PRIAS Personalized Biopsy Schedules

Firstname1 Lastname1

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse accumsan magna est, quis elementum leo laoreet eu. Donec sollicitudin elit non massa venenatis, in viverra dolor sagittis. Maecenas ac justo pulvinar, consectetur mauris hendrerit, vulputate lacus. Etiam tristique sapien quis sem commodo, et eleifend tortor viverra. In hac habitasse platea dictumst. Phasellus vel tempus risus, sit amet consectetur massa. Duis rutrum lectus eu ligula egestas iaculis. Sed condimentum, ipsum in dignissim condimentum, nisi turpis blandit massa, et aliquam magna ligula eget lacus. Donec ac eleifend nulla, quis cursus nisi. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse accumsan magna est, quis elementum leo laoreet eu. Donec sollicitudin elit non massa venenatis, in viverra dolor sagittis. Maecenas ac justo pulvinar, consectetur mauris hendrerit, vulputate lacus. Etiam tristique sapien quis sem commodo, et eleifend tortor viverra. In hac habitasse platea dictumst. Phasellus vel tempus risus, sit amet consectetur massa. Duis rutrum lectus eu ligula egestas iaculis. We demonstrate that personalized schedules are a very promising method to decide biopsy times for prostate cancer patients.

1 Introduction

Prostate cancer is the development of cancer in the prostate gland. With increase in life expectancy and increase in number of screening tests, an increase in diagnosis of low grade prostate cancers has been observed. Majority of these cancers have good long-term survival and in many cases the prostate cancer is (over) diagnosed solely due of screening. i.e. it wouldn't have shown any malignant symptoms for a long time otherwise. To avoid overtreatment, patients diagnosed with prostate cancer are often motivated to join active surveillance (AS) programs instead of taking immediate treatment. The goal of AS programs 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, with 10 corresponding 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(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 the gold standard in this regard (Bokhorst et al., 2016). If either the DRE or the Gleason score are found to be above the aforementioned threshold, then it is considered that the disease has progressed and 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 here onwards).

The reliability of Gleason score comes at a high cost. Biopsies are difficult to obtain, are painful and have serious side effects such as hematuria and sepsis for prostate cancer patients (Loeb et al., 2013). So much so, that PRIAS as well as majority of the AS programs around the world strongly adhere to the rule of not having more than 1 biopsy per year. Performing a biopsy every year has the advantage that it is possible to detect GR within 1 year since 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 the scenario where biopsy is performed every year. Despite this, several AS studies schedule biopsies for every patient annually (Tosoian et al., 2011; Welty et al., 2015). For patients enrolled in PRIAS the schedule is comparatively less rigorous. One biopsy is performed at the time of induction, and the rest are scheduled at 1, 4, 7, 10 years and every 5 years thereafter. For patients who have a PSA doubling time (PSA-DT) less than 10 years, repeat biopsy every year is advised.

The biopsy schedule of PRIAS program is less rigorous than other programs, and yet PRIAS 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. On the other hand even if patients comply with the schedule, be it annually or the schedule of PRIAS, it is not suitable for every patient. Patients whose cancers progress slowly often end up having biopsies when they are not needed. For patients who have faster progressions, crude measures such as PSA-DT are employed to decide if frequent biopsies are required. 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'.

Let us assume that we have a new patient j enrolled in the AS program. The most useful biopsy schedule for him will be the one with the least number of biopsies N_j^b and the smallest offset $O_j = T_j^o - T_j^*$ possible, where T_j^o is the time at which GR is observed and T_j^* is the actual time at which GR occurred. The search for the most useful biopsy schedule is the motivation behind this work. To this end, we have proposed alternative biopsy schedules, belonging to a class of schedules called personalized schedules. Personalized schedules are tailored separately for every patient and every disease. A simple example is the PRIAS schedule, which is personalized since it depends on the PSA-DT of the patient, an indicator of the health. More sophisticated personalized schedules have been developed in the past. For e.g. Bebu and Lachin, 2017 have proposed Markov models based cost optimized personalized schedules. 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. Rizopoulos et al., 2016 have proposed dynamic personalized schedules for longitudinal markers using joint models for time to event and longitudinal data(Rizopoulos, 2012; Tsiatis and Davidian, 2004).

The personalized schedules we have proposed in this paper, utilize joint models and are dynamic. i.e. at a time only one future visit is scheduled, based on all the information gathered up to that point in time. More specifically, we have proposed two types of personalized schedules. One based on expected time of GR of a patient and the second based on the risk of GR. We have also analyzed an approach where the two types of personalized schedules are combined. Both types of schedules not only consider a patient's measurable attributes such as age, but also latent patient to patient variations in health, which cannot be measured directly. Results from previous repeat biopsies of the patient and PSA measurements as well as the population level information about hazard of GR, is used by the personalized schedules we have proposed. It is important to note that a schedule for DRE measurements 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 disease progression or DRE crossing the threshold of cT2c.

Using joint models to model the association between PSA measurements and risk of GR has the advantage that this association is modeled via random effects, and therefore the models have an inherent patient specific nature. Secondly, they allow modeling the entire longitudinal history of PSA measurements, which is more sophisticated than PSA-DT. The use of PSA measurements in creating a personalized schedule is important because PSA is easy to measure, is cost effective and does not have any side effects. Secondly, in PRIAS Bokhorst et al., 2015 found that compliance rate for PSA measurements was as high as 91%. They also showed that there were more men who

had a Gleason score greater than 6 as well as PSA-DT less than 3 years compared to men who had Gleason > 6 as well as PSA-DT larger than 3 years. i.e. Information from PSA was found to be indicative of GR. Lastly, some patients/doctors in PRIAS did not comply with the biopsy schedule because they considered PSA to be stable. Therefore, if information from PSA is used in a methodical manner, it can lead to a more informative medical decision making process.

The rest of the paper is organized as follows. Section 2.1 covers briefly the joint modeling framework in context of the problem at hand. Section 2.2 details the personalized scheduling approaches we have proposed in this paper. In section 3 we demonstrate the efficacy of personalized schedules in a real world scenario by employing them for the patients from the PRIAS study. Lastly, in section 4, we present the results from a simulation study we conducted, to compare personalized schedules with the schedule of PRIAS study, as well with the most aggressive biopsy schedule of doing annual biopsies.

2 A framework for personalized biopsy schedules

The first step in creating a personalized schedule for biopsies is to come up with a model for Gleason scores, PSA levels and other patient specific characteristics. In PRIAS, PSA levels are measured at the time of induction, every 3 months for the first 2 years in the study and then every 6 months thereafter. Thus PSA levels can be modeled as a longitudinal outcome. As mentioned earlier, patients in PRIAS have a Gleason score of 6 or less at the time of induction in the study, and patients are removed from AS the first time GR takes place. Since our interest lies in finding the time of GR, we model it as a time to event outcome. While univariate modeling of the two aforementioned outcomes can be done separately using longitudinal model and survival models, it is important to note that they are not independent. To model the association between the two types of outcomes we use a joint model for time to event and longitudinal outcomes.

2.1 Joint model for risk of GR and PSA levels

Let T_i^* denote the true GR time for the i^{th} patient. Let the times at which biopsies are conducted for the i^{th} patient be denoted by $C_i = \{C_{i0}, C_{i1}, ...C_{ig_i}; C_{ij} < C_{ik}, \forall j < k\}$, where g_i is 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 $l_i = C_{i(g_i-1)}, r_i = C_{ig_i}$ if GR is observed, and $l_i = C_{ig_i}, r_i = \infty$ if patient drops out. The latter is also known as right censoring. Let $\boldsymbol{y}_i$ denote the $n_i \times 1$ longitudinal outcome vector for the PSA levels of the i^{th} patient. The population of interest is all the patients enrolled in AS. For a sample of n patients from this population the complete data is denoted by $\mathcal{D}_n = \{T_i, l_i, r_i, \boldsymbol{y}_i; i = 1, ...n\}$, where $T_i \epsilon (l_i, r_i]$ denotes the observed time of GR.

To model the evolution of the longitudinal outcome, which is PSA levels in the case at hand, the joint model utilizes a linear mixed effects model. The longitudinal outcome $y_i(t)$ at time t is modeled as:

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

where, $m_i(t)$ denotes the true and unobserved value of the longitudinal outcome at time t. The measurement error $\varepsilon_i(t) \sim N(0, \sigma^2)$ is assumed normally distributed with variance σ^2. $\boldsymbol{\beta}$ denotes the vector of the unknown fixed-effects parameters. $\boldsymbol{b}_i \sim N(0, \boldsymbol{D})$ denotes the $q \times 1$ vector of random effects, assumed normally distributed with mean zero, and $qtimesq$covariance matrix D. $\boldsymbol{b}_i$ and $\varepsilon_i(t)$ are assumed independent. $\boldsymbol{x}_i(t)$ and $\boldsymbol{z}_i(t)$ denote row vectors of the design matrices for the fixed and random effects, respectively. For non continuous longitudinal outcomes joint models utilize Generalized linear mixed models (Rizopoulos, 2012).

Since both PSA levels and Gleason scores are affected by the state of prostate cancer, they are inherently correlated with each other. To this end, joint models utilize a relative risk sub-model where the hazard of GR at any time point t, denoted by $h_i(t)$, depends on the history of true and unobserved values of PSA levels $\mathcal{M}_i(t) = \{m_i(v), 0 \leq v \leq t\}$ measured up to that time point.

Joint models offer flexibility in modeling this dependence. In its simplest form, the hazard may depend on instantaneous value of PSA $m_i(t)$ at time t. More sophisticated ones are dependence of hazard at time t on PSA-DT, PSA velocity $m_i'(t) = \dfrac{dm_i(t)}{dt}$, or even on the cumulative effect of PSA $\int_0^t m_i(s)\,ds$ up to t. The fact that any functional form of dependence is possible, is evident from the following equation:

$$h_i(t \mid M_i(t), \boldsymbol{w}_i) = h_0(t) e^{\boldsymbol{\gamma}^T \boldsymbol{w}_i + f\{M_i(t), \boldsymbol{b}_i, \boldsymbol{\alpha}\}}$$

where $h_0(t)$ is the baseline hazard at 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 function form of longitudinal outcome that is used in the linear predictor of the relative risk model.

While $\boldsymbol{\alpha}$ controls the strength of association between the hazard of GR and features of the PSA levels, the fact that both Gleason scores and PSA levels are internally related to a patient's health, is manifested by the random effects $\boldsymbol{b}_i$ in the model. The joint model postulates that given the random effects, time to GR and PSA levels measured at different time points are mutually independent. As mentioned earlier, in PRIAS study PSA-DT is used to decide the schedule of biopsies. Although PSA-DT is computed using observed PSA values, dependence on observed longitudinal history $\mathcal{Y}(t) = \{y_i(v), 0 \leq v \leq t\}$ at any time t, is not the same as dependence on patient's health. This because dependence on patient's health, manifested by $\boldsymbol{b}_i$ is same as dependence on future unobserved values of PSA. Thus the inference for the model parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\alpha}^T, \sigma^2, \{d_{jk} \mid j = k = 1, ...q\}\}^T$ doesn't change even if uncertainty in biopsy schedule C_i is not modeled. The kernel of the corresponding joint likelihood conditional on the random effects and the model parameters is given by:

$$p(T_i, l_i, r_i, \boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) \propto p(T_i \mid l_i, r_i, \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})$$

2.2 Personalized scheduling approaches

Once a joint model for GR and PSA levels is obtained, the next step is to use it to create personalized schedules for biopsies. In this section we present the various personalized biopsy scheduling approaches and their motivation. The personalized schedules that we propose are dynamic in nature and thus at any given time, only 1 future biopsy is scheduled. The age of the patient and entire PSA, repeat biopsy history up to that time point is considered while computing the time of next biopsy. To elucidate the scheduling methods, let us assume that the a personalized schedule is to be created 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 their last biopsy, performed at time t and that the PSA measurements are available for the patient up to a time point s. Combining these two pieces of information, the predictive distribution $g(T_j^*)$ for time to GR for this patient is given by (conditioning on baseline covariates $\boldsymbol{w}_i$ is dropped for notational simplicity here onwards):

$$g(T_j^*) = p(T_j^* \mid T_j^* > t, \mathcal{Y}_j(s), \mathcal{D}_n) \tag{1}$$

where $\mathcal{Y}_j(s)$ denotes the history of PSA measurements done up to time s. Given the predictive distribution, our goal is find the optimal time $u \geq \max(t, s)$ of the next biopsy. To this end, 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 $E_g[L(T_j^*, u)]$, where the expectation is taken w.r.t. the predictive distribution $g(T_j^*)$.

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

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.

2.2.1 Expected 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. In this regard, the squared loss function $L(T_j^*, u) = (T_j^* - u)^2$ has the property that the loss increases quadratically as the error $T_j^* - u$ increases. So the squared loss function satisfies our requirement of choosing a u as close to the true GR time as possible. The posterior expected loss is given by:

$$\begin{aligned} E_g[L(T_j^*, u)] &= E_g[(T_j^* - u)^2] \\ &= E_g[(T_j^*)^2] + u^2 - 2uE_g[T_j^*] \end{aligned} \tag{2}$$

The posterior expected loss in equation 2 attains its minimum at $u = E_g[T_j^*]$, also known as expected time of GR.

Estimation

Since there is no closed form solution available for $E_g[T_j^*]$, for its estimation we introduce a construct called dynamic survival probability (Rizopoulos, 2011). The dynamic survival probability $\pi_j(v \mid t, s)$ of patient j is the survival probability at time v, conditional on the observed PSA history $\mathcal{Y}_j(s)$ and the fact that the patient did not have GR up to t. It is given by:

$$\pi_j(v \mid t, s) = Pr(T_j^* \geq v \mid T_j^* > t, \mathcal{Y}_j(s), D_n), v \geq t \tag{3}$$

The relationship between expected time of GR and dynamic survival probability is given by:

$$E_g[T_j^*] = t + \int_t^\infty \pi_j(v \mid t, s)\, dv$$

Since the R package JMbayes already provides an implementation of $\pi_j(v \mid t, s)$, this approach was preferred over Monte Carlo methods to estimate $E_g[T_j^*]$ from the predictive distribution $g(T_j^*)$. A limitation of expected time of GR though, is that it is only useful when the variance of predictive distribution $g(T_j^*)$ is small. The variance is given by:

$$\begin{aligned} Var_g[T_j^*] &= E_g[{T_j^*}^2] - E_g[T_j^*]^2 \\ &= 2\int_t^\infty (v - t)\pi_j(v \mid t, s)\, dv - \left(\int_t^\infty \pi_j(v \mid t, s)\, dv \right)^2 \end{aligned} \tag{4}$$

It is to be noted that the posterior expected loss with the squared loss function is equal to the variance. Since the variance is same as expected loss....loss increases....intuitive??

2.2.2 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 of GR $1 - \pi_j(u \mid t, s)$ since the last biopsy. 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. Or in other words the dynamic survival probability $\pi_j(u \mid t, s)$ is below a threshold κ beyond u. In this regard, the posterior expected loss for the following multilinear loss function can be minimized to find the most optimal u:

$$L_{k_1,k_2}(T_j^*, u) = \begin{cases} k_2(T_j^* - u) & if(T_j^* > u) \\ k_1(u - T_j^*) & \text{otherwise} \end{cases} \tag{5}$$

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

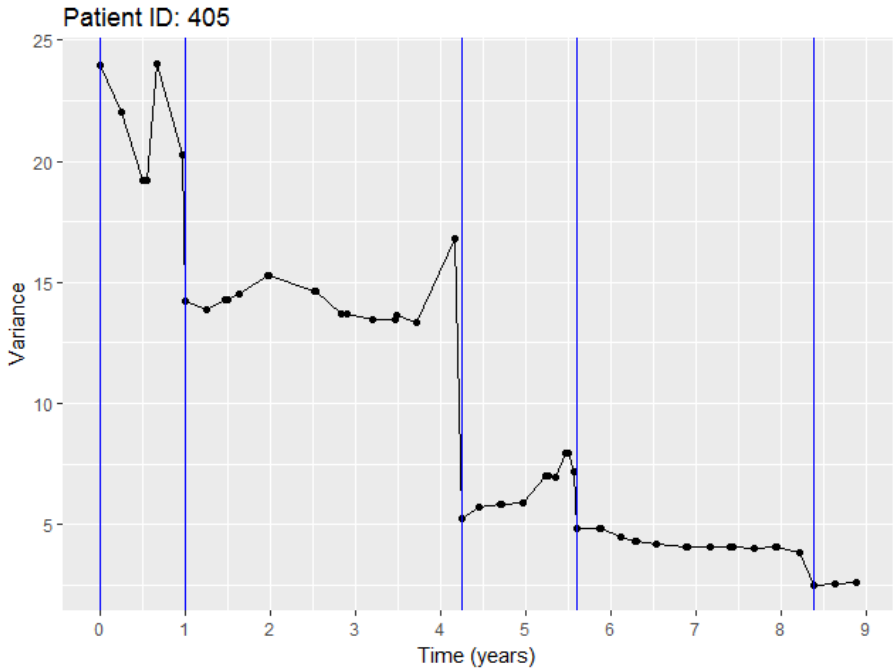


Figure 1: Variance of the predictive distribution $g(T_j^*)$ over a period of 10 years for patient 405. Blue vertical lines indicate biopsies.

Choice of κ

Since the value κ dictates the biopsy schedule, 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 50% then $1 - \kappa = 0.5$ and correspondingly, all $k_1, k_2 \mid k_1 = k_2$ can be used in equation 5 to calculate u.

While expert advice can be invaluable, it is also possible to automate the choice of κ. 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_a and in the cases at some future time point $t_b > t_a$, 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, 2014). The time window Δt can be either chosen on a clinical basis (such as 1 year in PRIAS **WHY 1 YEAR**) or it can be chosen at a point where $AUC(t, \Delta t, s)$ (Rizopoulos, 2014) is largest. i.e. Δt for which the model has the most discriminative capability at time t. The binary classification accuracy measures we maximize to select the threshold κ are the following (the binary classification measures are functions of $t, \Delta t, s$, although the notation is dropped for readability):

- Accuracy: $ACC = \dfrac{TP + TN}{TP + FP + TN + FN}$, where TP, FP, TN and FN are the number of true positives, false positives, true negatives and false negatives at time point t. In this case if $k_1 = TP + TN$ and $k_2 = FP + FN$, then $\arg\max_{k_1,k_2} ACC$ gives the optimal k_1, k_2 or equivalently the κ.
- Youden's index: $J = \text{Sensitivity} + \text{Specificity} - 1$,
  where sensitivity is defined as $Pr(\pi_j(t + \Delta t \mid t, s) \leq \kappa \mid T_j^* \epsilon (t, t + \Delta t])$ and specificity is defined as $Pr(\pi_j(t + \Delta t \mid t, s) > \kappa \mid T_j^* > t + \Delta t)$. In this case if $k_1 = FP \cdot TP - FN \cdot TN$ and $k_2 = (TP + FN)(FP + TN) - k_1$, then $\arg\max_{k_1,k_2} J$ gives the optimal k_1, k_2 or equivalently the κ.
- F1 Score: $F1 = \dfrac{2TP}{2TP + FP + FN}$. In this case if $k_1 = 2TP$ and $k_2 = FP + FN$, then

$\arg\max_{k_1,k_2} F1$ gives the optimal k_1, k_2 or equivalently the κ.

2.2.3 Mixed approach

2.2.4 Scheduling multiple biopsies

Scheduling biopsies using personalized schedules is an iterative process. Only one biopsy is scheduled at once, using all the information available up to that point in time. The information is manifested via the predictive distribution $g(T_j^*)$. It is important to note that new biopsy times are only proposed when the patient visits the hospital for a PSA measurement or for biopsy, since $g(T_j^*)$ is updated only at these time points. If a biopsy is conducted at time t, then scheduling the next biopsy at time $u \geq \max(t, s)$ for patient j is the goal. Although the predictive distribution $g(T_j^*)$ is updated with the new information $T_j^* > t$, there are a couple of restrictions on time u, namely:

1. Two consecutive biopsies should have a gap of at least 1 year between them. i.e. $u \geq t + 1$. Given the medical side effects of biopsies, the 1 year gap is strongly advised for patients enrolled in PRIAS. However, the personalized scheduling methods do not take this rule into account.

2. Although it is required that $u \geq \max(t, s)$, the personalized scheduling methods may propose to perform a biopsy at a time $u\epsilon(t, s]$. The reason is that, GR is not a terminating event, and so the patients continue to visit for PSA measurements. The support of the predictive distribution $g(T_j^*)$, however does not depend on last time of PSA measurement s, and remains (t, ∞).

Lastly, it is extremely likely that on consecutive visits to the hospital for PSA measurements, the personalized biopsy time is postponed or preponed. The choice of time of biopsy from consecutive visits is not clear. To resolve these issues, we propose to supplement the personalized scheduling methods with the following algorithm.

Definitions

Let us assume that patient j had their latest biopsy at time T_j^l, and the latest available PSA measurement is from the current visit to the hospital, at time T_j^{cv}. The predictive distribution for time of GR is given by $p(T_j^*|T_j^* > T_j^l, \mathcal{Y}_j(T_j^{cv}), \mathcal{D}_n)$. Further let,

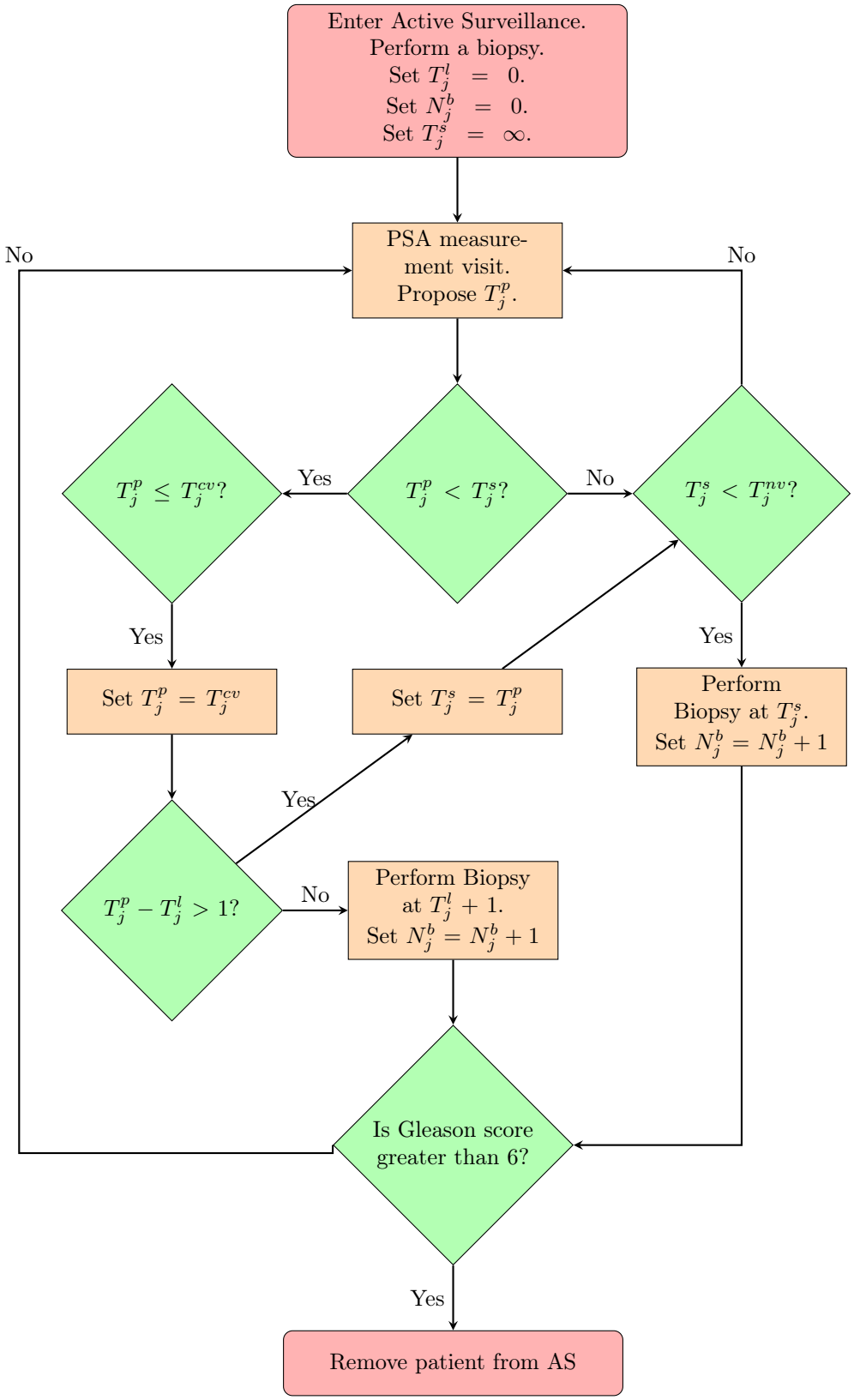
1. T_j^p denote the time at which a biopsy is proposed by personalized scheduling methods.

2. Let T_j^s denote the time at which a biopsy is scheduled. This is not necessarily equal to the proposed time T_j^p, but rather a time devoid of the problems mentioned earlier. This time is generated from the algorithm.

3. T_j^{nv} denote the time of the next visit for PSA.

4. N_j^b denote the counter for number of biopsies conducted up to time T_j^l, excluding the biopsy done at the time of induction in AS.

Algorithm



3 Personalized schedules for patients in PRIAS

To demonstrate how the personalized schedules described in section 2.2 work, we apply them to the patients in the PRIAS dataset. To this end, we divided the PRIAS data set into training(5938 subjects) and demonstration data sets (5 subjects). We fitted a joint model to the training data set and then used it to create a personalized schedule for subjects in demonstration data set. We

fit the joint model using the R package JMbayes (Rizopoulos, 2014), which uses the Bayesian methodology to estimate the model parameters.

3.1 Fitting the joint model to PRIAS dataset

The training data set contains information about 5938 prostate cancer patients who satisfied the conditions for enrollment in AS. 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 Gleason reclassification, 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 measurements we used $\log_2 PSA$ measurements instead of the raw data. The log transformation was done because the PSA scores took very large values at the onset of disease progression. This indicated that the underlying distribution for PSA scores was right skewed. The longitudinal sub-models of the joint model we fitted is given by:

$$\begin{aligned}
\log_2 PSA(t) &= m_i(t) + \varepsilon_i(t), \\
m_i(t) &= (\beta_0 + b_{i0}) + \beta_1(Age - 70) + \beta_2(Age - 70)^2 \\
&\quad + \sum_{k=1}^{4} \beta_{k+2} B_k(t, \mathcal{K}) + b_{i1} B_7(t, 0.5) + b_{i2} B_8(t, 0.5) \\
\varepsilon_i(t) &\sim N(0, \sigma^2), \\
\boldsymbol{b}_i &\sim N(0, \boldsymbol{D})
\end{aligned}$$

The evolution of PSA levels 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 $\mathcal{K} = \{0.5, 1.2, 2.5\}$ years, and boundary knots are at 0 and 7 years. For the random effects part there is only 1 internal knot at 0.5 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. The variable Age was median centered so avoid numerical instabilities while estimating the parameters in the model. For the survival sub-model the hazard function we fitted is given by:

$$h_i(t) = h_0(t) e^{\gamma_1(Age-70) + \gamma_2(Age-70)^2 \alpha_1 m_i(t) + \alpha_2 m_i'(t)}$$

where, α_1 and α_2 are measures of strength of association between hazard of Gleason reclassification and PSA value and PSA velocity, respectively. $h_0(t)$ is the baseline hazard at time t, and is modeled flexibly using P-splines(Eilers and Marx, 1996). Lastly, to fit the joint model we use the R package JMbayes Rizopoulos, 2014, which uses a Bayesian approach for parameter estimation. The parameter estimates for the survival sub-model were estimated using Bayesian ridge methodology (Andrinopoulou and Rizopoulos, 2016).

3.1.1 Parameter Estimates

The parameter estimates for the joint model we fitted to the PRIAS data set 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 over a period of 10 years for a patient who is 70 years old in Figure 2. It can be seen that the 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.

For the survival sub-model, the parameter estimates in Table 2 show that both the $\log_2 PSA$ value and $\log_2 PSA$ velocity are associated with time to Gleason reclassification. The effect is quite strong and if at any given time point the PSA becomes approximately 4 times of its value then the hazard of Gleason reclassification becomes 1.5 times of the original. This is valid under the condition that the $\log_2 PSA$ velocity remains the same. The effect of $\log_2 PSA$ velocity is far stronger, but it is not interpretable easily. The estimated values of the association parameters reinforces the idea that PSA measurements are indicative of Gleason reclassification. Lastly, for the effect of Age on hazard we can say that it can be safely ignored for all practical purposes.

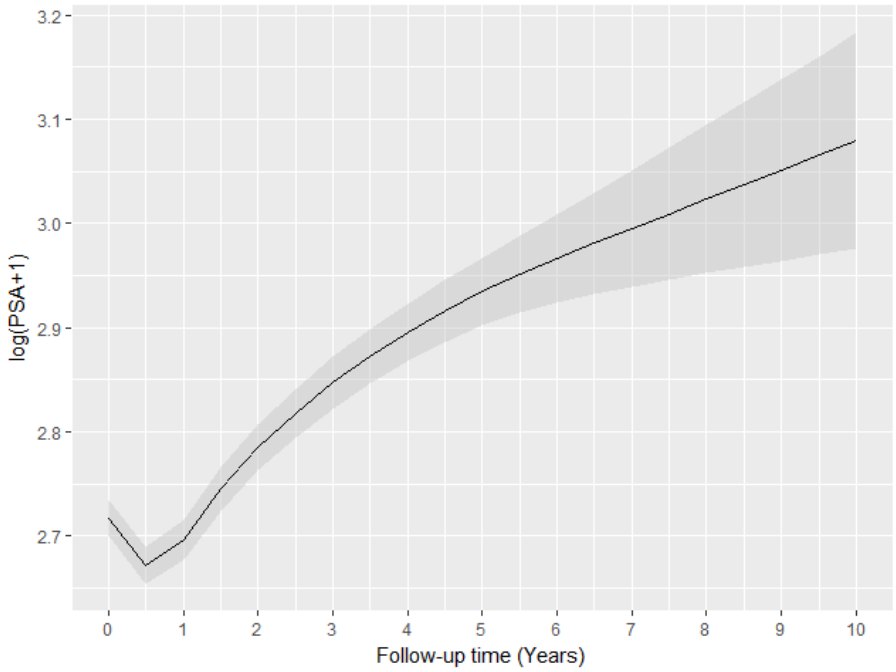


Figure 2: Fitted evolution of $\log_2 PSA$ over a period of 10 years, for a patient who was inducted in AS at the Age of 70 years.

	Mean	Std. Dev	2.5%	97.5%	P
Intercept	2.717	0.008	2.701	2.733	<0.000
(Age - 70)	0.003	0.001	0.001	0.005	0.002
(Age - 70) × (Age - 70)	-0.001	1×10^{-4}	-7×10^{-4}	-4×10^{-4}	<0.000
Spline: visitTimeYears[0, 0.5]	0.026	0.008	0.012	0.042	<0.000
Spline: visitTimeYears[0.5, 1.2]	0.208	0.013	0.184	0.233	<0.000
Spline: visitTimeYears[1.2, 2.5]	0.175	0.019	0.137	0.210	<0.000
Spline: visitTimeYears[2.5, 7]	0.309	0.028	0.256	0.366	<0.000
σ	0.273	0.001	0.271	0.275	<0.000

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

Variable	Mean	Std. Dev	2.5%	97.5%	P
Age - 70	0.036	0.007	0.023	0.050	<0.000
(Age - 70) × (Age - 70)	-0.002	0.001	-0.004	2×10^{-4}	0.016
$\log_2 PSA$	0.184	0.093	0.016	0.369	0.032
Slope: $\log_2 PSA$	1.937	0.278	1.420	2.525	<0.000

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

3.2 Demonstration of personalized schedules

The personalized schedules we propose in this work depend on this dynamic distribution for time to Gleason reclassification (Eq. 1). As more PSA measurements are taken and repeat biopsies are performed, the distribution changes updated accordingly. In this section, we demonstrate the personalized schedules based on this distribution for a set of demo patients.

It can be seen in Figure 3 that the patient 3174 had a biopsy at the time of induction and none after that. The PSA for the patient increases rapidly after the 2nd year. In response to this rapid increase, the proposed biopsy times based on conditional expected failure time also decrease accordingly from 11 years to around 3.4 years. The change for risk based methods is not so drastic

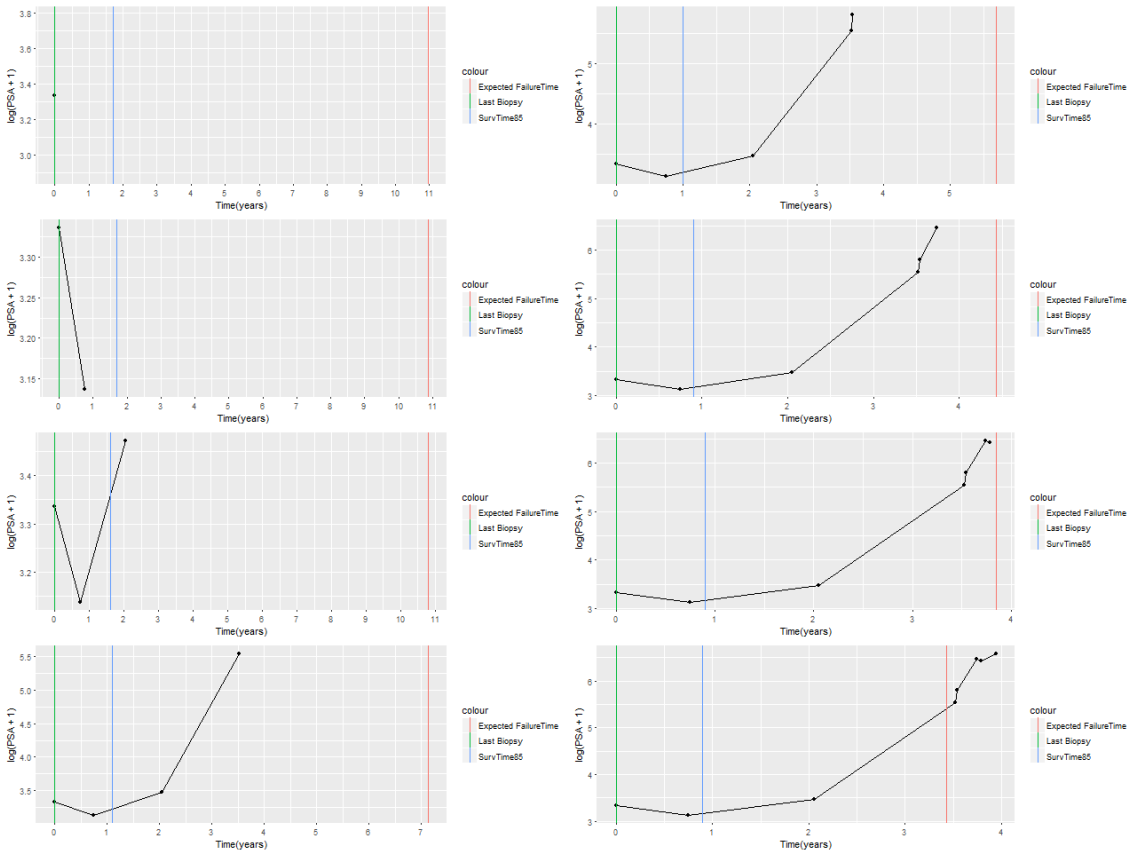


Figure 3: Proposed biopsy times for patient 3174 from PRIAS.

though. Further it can be seen that at the last visit for PSA, the proposed biopsy time is earlier than the last 5 PSA visits and similar is the case for the proposed biopsy time based on dynamic risk of failure. This is due to the fact that the last time of biopsy was time 0 (induction time) and thus the time to Gleason reclassification can take any value larger than 0. We discuss this issue in detail in section 4.1.

Figure 4 shows the PSA evolution and biopsy times for subject 911. It can be seen that this patient had 3 biopsies where Gleason reclassification did not happen. At year 2 when the patient's PSA increases rapidly the proposed failure times also decrease, whereas they increase over the next 1 year because the PSA also drops down in that time period. The fact that PSA velocity affects the biopsy times the most is also evident in the case of subject 2340, whose evolution is shown in Figure 5. Here the rate of change at each time point is not high, and even though the PSA value reaches as high as 25 it has no effect on proposed biopsy times. This is in accordance with the estimated strength of association between PSA velocity/value and hazard of time to Gleason reclassification.

An interesting observation we made while creating these schedules was that the variance of time to Gleason reclassification (Eq. **??**) was quite high, which essentially rules out the usefulness of conditional expected time to Gleason reclassification. Given the large difference in proposed biopsy times based on the former and methods based on dynamic risk of Gleason reclassification, one might conclude that the latter are more useful. However as we will see in the simulation study (Section 4) ahead, the usefulness of the two categories of methods depends on the distribution of time to Gleason reclassification.

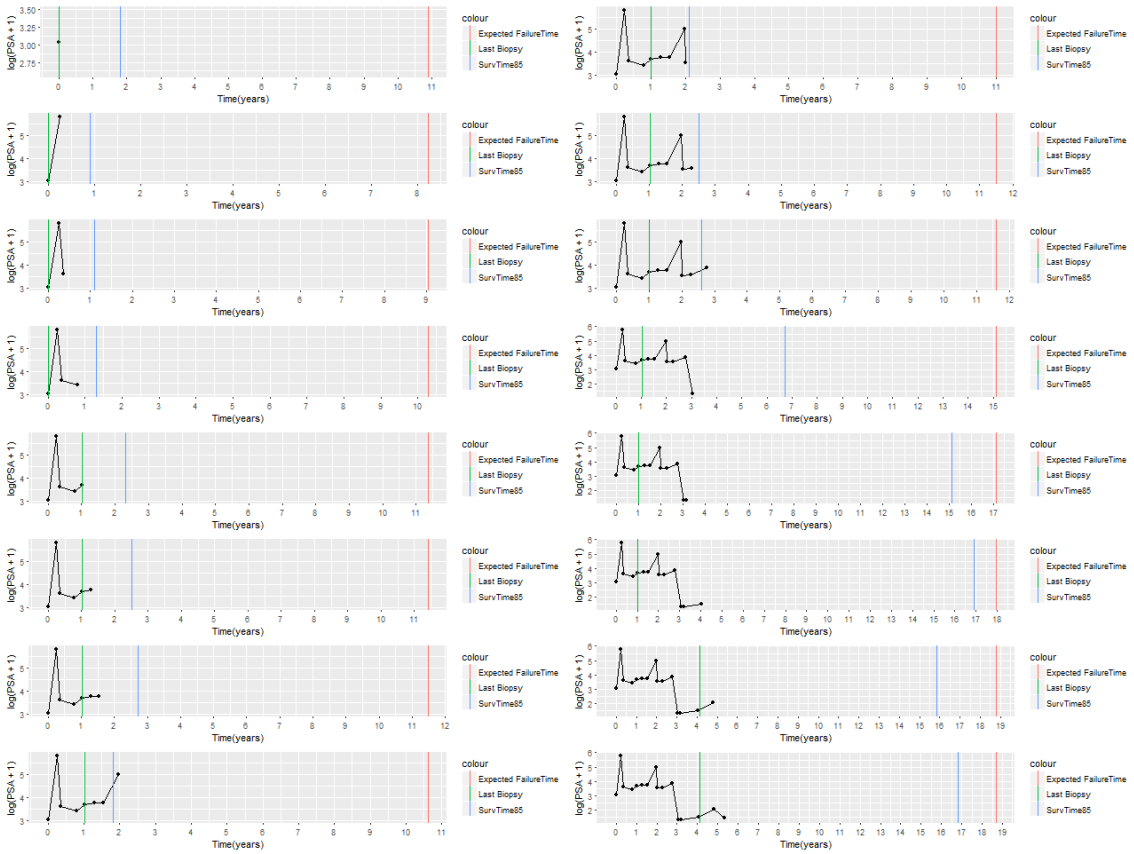


Figure 4: Proposed biopsy times for patient 3174 from PRIAS.

4 Simulation study

For contrasting the personalized scheduling approaches against the schedule of PRIAS and the schedule of doing biopsy every year, we performed a simulation study. We used the parameters from the joint model fitted to the original PRIAS data set to generate 60 data sets with 1000 patients each. We employed a Weibull baseline hazard while generating Gleason reclassification times for the simulated patients. We divided the 60 data sets into groups of 10 each, and within each group the parameter values for the Weibull baseline hazard were kept the same, while across the group they were different. The reason to create such groups was to ensure the consistency of results within in each group. Using different parameter values for the baseline hazard across the groups ensured that we tested not only the scenarios in which patients have early progression times but also those in which patients have late progression times. This is important because how late/early patients are observed to have Gleason reclassification, partially depends on the induction criteria of the AS program.

4.1 Simulation setup

We divided each of the 60 data sets into training (750 patients) and test (250 patients) parts. We fitted a joint model to the training data set and using it we generated biopsy schedules for patients in the test data set.

Evaluating performance of personalized biopsy schedules

Since we generate the true Gleason reclassification time for each of the patients in the test data set, we are able to calculate the offset for each patient and each scheduling method. While creating the personalized schedules we adhere to the PSA measurement schedule of PRIAS.

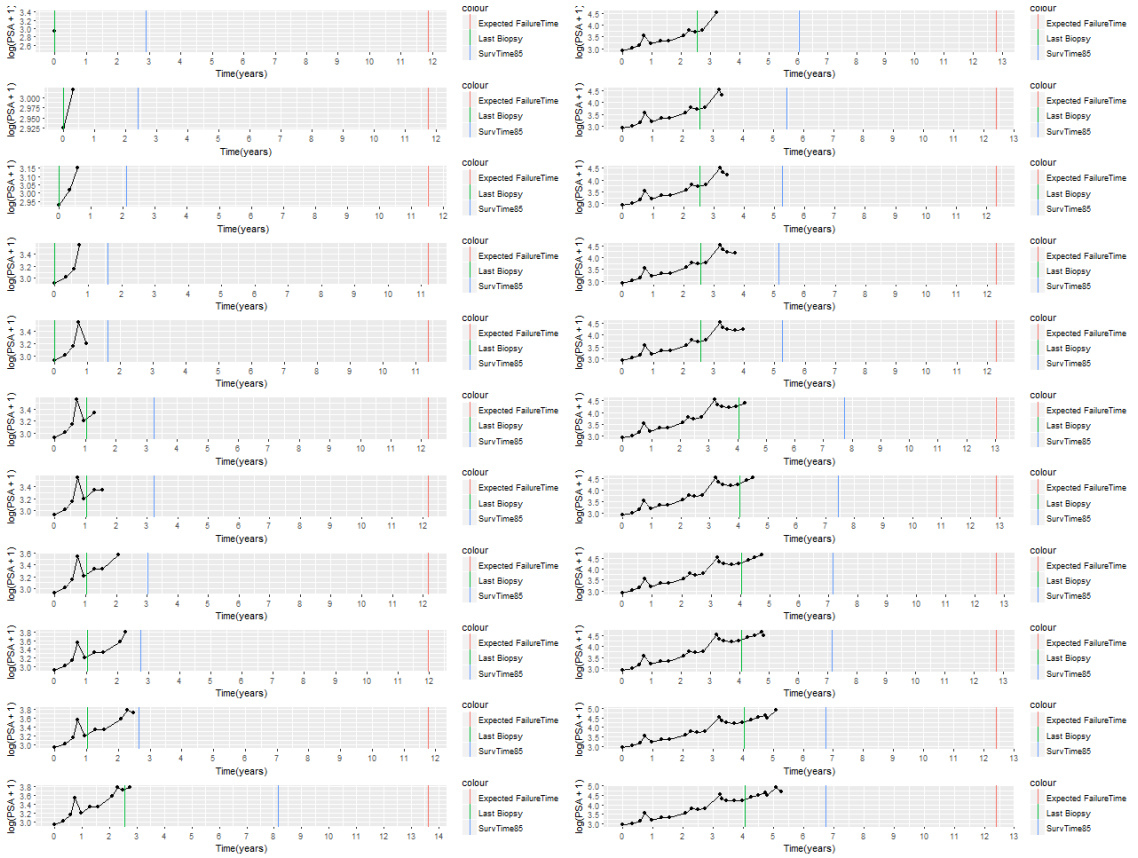


Figure 5: Proposed biopsy times for patient 3174 from PRIAS.

4.2 Results

Since the Gleason reclassification times varied widely across the simulation data sets, we discuss the results separately for the different scenarios. In total we have 6 scenarios corresponding to the 6 groups of 10 data sets. For brevity, we only chose 3 most dissimilar scenarios and further we present one set of results for each of the 3 scenarios. The complete set of results can be found at https://goo.gl/u6dg8G. At this moment, we will not compare our results against the PRIAS schedule, since the PRIAS schedule we used is fixed for all patients, whereas in reality it can be dynamic if the PSA doubling time is less than 10 years.

4.2.1 Scenario 1

In the first scenario the Gleason reclassification times for the patients are mostly between 7 and 11 years. The Gleason reclassification times of test patients from one of the data sets is shown in Figure 6. Figure 9 shows that, for this data set the approaches with the least number of biopsies, 2 in number, are conditional expected failure time, and dynamic risk with a threshold κ chosen such that accuracy (section 2.2.2) is maximized. The mean and median offset for the two approaches is around 9 months. However for a few outlying patients the offset is more than 24 months. If we compare this to the schedule of performing biopsy every year the mean offset is around 6 months and mean number of biopsies is 8.

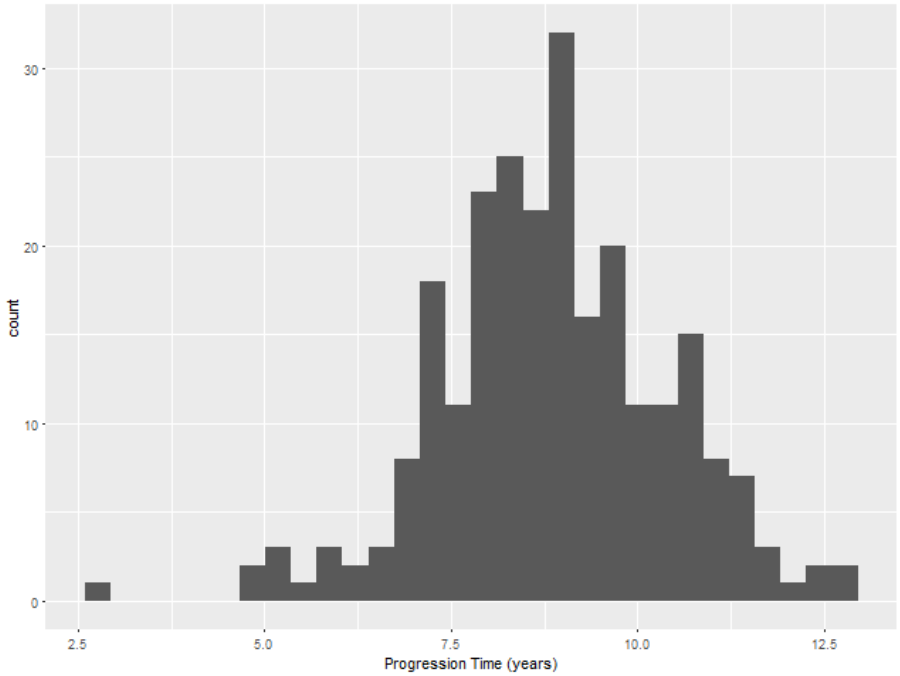


Figure 6: Gleason reclassification times for patients from the test data set of scenario 1.

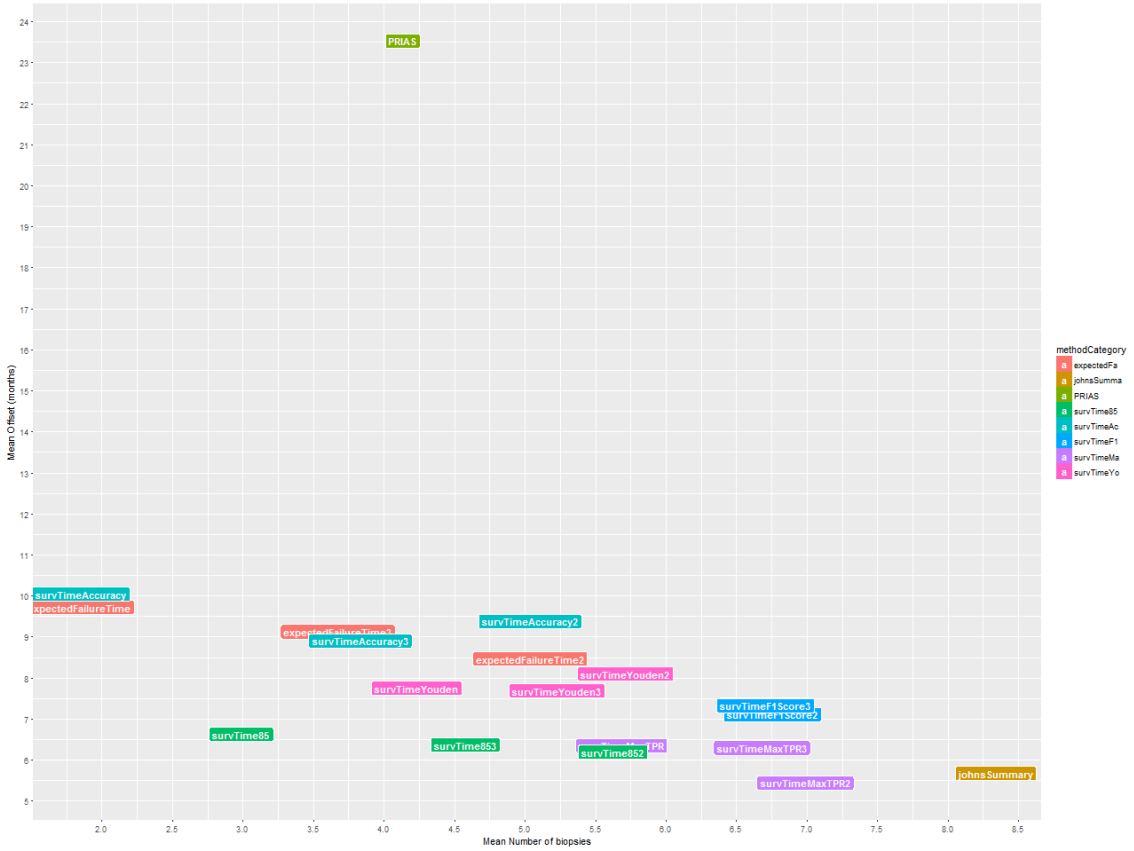


Figure 7: Mean number of biopsies against the mean offset (in years) for each of the approaches in scenario 1.

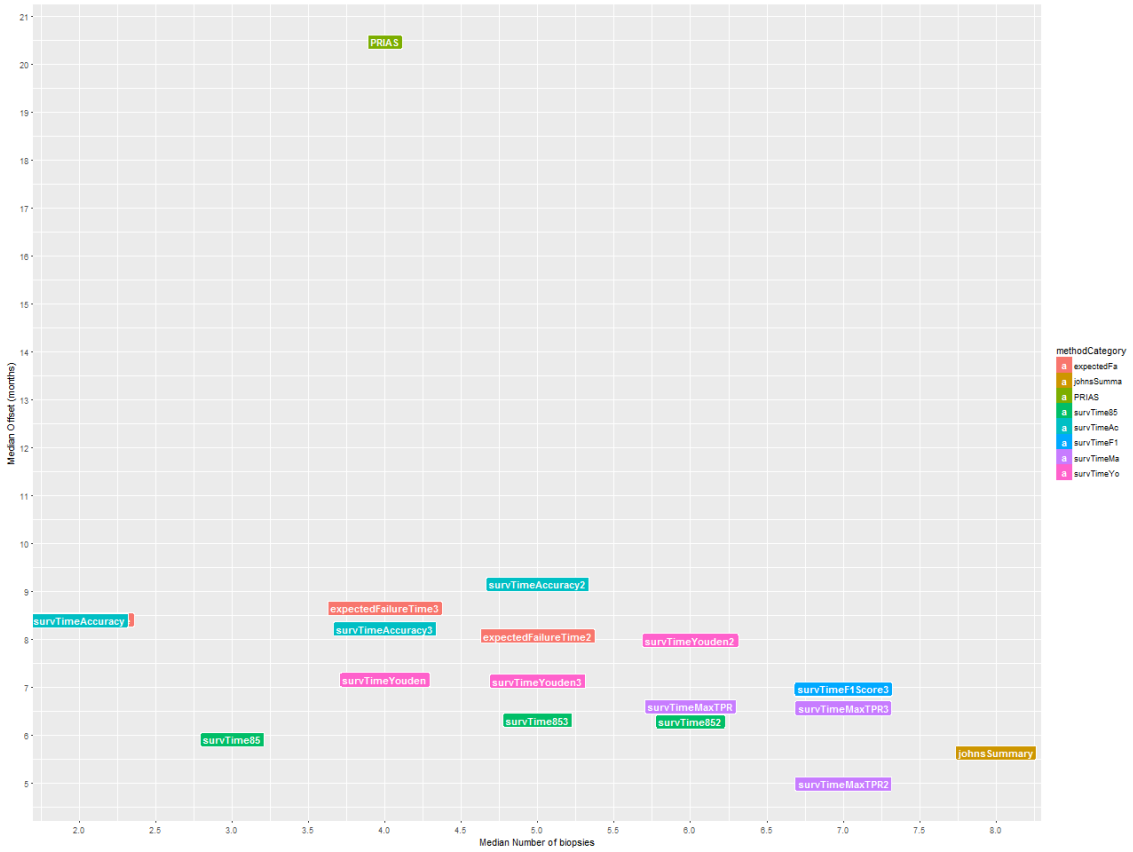


Figure 8: Median number of biopsies against the median offset (in years) for each of the approaches in scenario 1.

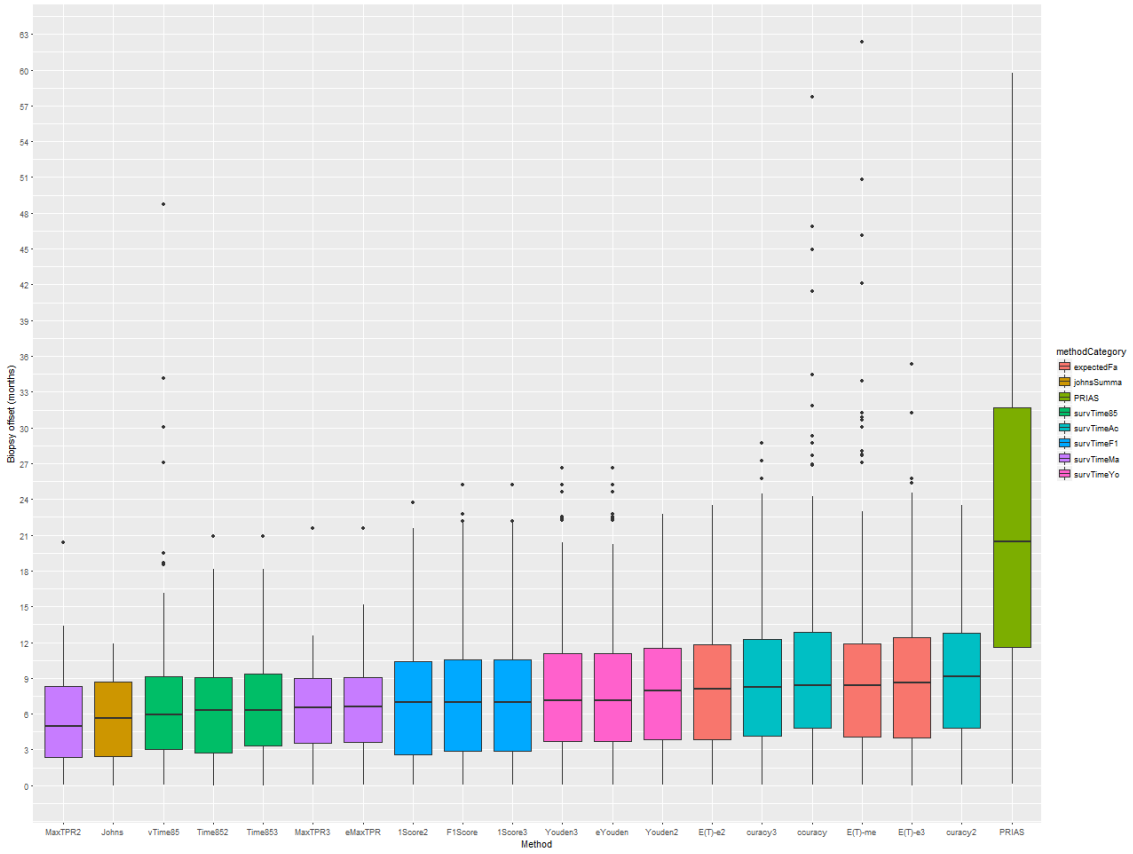


Figure 9: Boxplot for the offset corresponding to the various approaches in scenario 1.

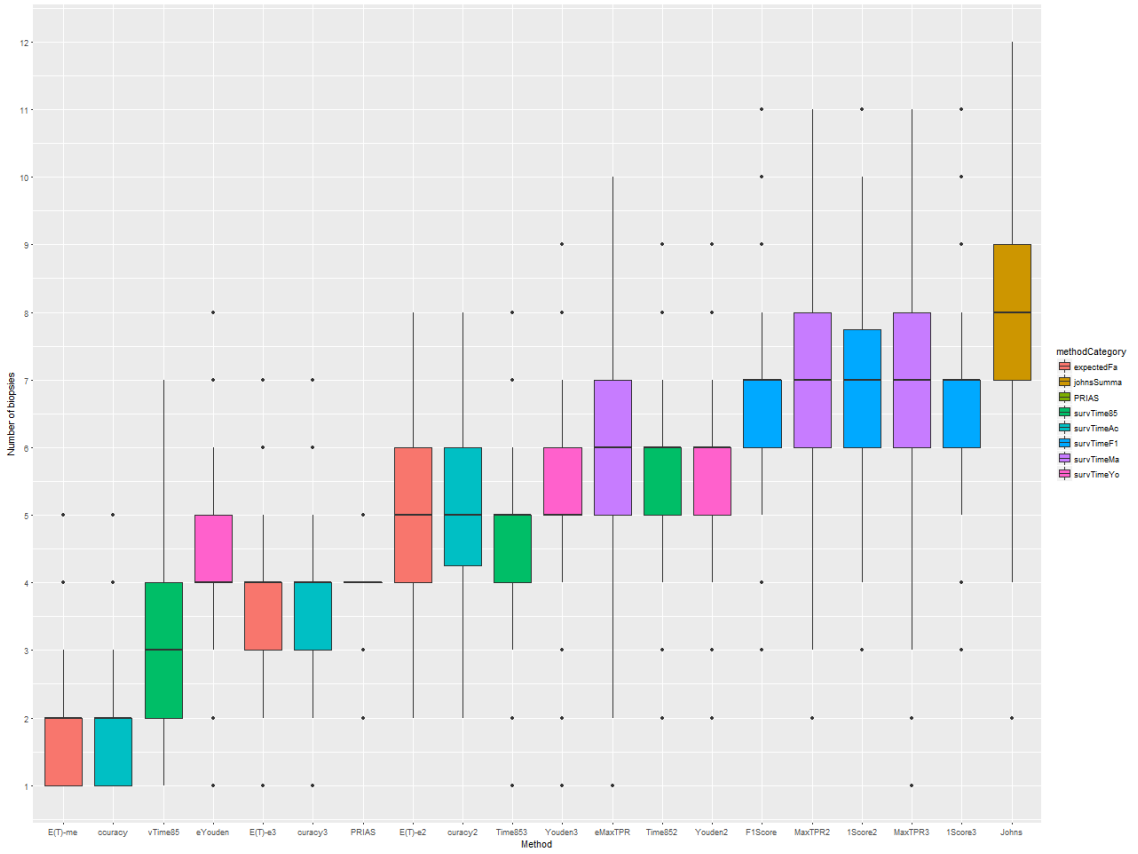


Figure 10: Boxplot for the number of biopsies corresponding to the various approaches in scenario 1.

4.2.2 Scenario 2

Unlike the previous scenario, the Gleason reclassification times for the patients in Scenario 2 are mostly between 0 and 5 years. The Gleason reclassification times of test patients from one of the data sets is shown in Figure 11. Figure 14 shows that, for this data set the approach with the least number of biopsies is conditional expected failure time. The mean and median number of biopsies are 1.5 and 1 respectively. While the mean and median offset is around 18 months, there is considerable variation in the offset for the various patients. The first and third quartiles are at 12 months and 27 months respectively. The large variation in offset is due to the fact that time to Gleason reclassification has larger variance since it is based on a small history of the patient. In such a scenario one might be inclined towards doing a biopsy every year as they have the least offset. However that approach leads to a very large number of biopsies as shown in Figure 15. Certain approaches based on dynamic risk perform better here, both in terms of number of biopsies as well as the offset. In particular if the choice of κ can be based on maximization of the Youden index or a fixed κ can be chosen.

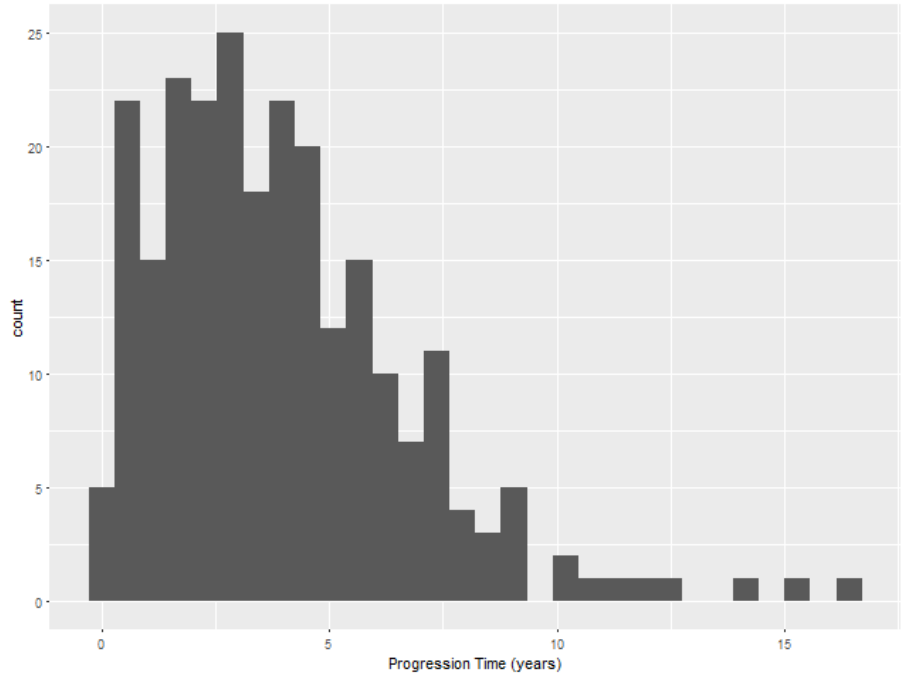


Figure 11: Gleason reclassification times for patients from the test data set of scenario 2.

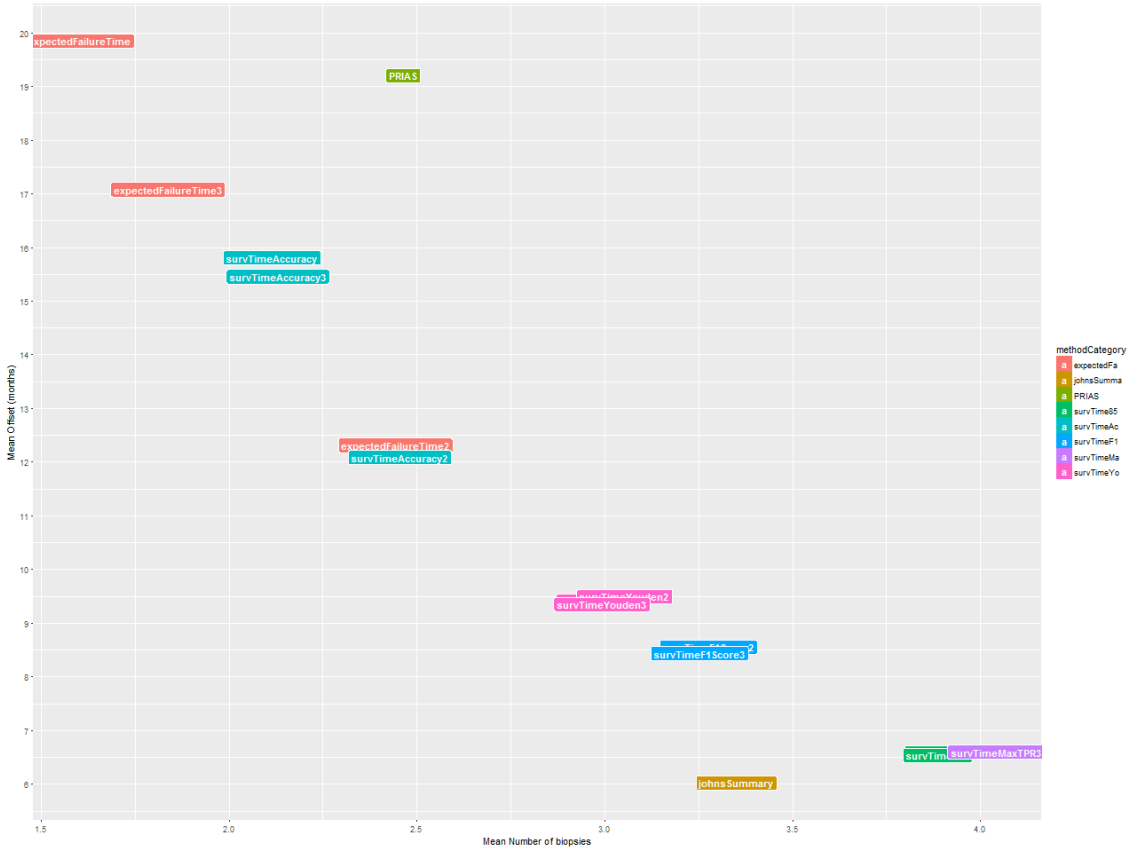


Figure 12: Mean number of biopsies against the mean offset (in years) for each of the approaches in scenario 2.

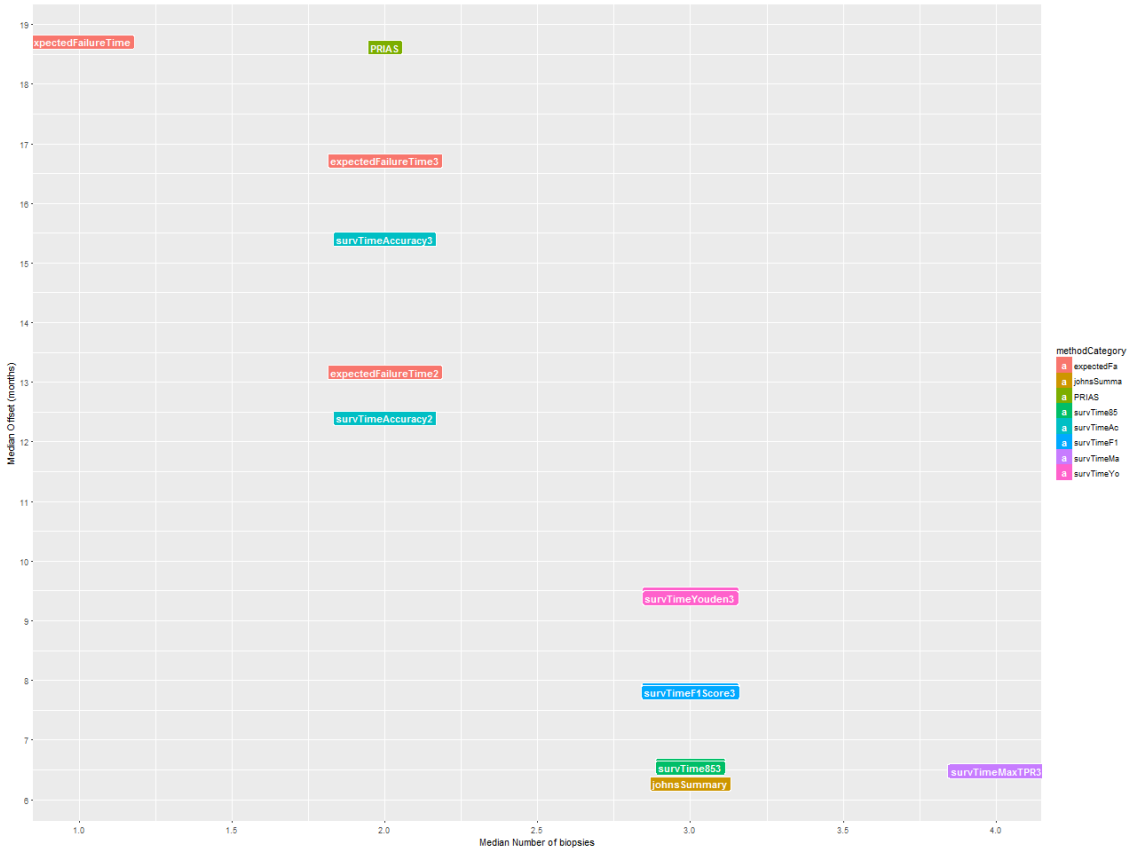


Figure 13: Median number of biopsies against the median offset (in years) for each of the approaches in scenario 2.

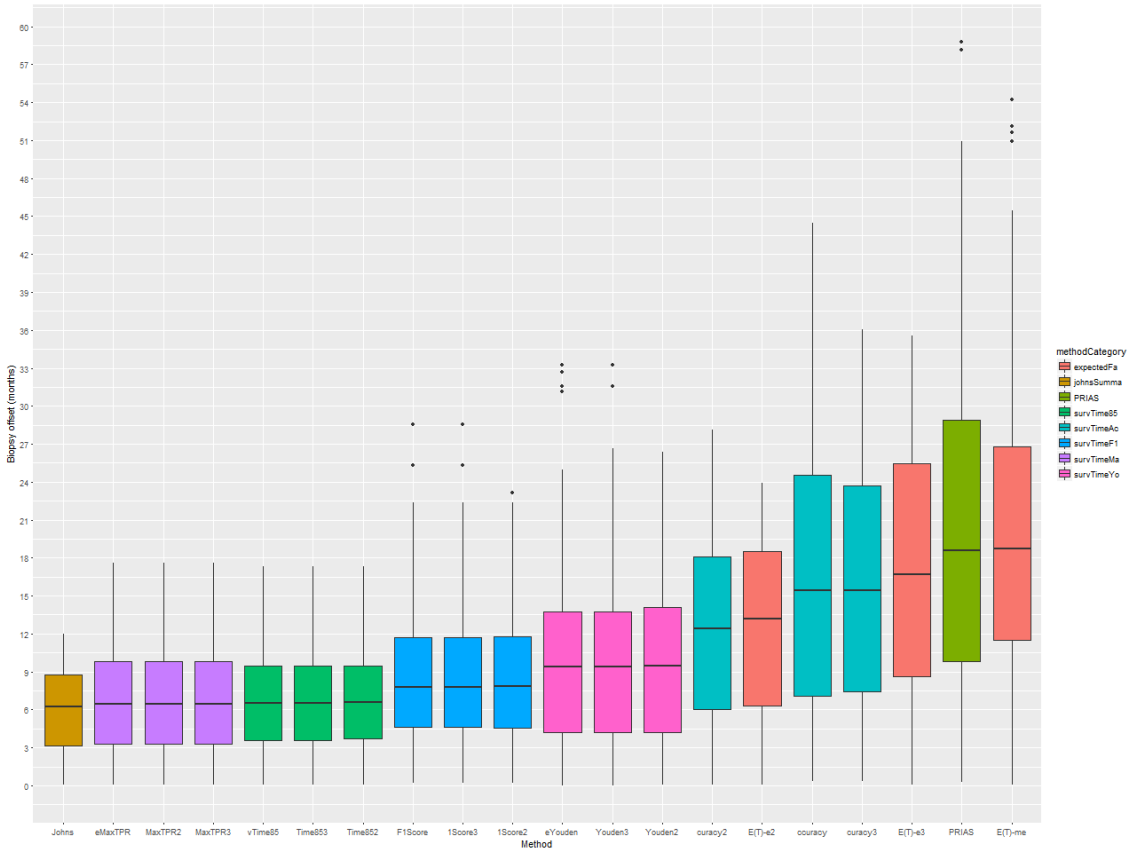


Figure 14: Boxplot for the offset corresponding to the various approaches in scenario 2.

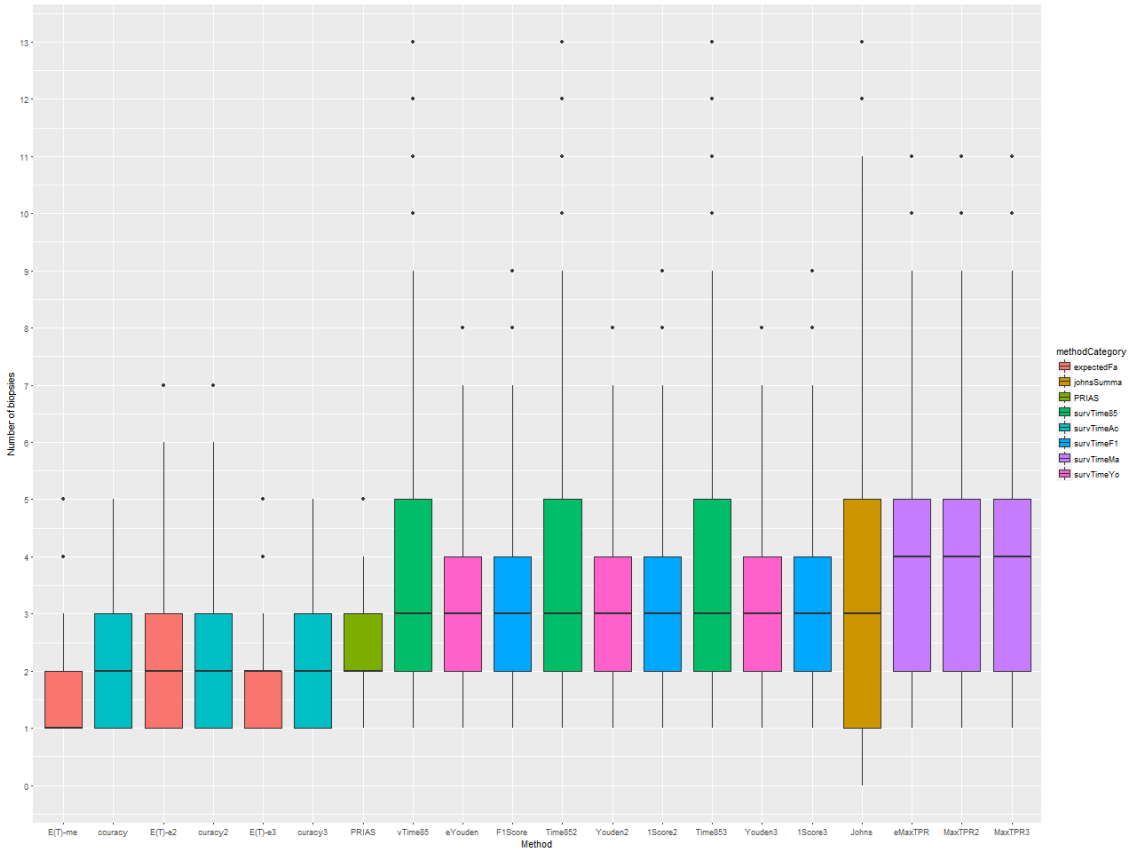


Figure 15: Boxplot for the number of biopsies corresponding to the various approaches in scenario 2.

4.2.3 Scenario 3

In scenario 3, the Gleason reclassification times for the patients are widely spread, mostly between 2 and 9 years. i.e. patients fail late as well early. The Gleason reclassification times of test patients from one of the data sets is shown in Figure 16. Figure 19 shows that, for this data set the approach with the least number of biopsies is conditional expected failure time. The mean and median number of biopsies are 1.5 and 1 respectively. This method also has considerable variation in the offset for the various patients. The first and third quartiles are at 9 months and 27 months respectively. A more detailed analysis revealed that the variation in offset was higher for subjects with earlier failure times. For e.g. the mean and median offset in subjects with reclassification times more than 4 years was nearly 13.5 months. For others it was nearly 39.5 months. This is once again is due to the fact that time to Gleason reclassification has larger variance since it is based on a small history of the patient. And thus usefulness of conditional expected failure time is questionable.

Dynamic risk based methods once again provide an alternative. If a fixed κ of 0.15 is used then although the offset will be low, there is very high variation in number of biopsies. i.e. people who have reclassifications early will benefit but people who have reclassifications later will have too many biopsies. If κ is chosen such that accuracy is maximized, and a biopsy is done whenever there is a gap of 2 years then a reasonable offset is obtained. The first and third quartiles are almost 4.5 and 18 months. The first and third quartiles for number of biopsies are 3 and 4.

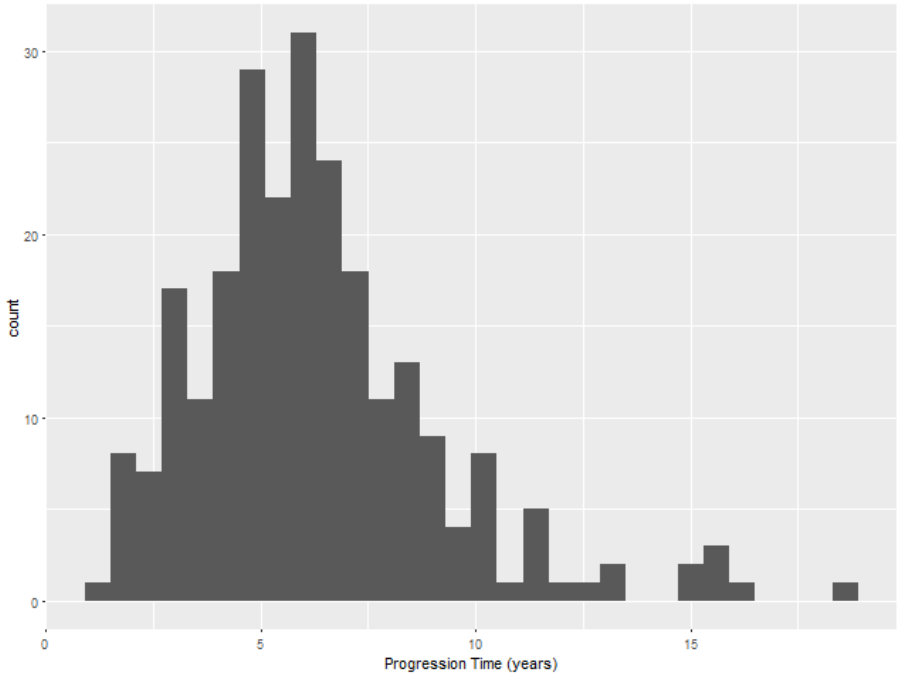


Figure 16: Gleason reclassification times for patients from the test data set of scenario 3.

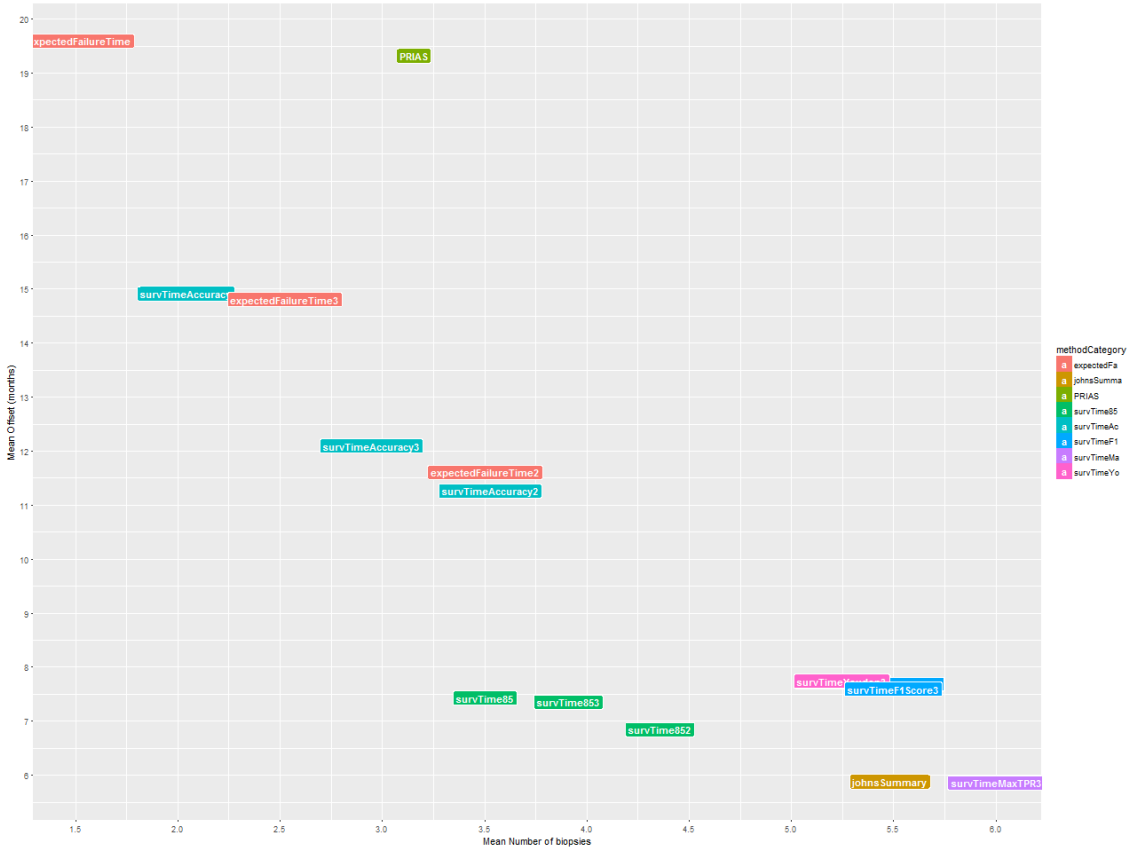


Figure 17: Mean number of biopsies against the mean offset (in years) for each of the approaches in scenario 3.

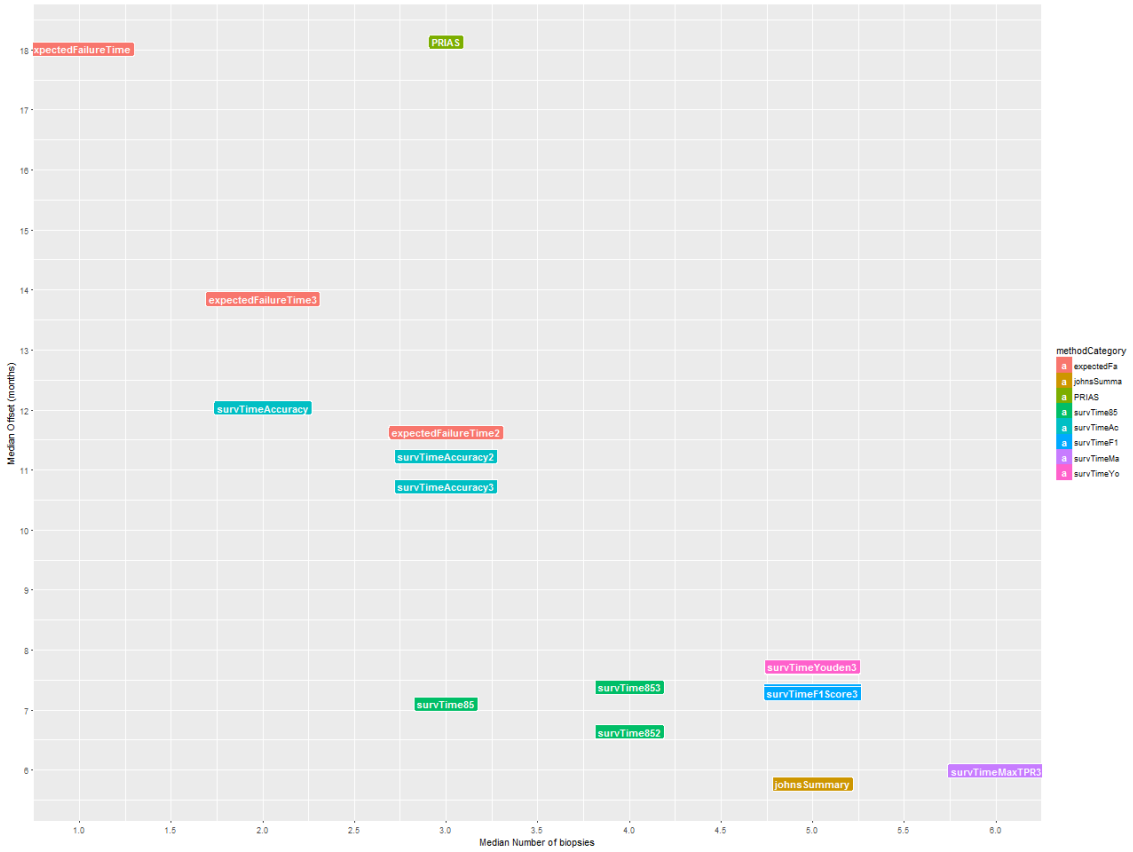


Figure 18: Median number of biopsies against the median offset (in years) for each of the approaches in scenario 3.

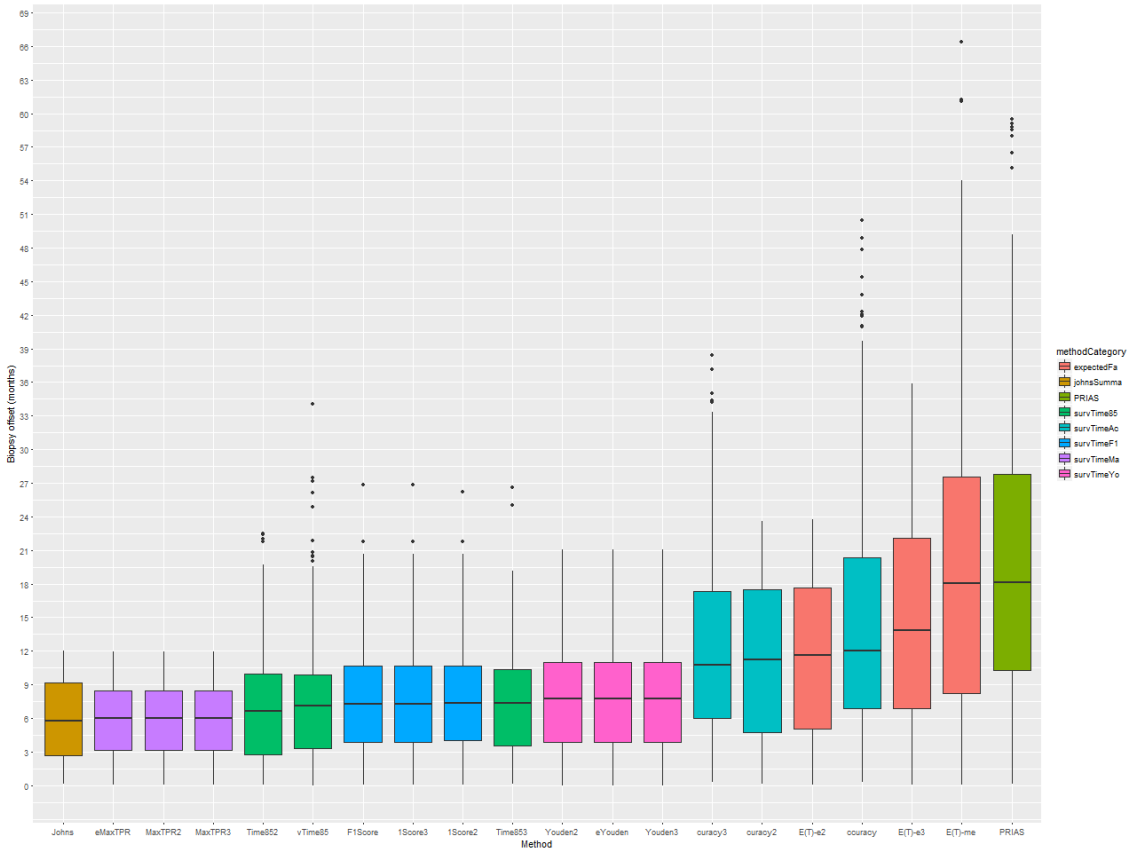


Figure 19: Boxplot for the offset corresponding to the various approaches in scenario 3.

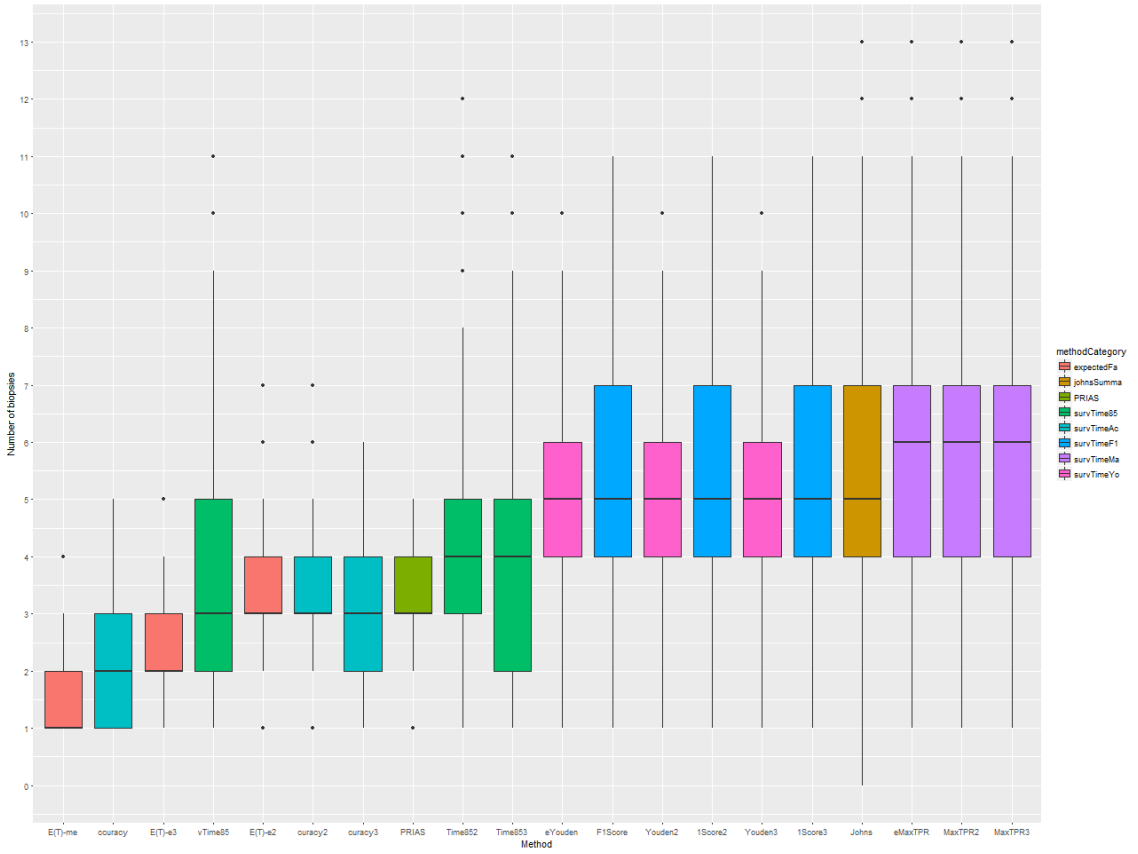


Figure 20: Boxplot for the number of biopsies corresponding to the various approaches in scenario 3.

4.3 A mixed approach: E(T) vs dynamic risk

While we did create biopsy schedules using conditional expected Gleason reclassification time, and dynamic risk of Gleason reclassification based approaches, we saw that no single approach works best in all scenarios. In particular, we observed that when the failure times are early dynamic risk based approaches perform better in terms of offset than conditional expected Gleason reclassification time. The latter works best when the Gleason reclassification times are later. This has motivated us to use a mixed approach, where we schedule biopsies using dynamic risk based approaches if the variance (Eq. **??**) of time to Gleason reclassification is higher than a certain threshold, otherwise we use conditional expected Gleason reclassification time. This is still work in progress.

References

Andrinopoulou, Eleni-Rosalina and Dimitris Rizopoulos (2016). "Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures". In: *Statistics in Medicine* 35.26, pp. 4813–4823.

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.

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.

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.

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.

— (2014). "The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC". In: *arXiv preprint arXiv:1404.7625*.

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. eprint: /oup/backfile/content_public/journal/biostatistics/17/1/10.1093_biostatistics_kxv031/3/kxv031.pdf.

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

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

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 Heading

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse accumsan magna est, quis elementum leo laoreet eu. Donec sollicitudin elit non massa venenatis, in viverra dolor sagittis. Maecenas ac justo pulvinar, consectetur mauris hendrerit, vulputate lacus. Etiam tristique sapien quis sem commodo, et eleifend tortor viverra. In hac habitasse platea dictumst. Phasellus vel tempus risus, sit amet consectetur massa. Duis rutrum lectus eu ligula egestas iaculis. Sed condi-

mentum, ipsum in dignissim condimentum, nisi turpis blandit massa, et aliquam magna ligula eget lacus. Donec ac eleifend nulla, quis cursus nisi. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse accumsan magna est, quis elementum leo laoreet eu. Donec sollicitudin elit non massa venenatis, in viverra dolor sagittis. Maecenas ac justo pulvinar, consectetur mauris hendrerit, vulputate lacus. Etiam tristique sapien quis sem commodo, et eleifend tortor viverra. In hac habitasse platea dictumst. Phasellus vel tempus risus, sit amet consectetur massa. Duis rutrum lectus eu ligula egestas iaculis. Sed condimentum, ipsum in dignissim condimentum, nisi turpis blandit massa, et aliquam magna ligula eget lacus. Donec ac eleifend nulla, quis cursus nisi.