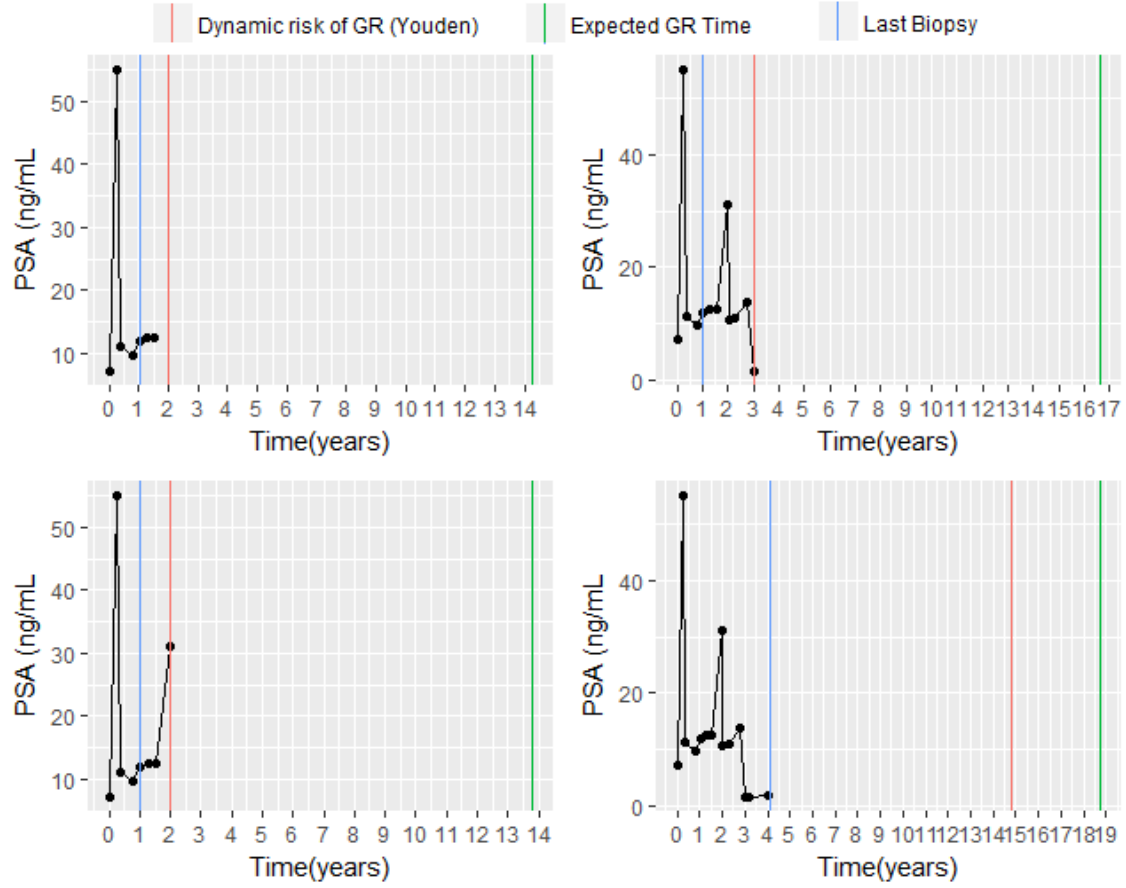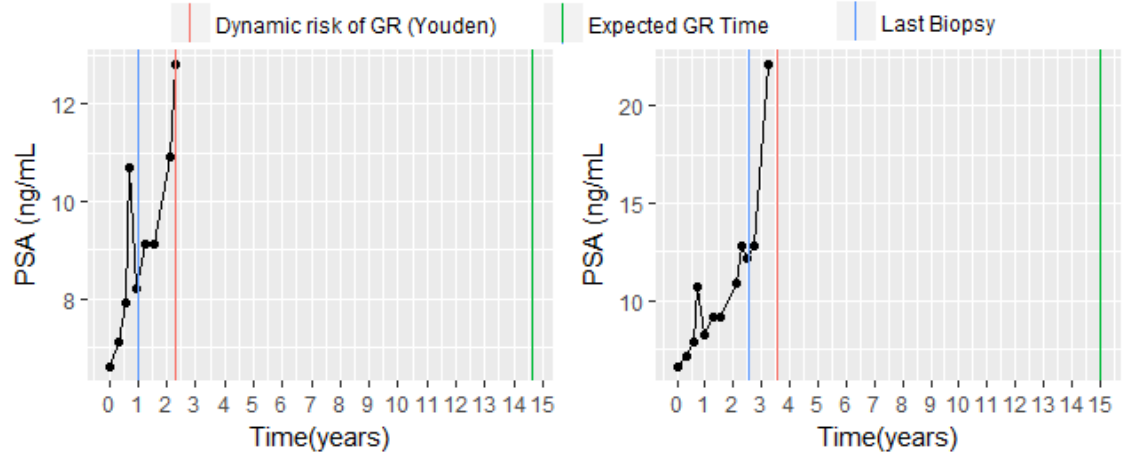
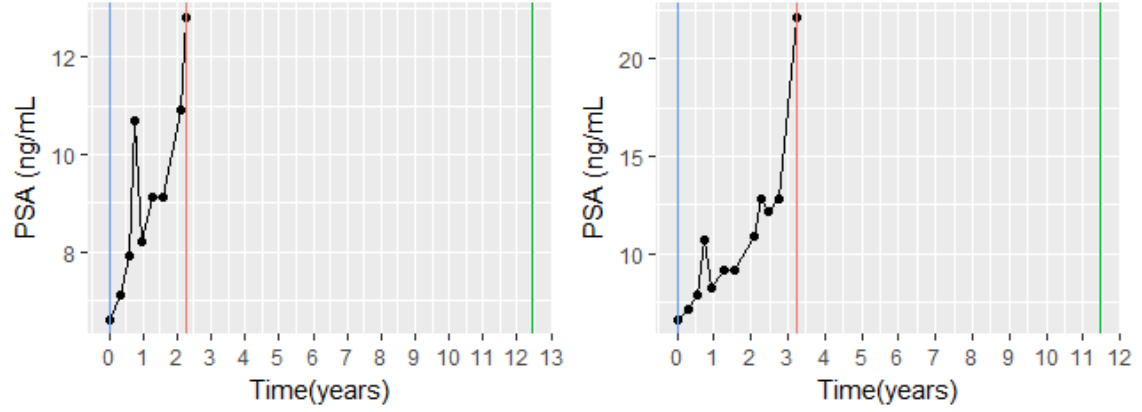Figure 5: Proposed biopsy times for patient 911 from PRIAS.

of GR, we can see that for the same κ, the corresponding biopsy time is 3.5 years when results from repeat biopsy are considered and it is 3.25 years when repeat biopsies are ignored.

5 Simulation study

The application of personalized schedules for patients from PRIAS program demonstrated that personalized schedules adapt according to the PSA and repeat biopsy history of each patient. However, the patients in PRIAS have already had their biopsies as per the PRIAS schedule, and hence evaluation of the efficacy of personalized schedules against the PRIAS schedule was not possible. To this end, we have performed a simulation study to compare 3 broad categories of schedules: Personalized schedules, PRIAS schedule and annual schedule. To compare the schedules we employ them for simulated patients enrolled in a hypothetical AS program, with the same entrance criteria as PRIAS. For these patients we simulate the evolution of PSA and the time of GR. Since our interest is only in the schedule of biopsies, we keep a fixed schedule for measurement of PSA levels. The hypothetical repeat biopsies are conducted until the GR is detected. Although Gleason scores are susceptible to inter-observer variation (Carlson et al., 1998), we assume that

(a) No repeat biopsy is ignored.

(b) All repeat biopsies are ignored.

Figure 6: Proposed biopsy times for patient 2340 from PRIAS.

any biopsy conducted after the true GR time of a patient will lead to GR detection with 100% certainty. We next present the details of the simulation study and the biopsy schedule evaluation criteria.

5.1 Simulation setup

5.1.1 Patient population

For the simulation study we first select a population $\mathcal{P}$ of patients enrolled in AS. We assume that the PSA and hazard of GR for the patients from this population follows a joint model of the form postulated in Section 4.1, with parameters $\boldsymbol{\theta}^{\mathcal{P}}$. These parameters are selected to be equal to the posterior mean of parameters $E[\boldsymbol{\theta} \mid \mathcal{D}^{PRIAS}]$ estimated from the joint model fitted to PRIAS dataset (Section 4.1.1). To demonstrate the efficacy of personalized schedules for patients with early as well late failure times, we generated the patients in population $\mathcal{P}$ from 3 equal sized subgroups G_1, G_2, G_3. The baseline hazard for each of the three subgroups was assumed to be the

18

hazard function of a Weibull distribution. The shape and scale parameters (k, λ) for this Weibull distribution are $(1.5, 4)$, $(3, 5)$ and $(4.5, 6)$ for G_1, G_2 and G_3 respectively. The effect of these parameters is that the variance in GR times is highest for G_1 and lowest for G_3, while the average GR is lowest in G_1 and highest in G_3.

5.1.2 Generating PSA values and GR times

From the population $\mathcal{P}$ we randomly sample a total of 408 datasets with 1000 patients each. For each of the simulated patients the PSA measurement schedule is same as that of PRIAS, i.e. every 3 months for first 2 years and every 6 months thereafter. Each simulated dataset is split into a training (750 patients) and a test (250 patients) part. Keeping in line with the notation for joint model in Section 2.1, the observed data in the k-th training dataset can be represented as $\mathcal{D}^k = \{l_{ki}, r_{ki}, \boldsymbol{y}_{ki}; i = 1, \ldots, 750\}$, where $\boldsymbol{y}_{ki}$ denotes the PSA levels for the i-th patient in the k-th training dataset. For a patient in the training dataset we generate a true event time T^*_{ki} as well as a random and non-informative censoring time C_{ki}. When $T_{ki} < C^*_{ki}$, i.e. when the event is observed, then $l_{ki} = r_{ki} = T^*_{ki}$. When $C_{ki} < T^*_{ki}$, then $l_{ki} = C_{ki}$ and $r_{ki} = \infty$. For the patients in the test datasets censoring time is not generated.

5.1.3 Personalized schedules for test patients

We create personalized biopsy schedules only for the patients in the test dataset. To this end we first fit a joint model with the same specification as for PRIAS. to the training dataset. To model the baseline hazard we use a P-splines approach (Section 2.1). From the fitted joint model we obtain posterior distribution of parameters $p(\boldsymbol{\theta} \mid \mathcal{D}^k)$, which is required for the posterior predictive distribution $g(T^*_{kj})$ of the j-th test patient from the k-th dataset. While the posterior predictive distribution is sufficient for scheduling biopsies based on expected time of GR, the choice of time window Δt (Section 3.2.2) has to be made for scheduling biopsies on the basis of dynamic risk of GR. In PRIAS and in most AS programs biopsies are done at a gap of 1 to 3 years. A gap of 1 year between biopsies detects GR the earliest, and in worst case the detection of GR can be delayed by 1 year. Being a clinically relevant period of time to differentiate between patients who obtain GR and those who don't, we choose a Δt of 1 year. Further, in schedules based on dynamic risk of GR, for a test patient j, we choose a value of κ which maximizes a binary classification measure (Section 3.4.2) at the last known repeat biopsy time t of the patient. To this end the binary classification measures are first computed over a fine grid of κ values in the interval $[0, 1]$ using the training dataset and then the most optimal κ is chosen.

To create personalized schedules we employ the algorithm described in Section 3.5. The al-

gorithm is run for 7 different settings, one each corresponding to the following: PRIAS schedule, annual schedule, expected time of GR, median time of GR, dynamic risk of GR with κ chosen such that a) Youden's J is maximized, b) F_1-Score is maximized. In addition to these a mixed approach (Section 3.3) is also employed where a choice between median time of GR and dynamic risk of GR (Youden's J maximized) is made before scheduling a biopsy.

5.2 Evaluating efficacy of scheduling methods

For a particular biopsy scheduling method S, the first criteria in the evaluation of efficacy of the method is the number of repeat biopsies N^{bS} the corresponding schedule takes before GR is detected for a patient from the population. As we discussed earlier, the less the N^{bS} the better it is for the patients. Our interest lies in the marginal distribution of N^{bS} for the population $\mathcal{P}$. In its simplest form, the mean $E[N^{bS}]$ and variance $\text{var}[N^{bS}]$ of the aforementioned distribution are indicators of the performance of the scheduling algorithm. More specifically, a low mean as well as low variance is desired. Other quantiles of the distribution may also be used evaluation criteria. For example a method which takes less than 2 (say) biopsies in 95% cases may be preferred.

The second criteria in evaluation of efficacy of a schedule is the offset. Given a scheduling method S, the offset for a particular patient j is defined as $O_j^S = T_{jN_j^{bS}}^S - T_j^*$, where N_j^{bS} is the number of biopsies required for patient j before GR is detected and $T_{jN_j^{bS}}^S > T_j^*$ is the time at which GR is detected by the scheduling method S. Once again the interest lies in both the mean $E[O^S]$ and variance $\text{var}[O^S]$ of the marginal distribution of offset.

5.2.1 Finding the most optimal schedule

Given the multiple criteria for efficacy of a schedule the next step is to find the most optimal schedule. Using principles from compound optimal designs (Läuter, 1976) we propose to choose a scheduling method S which minimizes the following loss function:

$$L(S) = \sum_{g=1}^{G} \lambda_g \mathcal{G}_g(N^{bS})^{d_g=1} \mathcal{G}_g(O^S)^{d_g=0} \tag{11}$$

where $\mathcal{G}_g(\cdot)$ is either a function of number of biopsies or of the offset, and d_g is the corresponding indicator for this choice. Some examples of $\mathcal{G}_g(\cdot)$ are mean, median, variance and quantile function. Constants $\lambda_1, \ldots, \lambda_G$, where $\lambda_g \epsilon [0, 1]$ and $\sum_{g=1}^{G} \lambda_g = 1$, are weights to differentially weigh-in the contribution of each of the G evaluation criteria manifested via the functions $\mathcal{G}_g(\cdot)$. An example loss function is:

$$L(S) = \lambda_1 E[N^{bS}] + \lambda_2 E[O^S] \tag{12}$$

The choice of λ_1 and λ_2 is not easy. This because biopsies have serious medical side effects and the cost of an extra biopsy cannot be quantified easily. To obviate this issue we utilize the equivalence between compound and constrained optimal designs (Cook and Wong, 1994). More specifically, it can be shown that for any λ_1 and λ_2 there exists a constant $C > 0$ for which minimization of loss function in (12) is equivalent to minimization of the same, subject to the constraint that $E[O^S] < C$. i.e. The optimal schedule is the one with the least number of biopsies and an offset less than C. The choice of C now can be based on the protocol of AS program. For example in PRIAS the maximum gap between 2 repeat biopsies is 3 years, i.e. If GR is detected within 3 years of its occurrence, it is acceptable. For the optimization in case of more than 2 criteria, such as in (11), the solution can be found by minimizing $\mathcal{G}_G(\cdot)$ under the constraint $\mathcal{G}_g < C_g; g = 1, \ldots, G-1$.

For the loss function in (12), we estimate $E[N^{bS}]$, $\mathrm{var}[N^{bS}]$, $E[O^S]$ and $\mathrm{var}[O^S]$ using pooled estimates of each from the 254 repetitions of the simulation study. The estimates are calculated separately for each of the 7 methods mentioned in Section 5.1.3. The pooled estimates for a scheduling method S are calculated as following:

$$\widehat{E[O^S]} = \frac{\sum_{k=1}^{254} n_k \widehat{E[O_k^S]}}{\sum_{k=1}^{254} n_k},$$

$$\widehat{\mathrm{var}[O^S]} = \frac{\sum_{k=1}^{254} (n_k - 1)\widehat{\mathrm{var}[O_k^S]}}{\sum_{k=1}^{254} (n_k - 1)},$$

where n_k are the number of test patients in the k-th simulation, $\widehat{E[O_k^S]} = \frac{\sum_{j=1}^{n_k} O_{kj}^S}{n_k}$ and $\widehat{\mathrm{var}[O_k^S]} = \frac{\sum_{j=1}^{n_k} \left\{ O_{kj}^S - \widehat{E[O_k^S]} \right\}^2}{n_k - 1}$ are the estimated mean offset and estimated variance of the offset for the k-th simulation, respectively. The estimates for number of biopsies N^{bS} are calculated similarly.

5.3 Results

From the simulations we calculated the pooled estimates of the mean and variance of number of biopsies/offset for the entire sample. The estimates are plotted in Figure 7 and also summarized in Table 3. From the figure it is evident that those schedules which conduct less biopsies on average, have a higher average offset, and vice versa. For example, the annual schedule conducts 5.2 biopsies on average, which is the highest among all schedules, however it has the least average offset of 6 months as well. On the other hand the schedule based on expected time of GR conducts only 1.9 biopsies on average, the least among all schedules but it also has the highest average offset of 15 months. The schedule based on median time of GR performs almost the same as that based on expected time of GR. As mentioned earlier the variance in number of biopsies and offset are important as well. In this regard annual schedule has the largest $\mathrm{var}[N^{bS}]$ since it attempts to

contain the offset within an year, and consequently it has the least $\text{var}[O^S]$. Schedules based on expected and median time of GR perform the opposite in terms of variance.
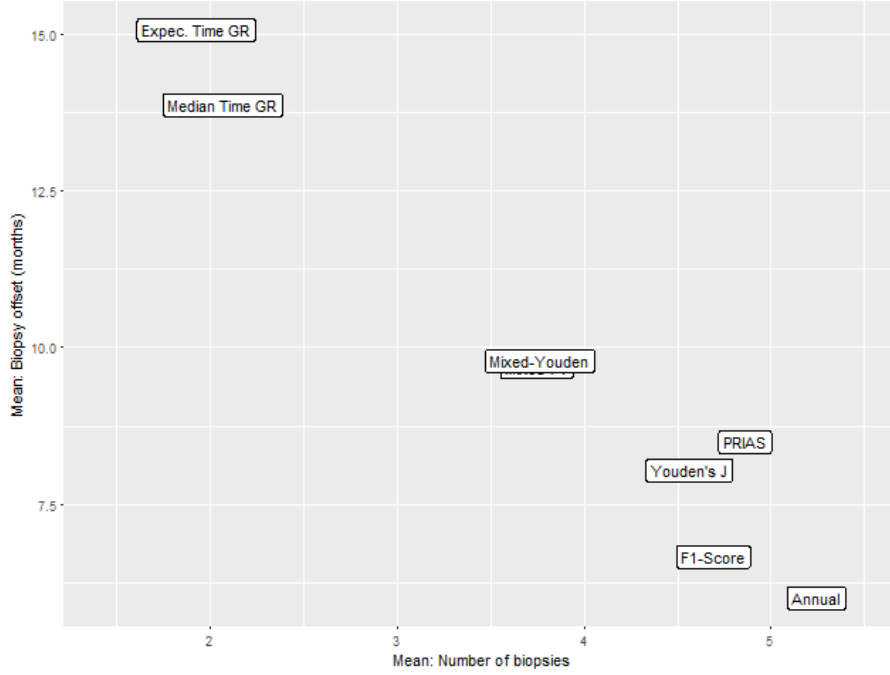

Figure 7: Estimated mean number of biopsies and mean offset (months) for the 7 scheduling methods using all patients. Method names are abbreviated for ease of graphing.

Table 3: Pooled estimates of mean and variance of number of biopsies and offset for all patients.

Schedule	Total Patients	$E[N^{bS}]$	$E[O^S]$	$\text{var}[N^{bS}]$	$\text{var}[O^S]$
Annual	63386	5.237	6.001	6.421	11.874
PRIAS	63386	4.858	8.464	5.523	74.223
Expected time of GR	63386	1.923	15.066	1.422	146.319
Median time of GR	63386	2.066	13.899	2.003	139.773
F_1-Score	63386	4.683	6.654	4.807	18.834
Youden's J	63386	4.560	8.030	4.006	118.908
Mixed approach	63386	3.763	9.744	2.878	58.356

We observe that the PRIAS schedule performs more or less the same as annual schedule. Despite this the latter may be preferred over PRIAS since it conducts only 0.38 biopsies more on average, however unlike PRIAS it has very low variance of offset, thus guaranteeing early detection for everyone. If we compare the PRIAS schedule with dynamic risk of GR based schedules, we can see that the schedule where κ is chosen after maximizing F_1-Score, performs better than in PRIAS schedule in all aspects. The schedule where κ is chosen after maximizing Youden's J has a very large $\text{var}[O^S]$ and hence is not preferable over PRIAS. The mixed approach combines the benefits of methods with low $E[N^{bS}]$ and $\text{var}[N^{bS}]$, and those methods with low $E[O^S]$ and $\text{var}[O^S]$. It conducts 1.5 less biopsies than annual schedule on average and at 9.7 months the mean offset is less than an year.

Table 4: Pooled estimates of mean and variance of number of biopsies and offset for subgroup G_1.

Schedule	Total Patients	$E[N^{bS}]$	$E[O^S]$	var$[N^{bS}]$	var$[O^S]$
Annual	21004	4.306	6.024	9.788	11.747
PRIAS	21004	4.032	7.951	8.221	63.528
Expected time of GR	21001	1.922	15.114	1.441	149.167
Median time of GR	20937	2.068	13.87	1.999	138.396
F_1-Score	21061	4.689	6.648	4.863	18.745
Youden's J	21017	4.581	8.045	3.979	121.309
Mixed approach	21004	3.252	10.361	4.611	73.781

Table 5: Pooled estimates of mean and variance of number of biopsies and offset for subgroup G_2.

Schedule	Total Patients	$E[N^{bS}]$	$E[O^S]$	var$[N^{bS}]$	var$[O^S]$
Annual	21160	5.181	5.95	4.567	12.03
PRIAS	21160	4.817	8.569	3.98	75.716
Expected time of GR	21151	1.927	15.078	1.447	144.333
Median time of GR	21189	2.062	13.947	1.994	140.633
F_1-Score	21133	4.666	6.663	4.726	18.956
Youden's J	21167	4.557	8.024	3.98	122.84
Mixed approach	21160	3.702	10.359	1.869	60.415

Table 6: Pooled estimates of mean and variance of number of biopsies and offset for subgroup G_3.

Schedule	Total Patients	$E[N^{bS}]$	$E[O^S]$	var$[N^{bS}]$	var$[O^S]$
Annual	21222	6.214	6.03	3.118	11.851
PRIAS	21222	5.717	8.866	2.977	82.915
Expected time of GR	21234	1.921	15.006	1.375	145.426
Median time of GR	21260	2.07	13.879	2.016	140.14
F_1-Score	21192	4.695	6.65	4.841	18.789
Youden's J	21202	4.541	8.02	4.061	112.559
Mixed approach	21222	4.33	8.521	1.581	38.586

We next check the performance of these methods for each of the 3 subgroups G_1, G_2 and G_3. Estimates of $E[N^{bS}]$, var$[N^{bS}]$, $E[O^S]$ and var$[O^S]$ for the 3 subgroups are presented in Table 4, Table 5 and Table 6. We observe that all of the schedules which are based on personalized methods, i.e. expected time of GR, median time of GR and dynamic risk of GR based schedules perform the same across the subgroups, with trivial differences in estimates. On the other hand, the annual schedule conducts 6 biopsies on average for patients in G_3 as compared to 4 for patients in G_1. It also has var$[N^{bS}]$ 3 times more for patients in G_1 compared to G_3. This can be attributed to the former having higher variance in GR times. However for annual schedule the $E[O^S]$ and var$[O^S]$ remain almost the same in all groups and it always detects GR within an year of the occurrence. The PRIAS schedule differs for the 3 subgroups as well. For number of biopsies the dynamics are similar to that of annual schedule. However for offset, the PRIAS schedule has high $E[O^S]$ and var$[O^S]$ for patients from G_3, i.e. patients who obtain GR later. As for the mixed approach, we

observe that it conducts more biopsies on average for patients from G_3, however it also has the least $E[O^S]$, $\mathrm{var}[O^S]$ and $\mathrm{var}[N^{bS}]$ for the same group.

To assess the methods further, we combined data from all of the 63386 patients, and also plotted the box plots for number of biopsies and offset in Figure 8a and Figure 8b respectively. Based on the combined data, we observe that both expected and median failure time of GR based schedules have 91.7% and 92.5% of patients below offset cutoff of 36 months, respectively. They also have 80.5% and 82.3% of patients below a cutoff of 24 months. Thus they seem to be quite practical. The mixed approach offers another practically viable solution, since neither it has large $\mathrm{var}[N^{bS}]$, nor $\mathrm{var}[O^S]$. The estimated $E[N^{bS}]$ is 3.8 and the estimated $E[O^S]$ is 9.7 months. For 99.9% patients it has an offset below 36 months and for 95% patients it has an offset below 24 months. Given this offset and the fact that it conducts much less biopsies than PRIAS schedule, annual schedule, and dynamic risk of GR based schedules, it is preferable over them.

(a) Boxplot for number of biopsies.

(b) Boxplot for offset (months)
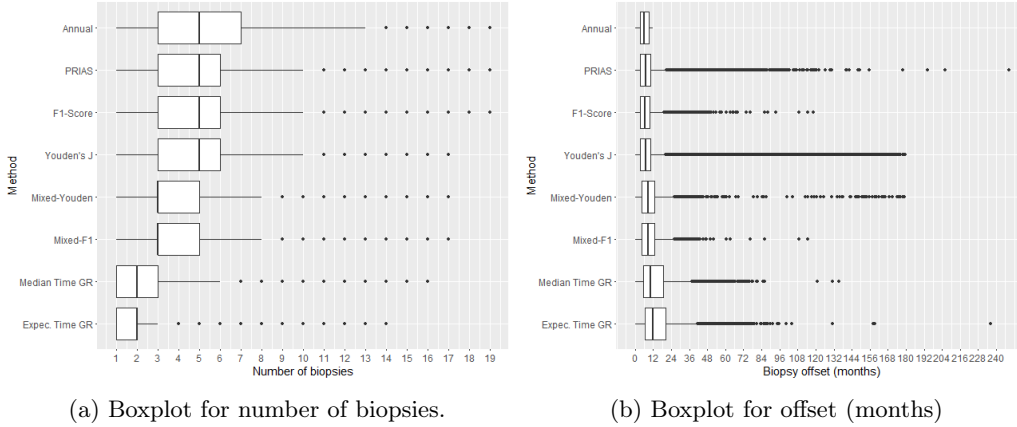
Figure 8: Boxplot for number of biopsies and offset (months), for all of the 63386 patients in the 254 simulated datasets. Method names are abbreviated for ease of graphing.

6 Discussion

In this paper we presented personalized biopsy scheduling methods for patients enrolled in AS programs. The problem at hand was that the AS patients have to undergo repeat biopsies frequently, which causes medical side effects and also brings financial burden. On top of that the existing schedules such as PRIAS schedule had high patient non-compliance because of frequent biopsies and crude analysis of PSA. To approach these problems, we first came up with a joint model to combine the information from PSA as well as repeat biopsies in a more sophisticated manner than the existing PRIAS schedule. Secondly, using the information from the model, we proposed personalized schedules tailored for individual patients. We proposed 2 different class of personalized schedules: those based on central tendency of the distribution of time of GR for indi-

vidual patient, and those based on dynamic risk of GR. In addition we also proposed a combination (mixed approach) of these 2 approaches. Lastly, we also proposed criteria for evaluation of various scheduling methods.

We demonstrated using PRIAS dataset that the personalized schedules adjust the time of biopsy with results from repeat biopsies and PSA profile even when the two are not in complete concordance with each other. Secondly from the simulation study we observed that personalized scheduling method based on dynamic risk of GR (F_1-Score) performed better than PRIAS schedule in terms of both mean and variance of number of biopsies and offset. We also observed that the PRIAS schedule performed quite similar to the annual schedule and the latter may be preferred over it since the latter only conducted 0.38 biopsies more on average but it always detected GR within an year of occurrence. The schedules based on expected and median time to GR conducted only 2 biopsies on average, which is very promising compared to PRIAS and annual schedule which conducted 4.9 and 5.2 biopsies on average respectively. In addition, for schedules based on expected and median time of GR, approximately 92% of the patients had an offset less than 36 months, which is the maximum acceptable offset in PRIAS. If a stronger restriction is prescribed for the offset, then we propose that the mixed approach be used since it offers the best of the two worlds, i.e. not too many biopsies and not too high offset. Lastly, we observed that the personalized methods performed the same for all sub-groups in our population, however the performance of annual and PRIAS schedule was dependent on the failure times of the patients.

While each of the personalized methods has their own disadvantages and advantages, they also offer multiple choices to the AS programs to choose the one as per their requirements, instead of choosing a common fixed schedule for all patients. In this regard, there is potential to develop and analyze more personalized schedules. For e.g. using loss functions which asymmetrically penalize overshooting/undershooting the target GR time can be interesting. Another option is to choose κ on the basis of other binary classification accuracy measures which were not discussed in this paper. While in this work we considered that GR time was interval censored, however the Gleason scores are susceptible to inter-observer variation. Models and schedules which account for error in measurement of time of GR, will be interesting to investigate further.

Acknowledgement

study.

References

Bebu, Ionut and John M. Lachin (2017). "Optimal screening schedules for disease progression with application to diabetic retinopathy". In: *Biostatistics*. DOI: 10.1093/biostatistics/kxv009.

Berger, James O (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Bokhorst, Leonard P et al. (2015). "Compliance rates with the Prostate Cancer Research International Active Surveillance (PRIAS) protocol and disease reclassification in noncompliers". In: *European urology* 68.5, pp. 814–821.

Bokhorst, Leonard P et al. (2016). "A decade of active surveillance in the PRIAS study: an update and evaluation of the criteria used to recommend a switch to active treatment". In: *European Urology* 70.6, pp. 954–960.

Brown, Elizabeth R (2009). "Assessing the association between trends in a biomarker and risk of event with an application in pediatric HIV/AIDS". In: *The annals of applied statistics* 3.3, p. 1163.

Carlson, Grant D et al. (1998). "Accuracy of biopsy Gleason scores from a large uropathology laboratory: use of a diagnostic protocol to minimize observer variability". In: *Urology* 51.4, pp. 525–529.

Cook, R Dennis and Weng Kee Wong (1994). "On the equivalence of constrained and compound optimal designs". In: *Journal of the American Statistical Association* 89.426, pp. 687–692.

Eilers, Paul HC and Brian D Marx (1996). "Flexible smoothing with B-splines and penalties". In: *Statistical science*, pp. 89–102.

Keegan, Kirk A et al. (2012). "Active surveillance for prostate cancer compared with immediate treatment". In: *Cancer* 118.14, pp. 3512–3518.

Lang, Stefan and Andreas Brezger (2004). "Bayesian P-splines". In: *Journal of computational and graphical statistics* 13.1, pp. 183–212.

Läuter, E (1976). "Optimal multipurpose designs for regression models". In: *Mathematische Operationsforschung und Statistik* 7.1, pp. 51–68.

Loeb, Stacy et al. (2013). "Systematic review of complications of prostate biopsy". In: *European urology* 64.6, pp. 876–892.

López-Ratón, Mónica et al. (2014). "OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests". In: *Journal of Statistical Software* 61.8, pp. 1–36.

O'Mahony, James F et al. (2015). "The Influence of Disease Risk on the Optimal Time Interval between Screens for the Early Detection of Cancer: A Mathematical Approach". In: *Medical Decision Making* 35.2, pp. 183–195.

Parmigiani, Giovanni (1998). "Designing observation times for interval censored data". In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 446–458.

Potosky, AL et al. (1995). "The role of increasing detection in the rising incidence of prostate cancer". In: *JAMA* 273.7, pp. 548–552. DOI: 10.1001/jama.1995.03520310046028.

Rizopoulos, Dimitris (2011). "Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data". In: *Biometrics* 67.3, pp. 819–829.

— (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

— (2016). "The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC". In: *Journal of Statistical Software, Articles* 72.7, pp. 1–46. DOI: 10.18637/jss.v072.i07.

Rizopoulos, Dimitris et al. (2014). "Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging". In: *Journal of the American Statistical Association* 109.508, pp. 1385–1397.

Rizopoulos, Dimitris et al. (2016). "Personalized screening intervals for biomarkers using joint models for longitudinal and survival data". In: *Biostatistics* 17.1, p. 149. DOI: 10.1093/biostatistics/kxv031.

Robert, Christian (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal (2017). "Cancer statistics, 2017". In: *CA: A Cancer Journal for Clinicians* 67.1, pp. 7–30. DOI: 10.3322/caac.21387.

Sokolova, Marina and Guy Lapalme (2009). "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4, pp. 427–437.

Taylor, Jeremy MG et al. (2013). "Real-time individual predictions of prostate cancer recurrence using joint models". In: *Biometrics* 69.1, pp. 206–213.

Torre, Lindsey A et al. (2015). "Global cancer statistics, 2012". In: *CA: a cancer journal for clinicians* 65.2, pp. 87–108.

Tosoian, Jeffrey J et al. (2011). "Active surveillance program for prostate cancer: an update of the Johns Hopkins experience". In: *Journal of Clinical Oncology* 29.16, pp. 2185–2190.

Tsiatis, Anastasios A and Marie Davidian (2004). "Joint modeling of longitudinal and time-to-event data: an overview". In: *Statistica Sinica*, pp. 809–834.

Welty, Christopher J et al. (2015). "Extended followup and risk factors for disease reclassification in a large active surveillance cohort for localized prostate cancer". In: *The Journal of urology* 193.3, pp. 807–811.

A   Appendix

A.1   Personalized schedules for repeat biopsies

In this section we present the derivations for Equation 7 and Equation 8. To this end, we first expand the formula for dynamic survival probability presented in Equation 5.

$$
\begin{aligned}
\pi_j(u \mid t, s) &= \Pr\{T_j^* \geq u \mid T_j^* > t, \mathcal{Y}_j(s), D_n) \\
&= \int \int \Pr\{T_j^* \geq u \mid T_j^* > t, \boldsymbol{b}_j, \boldsymbol{\theta}\} p(\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(s), \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}_n) \, \mathrm{d}\boldsymbol{b}_j \, \mathrm{d}\boldsymbol{\theta} \\
&= \int \int \frac{\exp\{-H_j(u|\boldsymbol{b}_j, \boldsymbol{\theta})\}}{\exp\{-H_j(t|\boldsymbol{b}_j, \boldsymbol{\theta})\}} p(\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(s), \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}_n) \, \mathrm{d}\boldsymbol{b}_j \, \mathrm{d}\boldsymbol{\theta} \quad (13)
\end{aligned}
$$

where $H_j(u|\boldsymbol{b}_j, \boldsymbol{\theta}) = \int_0^u h_i(s \mid \boldsymbol{b}_j, \boldsymbol{\theta}) \, \mathrm{d}s$ is the cumulative hazard up to time point u.

A.1.1   Derivation of Equation 7

$$
E_g[T_j^*] = \int_t^\infty T_j^* g(T_j^*) \, \mathrm{d}T_j^*
$$

Using integration by parts, wherein $\dfrac{\mathrm{d}\{-\pi_j(T_j^* \mid t, s)\}}{\mathrm{d}T_j^*} = g(T_j^*)$

$$
\begin{aligned}
&= \left[-T_j^* \pi_j(T_j^* \mid t, s) \right]_t^\infty + \int_t^\infty \pi_j(T_j^* \mid t, s) \frac{\mathrm{d}(T_j^*)}{\mathrm{d}T_j^*} \, \mathrm{d}T_j^* \\
&= t\pi_j(t \mid t, s) - \lim_{T_j^* \to \infty} T_j^* \pi_j(T_j^* \mid t, s) + \int_t^\infty \pi_j(T_j^* \mid t, s) \, \mathrm{d}T_j^*
\end{aligned}
$$

where $\pi_j(t \mid t, s) = \Pr\{T_j^* \geq t \mid T_j^* > t, \mathcal{Y}_j(s), D_n\} = 1$. As for $\lim_{T_j^* \to \infty} T_j^* \pi_j(T_j^* \mid t, s)$, the limit can be interchanged with the integral in Equation 13, because as $T_j^* \to \infty$ the integrand in the equation converges uniformly on the domain of $(\boldsymbol{b}_j, \boldsymbol{\theta})$. Thus,

$$
\lim_{T_j^* \to \infty} T_j^* \pi_j(T_j^* \mid t, s) = \int \int \lim_{T_j^* \to \infty} \frac{T_j^*}{\exp\{H_j(T_j^*|\boldsymbol{b}_j, \boldsymbol{\theta})\}} \frac{p(\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(s), \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}_n)}{\exp\{-H_j(t|\boldsymbol{b}_j, \boldsymbol{\theta})\}} \, \mathrm{d}\boldsymbol{b}_j \, \mathrm{d}\boldsymbol{\theta}
$$

Using L'Hospital's rule

$$
\begin{aligned}
\lim_{T_j^* \to \infty} T_j^* \pi_j(T_j^* \mid t, s) &= \int \int \frac{1}{\lim_{T_j^* \to \infty} \exp\{H_j(T_j^*|\boldsymbol{b}_j, \boldsymbol{\theta})\} H_j'(T_j^*|\boldsymbol{b}_j, \boldsymbol{\theta})} \frac{p(\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(s), \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}_n)}{\exp\{-H_j(t|\boldsymbol{b}_j, \boldsymbol{\theta})\}} \, \mathrm{d}\boldsymbol{b}_j \, \mathrm{d}\boldsymbol{\theta} \\
&= \int \int 0 \frac{p(\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(s), \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}_n)}{\exp\{-H_j(t|\boldsymbol{b}_j, \boldsymbol{\theta})\}} \, \mathrm{d}\boldsymbol{b}_j \, \mathrm{d}\boldsymbol{\theta} \\
&= 0
\end{aligned}
$$

In light of these results, we obtain:

$$
E_g[T_j^*] = t + \int_t^\infty \pi_j(T_j^* \mid t, s) \, \mathrm{d}T_j^*
$$

A.1.2 Derivation of Equation 8

Since $\text{var}_g[T_j^*] = E_g[\{T_j^*\}^2] - E_g[T_j^*]^2$, we first show the derivation for $E_g[(T_j^*)^2]$.

$$E_g[(T_j^*)^2] = \int_t^\infty (T_j^*)^2 g(T_j^*)\, \mathrm{d}T_j^*$$

Using integration by parts, wherein $\dfrac{\mathrm{d}\{-\pi_j(T_j^* \mid t,s)\}}{\mathrm{d}T_j^*} = g(T_j^*)$

$$E_g\big[\{T_j^*\}^2\big] = \Big[-\{T_j^*\}^2 \pi_j(T_j^* \mid t,s)\Big]_t^\infty + \int_t^\infty \pi_j(T_j^* \mid t,s)\frac{\mathrm{d}(T_j^*)^2}{\mathrm{d}T_j^*}\,\mathrm{d}T_j^*$$

$$= t^2 \pi_j(t \mid t,s) - \lim_{T_j^* \to \infty}(T_j^*)^2 \pi_j(T_j^* \mid t,s) + 2\int_t^\infty T_j^* \pi_j(T_j^* \mid t,s)\,\mathrm{d}T_j^*$$

$$= t^2 + 2\int_t^\infty T_j^* \pi_j(T_j^* \mid t,s)\,\mathrm{d}T_j^*$$

Therefore,

$$\text{var}_g[T_j^*] = t^2 + 2\int_t^\infty T_j^* \pi_j(T_j^* \mid t,s)\,\mathrm{d}T_j^* - \bigg[t^2 + \Big\{\int_t^\infty \pi_j(T_j^* \mid t,s)\,\mathrm{d}T_j^*\Big\}^2 + 2t\int_t^\infty \pi_j(T_j^* \mid t,s)\,\mathrm{d}T_j^*\bigg]$$

$$= 2\int_t^\infty (T_j^* - t)\pi_j(T_j^* \mid t,s)\,\mathrm{d}T_j^* - \Big\{\int_t^\infty \pi_j(T_j^* \mid t,s)\,\mathrm{d}T_j^*\Big\}^2$$

A.2 Simulation study

This section presents the figures and tables for related to the simulation study results from Section 5. More specifically:

- Figure 9a, Figure 9b and Figure 9c show plots of average number of biopsies against average offset for each of the 7 scheduling methods present in the simulation study.

- Figure 10a and Figure 10b show boxplots of number of biopsies and offset for each of the 6 scheduling methods for the patients in subgroup G_1. Similar plots for subgroup G_2 and G_3 are shown in Figure 11 and Figure 12 respectively.

- Figure 13a and Figure 13b show the variation in estimated mean number of biopsies and estimated mean offset computed during the 254 simulations, for patients from all subgroups. The same plots for each of the subgroups are shown in Figure 14, Figure 15 and Figure 16.

(a) Plot for subgroup G_1

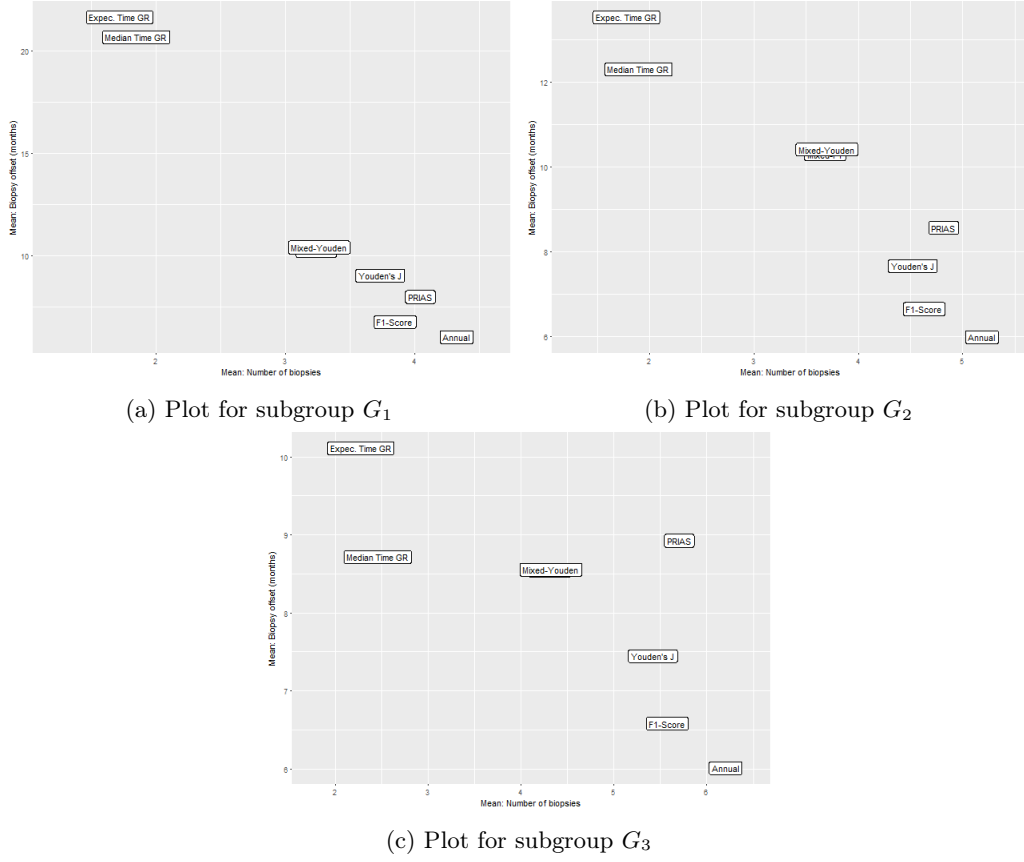(b) Plot for subgroup G_2

(c) Plot for subgroup G_3

Figure 9: Estimated mean number of biopsies and mean offset (months) for the 7 scheduling methods for the 3 sub-groups. Method names are abbreviated for ease of graphing.
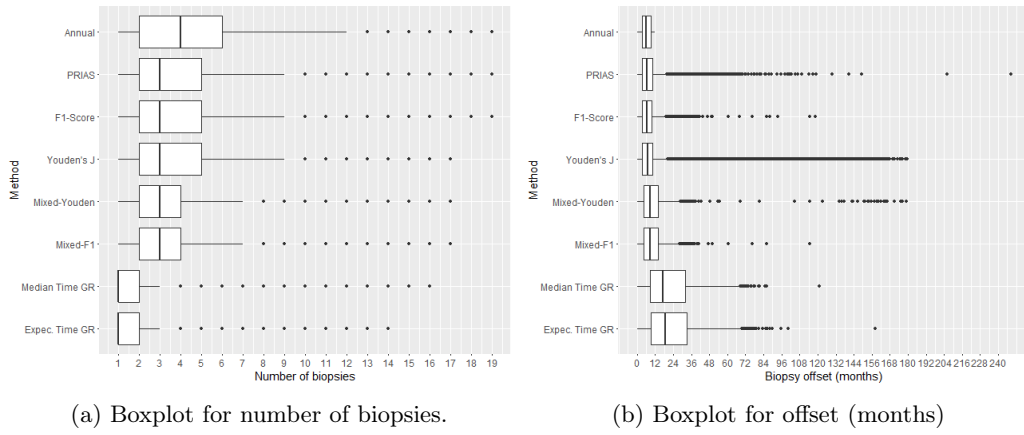

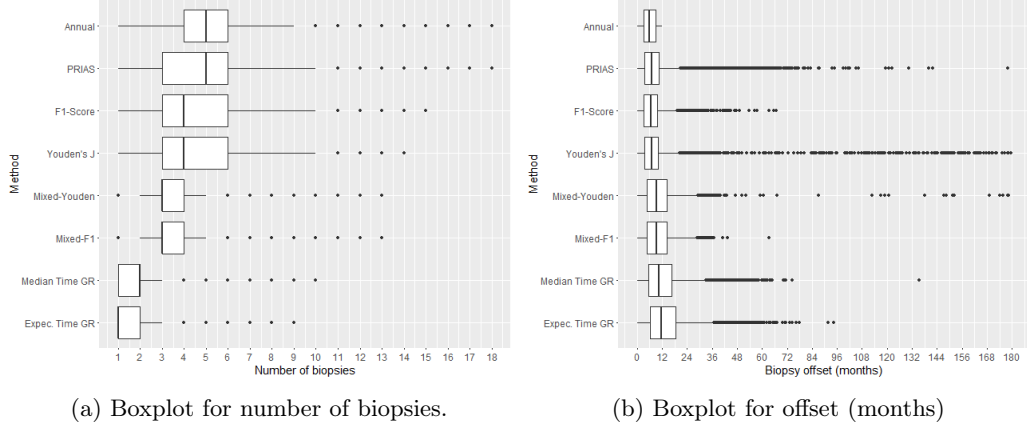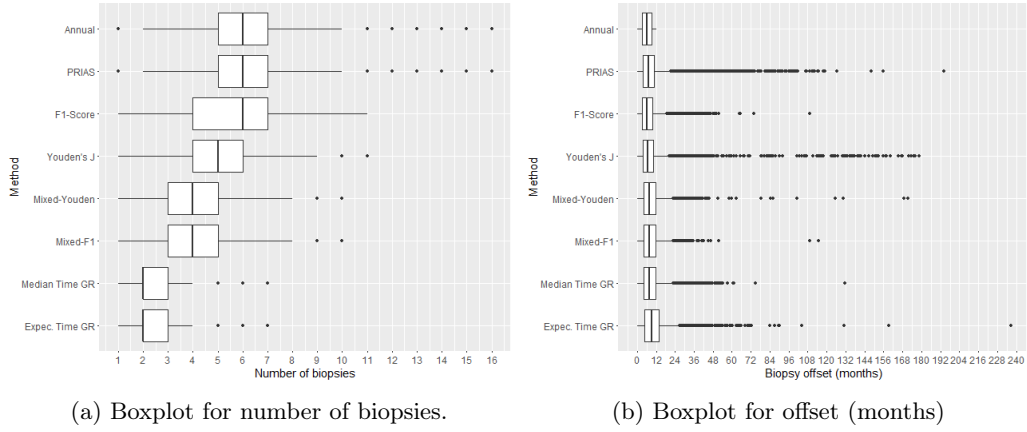(a) Boxplot for number of biopsies.

(b) Boxplot for offset (months)

Figure 10: Boxplot for number of biopsies and offset (months), for all patients in subgroup G_1 across all simulations. Method names are abbreviated for ease of graphing.
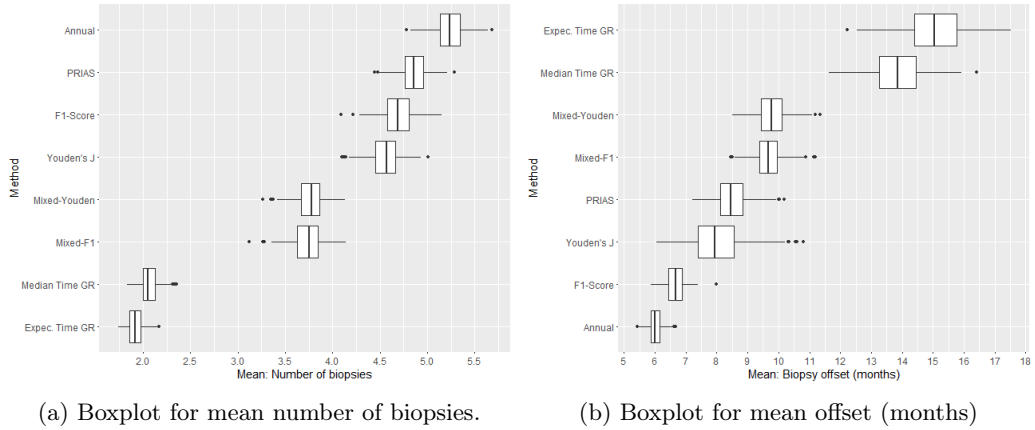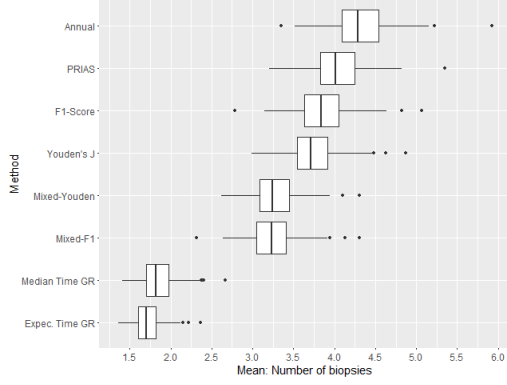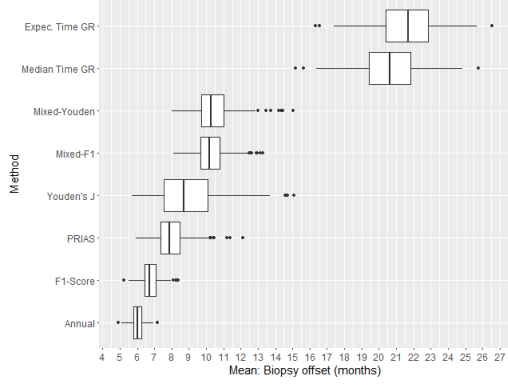
(a) Boxplot for number of biopsies.

(b) Boxplot for offset (months)

Figure 11: Boxplot for number of biopsies and offset (months), for all patients in subgroup G_2 across all simulations. Method names are abbreviated for ease of graphing.

(a) Boxplot for number of biopsies.

(b) Boxplot for offset (months)

Figure 12: Boxplot for number of biopsies and offset (months), for all patients in subgroup G_3 across all simulations. Method names are abbreviated for ease of graphing.
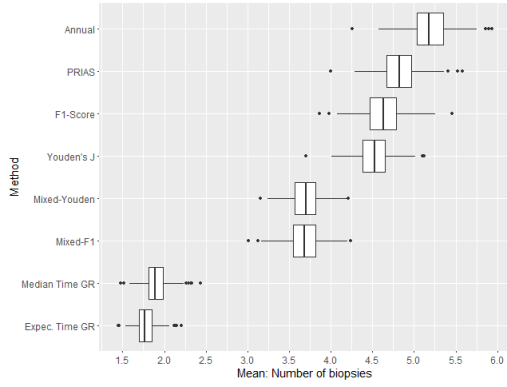

(a) Boxplot for mean number of biopsies.

(b) Boxplot for mean offset (months)

Figure 13: Boxplot showing variation in mean number of biopsies and mean offset (months) for all of the patients across the 254 simulations.
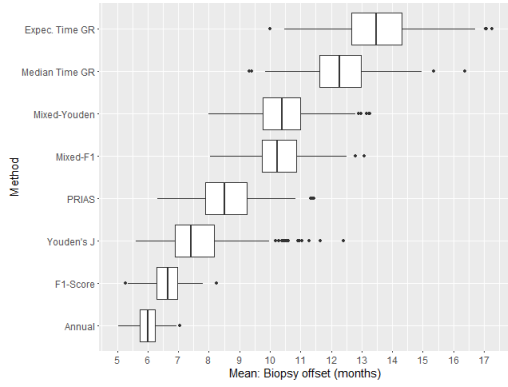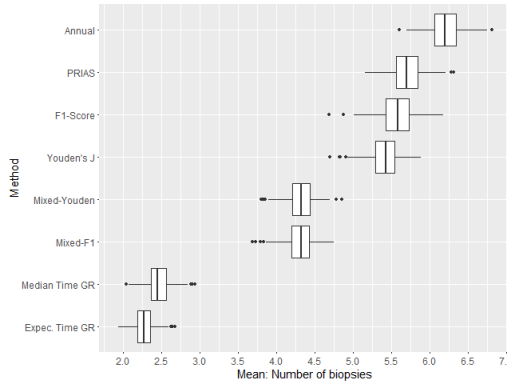
(a) Boxplot for mean number of biopsies.

(b) Boxplot for mean offset (months)

Figure 14: Boxplot showing variation in mean number of biopsies and mean offset (months) for patients in subgroup G_1 across the 254 simulations.

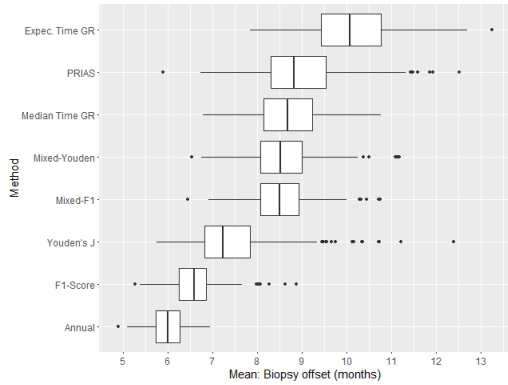(a) Boxplot for mean number of biopsies.

(b) Boxplot for mean offset (months)

Figure 15: Boxplot showing variation in mean number of biopsies and mean offset (months) for patients in subgroup G_2 across the 254 simulations.

(a) Boxplot for mean number of biopsies.

(b) Boxplot for mean offset (months)

Figure 16: Boxplot showing variation in mean number of biopsies and mean offset (months) for patients in subgroup G_3 across the 254 simulations.