Department of Biostatistics
Erasmus University Medical Center
PO Box 2040, 3000 CA Rotterdam
the Netherlands

August 15, 2021

Professor Els Goetghebeur
Department of Applied Mathematics, Computer Science and Statistics
Universiteit Gent
Krijgslaan 281, 9000 Gent
Belgium

Dear Professor Goetghebeur,

We are writing to you with respect to the manuscript #SIM-21-0304, titled "Shared Decision Making of Burdensome Surveillance Tests Using Personalized Schedules and Their Burden and Benefit" submitted to *Statistics in Medicine* and the reports we received after its review. We would like to thank you for giving us the opportunity to submit a revised version of our paper that tackles the weaknesses of the previous version.

Following the recommendations from the Reviewers, we have made changes in the revised version of the manuscript. In particular, we have added explanations for the model equations to make them accessible to a wider audience. We have also discussed the impact of the increasing use of MRI on the findings reported in our paper. The "clean" version of the manuscript is titled "main_manuscript.pdf", the manuscript with highlighted changes is titled "edited_manuscript.pdf", and the appendices are compiled into a file titled "supplementary.pdf".

Please find enclosed a detailed point-by-point response to the Reviewers' comments.

Yours sincerely,

the Authors

Response to 1st Referee's Comments

We would like to thank the Referee for their constructive comments, which have allowed us to considerably improve our paper. The main differences of the new version of the manuscript compared to the previous one can be found in the section titled "Study Population".

You may find below our responses to the specific issues raised.

1. **One major concern centers on the true time of progression T_j^*, which is only known in simulation but unknown in real applications. As on P20, it says that "Due to the periodical nature of schedules, the actual time delay in detecting progression cannot be observed in real-world surveillance". Hence, it is only known that T_j^* is located between two scheduled visits (i.e., subject to interval censoring) and its exact time is unknown. In this case, how to calculate the expected time delay in detecting progression?**

   We thank the referee for raising this point. To answer this question, we first denote a schedule $S_j = \{s_1, \ldots, s_{N_j}\}$ that subject j will follow. For subjects who have not progressed the delay calculation we presented in original manuscript on page-15 was:

   > ..., the random variable time delay is equal to the difference between the time of the test at which progression is observed and the true time of progression T_j^*, and is given by,
   >
   > $$\mathcal{D}_j(S_j^\kappa) = \begin{cases} s_1 - T_j^*, & \text{if } t < T_j^* \leq s_1, \\ s_2 - T_j^*, & \text{if } s_1 < T_j^* \leq s_2, \\ \vdots & \\ s_{N_j} - T_j^*, & \text{if } s_{N_j-1} < T_j^* \leq s_{N_j}, \end{cases}$$
   >
   > The expected time delay in detecting progression is the expected value of $\mathcal{D}_j(S_j^\kappa)$, given by the expression,
   >
   > $$E\{\mathcal{D}_j(S_j^\kappa)\} = \sum_{n=1}^{N_j} \Big\{ s_n - E(T_j^* \mid s_{n-1}, s_n, v) \Big\} \times \Pr(s_{n-1} < T_j^* \leq s_n \mid T_j^* \leq s_N),$$
   >
   > where $E(T_j^* \mid s_{n-1}, s_n, v)$ denotes the conditional expected time of progression for the scenario $s_{n-1} < T_j^* \leq s_n$ and is calculated as the area under the corresponding

survival curve,

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

However, instead of subjects who have not progressed, the referee asked "how to calculate the expected time delay in detecting progression when it is only known that T_j^* is located between two scheduled visits (i.e., subject to interval censoring) and its exact time is unknown". In mathematical notation the question is "what is $E\big\{\mathcal{D}_j(S_j)\big\}$ if it is known that $s_{n-1} < T_j^* \leq s_n$"? Using the delay equation from the original manuscript (in the shaded area above) and dropping the $\sum_{n=1}^{N_j}$ term from that equation, the delay in this specific situation is given by:

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

Here, $\Pr(s_{n-1} < T_j^* \leq s_n \mid T_j^* \leq s_N) = 1$ because we know that $s_{n-1} < T_j^* \leq s_n$. Thus,

$$E\big\{\mathcal{D}_j(S_j)\big\} = s_n - E(T_j^* \mid s_{n-1}, s_n, v), \tag{1}$$

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

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

On page 20, to compare schedules, the reason we did not use the estimated delay in (1) for the real subjects was that real subjects cannot undergo multiple schedules again and again. Nor can we estimate different delays for different schedules. Hence we instead conducted a realistic simulation study, where we were able to try different schedules per subject. In addition, in the simulation setting we also knew the simulated times of progression T_j^* of subjects. Consequently, when a simulated subject progressed between two consecutive visits, i.e., $s_{n-1} < T_j^* \leq s_n$, we knew that the true delay was $D_j = s_n - T_j^*$. We had no need to estimate it as per (1). Although, in a real world situation the formula in (1) can be used retrospectively, to estimate the delay that occurred knowing that the subject progressed in the interval $s_{n-1} < T_j^* \leq s_n$.

2. **The time-dependent AUC in Supplementary Table 8 is quite low (i.e., between 0.61 and 0.68), suggesting the poor predictive performance of either the model**

or the data or both. In addition, the time-dependent mean absolute prediction error (MAPE) was moderate to large, further indicating the insufficiency of the model. More detailed investigation of model fitting and selection is warranted.

We thank the referee for giving us the opportunity to provide our motivation for the choice of the model. The referee has correctly noted that the overall predictive performance of the model can be improved. This performance pertains to predicting the time of cancer progression. During the development of the methodology we had identified three areas of improvement for our model. First, the current model assumes that cancer progression can be measured perfectly whereas it is prone to inter-observer variation. Quoting from the discussion section of our manuscript:

> ... the detection of progression is susceptible to inter-observer variation, e.g., pathologists may grade the same biopsy differently. Progression is sometimes kobscured due to sampling error, e.g., biopsy results vary based on location and number of biopsy cores. Although models that account for inter-observer variation (Balasubramanian and Lagakos, 2003) and sampling error (Coley et al., 2017) will provide better risk estimates, the methodology for obtained personalized schedules can remain the same. ...

Second, we overestimate the risk of disease progression by considering all events other than progression as non-informative censoring. Quoting from the discussion section of our manuscript:

> ... The proposed joint model assumed all events other than progression to be non-informative censoring, and consequently the cumulative-risk of progression is overestimated. Better estimates may be obtained by using models that account for competing risks. ...

Third, our time-to-event sub-model consists of only three predictors, namely patient-specific instantaneous prostate-specific antigen (PSA) value, instantaneous PSA velocity, and instantaneous log odds of obtaining a digital rectal examination (DRE) measurement larger than T1c. Indeed one can also add area under the PSA, and/or velocity of the log odds of DRE being larger than T1c, and/or the random-effects from the longitudinal sub-models, and/or lagged effects of the aforementioned predictors. Our choice of predictors, though, was mainly driven by clinical literature and the PRIAS dataset's study protocol. More specifically, the

PRIAS protocol uses PSA-doubling time (inverse of the slope of the regression line through observed PSA), last observed PSA and DRE as predictors of progression. Our three predictors improve upon these predictors because we do not rely only on last observed values, but rather use all observed data to obtain a fitted instantaneous value and velocity.

Indeed we could have developed a more sophisticated model that would have covered all three aforementioned areas of improvement, but the methodology we proposed would not have changed at all, as it only relies on risk predictions. Such risk predictions may be obtained from any other model as well. Although, with a better model we would have had the advantage to argue stronger in favor of personalized schedules, but that could have had also partially obscured the fact that personalized schedules are only as good as the data and the model. Thus, despite in much agreement with the referee about the model's performance, we respectfully argue that the purpose of the model fitted to the PRIAS study's dataset is only illustration and not supporting direct use of the model for PRIAS subjects. Rather any model developed for PRIAS subjects must be externally validated (not internally like us) and patient considerations must be taken into account before employing personalized schedule based tests for them. Enabling this shared-decision making has been the motivation and novelty of our methodology. Specifically, for any schedule, whether it is model based or not, we are able to provide the consequences of following the schedule in terms of personalized expected number of tests (burden) and personalized expected delay in detecting progression (benefit).

We have now updated the discussion section of the manuscript and have mentioned that the model still needs improvements and external validation. It currently reads as:

> ... While based on these arguments we propose the use of joint models for predicting risks, the methodology in Section 3 can be used with any other model that provides risk estimates for progression. ...
>
> ...
>
> ... The simulation study results are by no means the performance-limit of the personalized schedules. Instead, models with higher predictive accuracy and discrimination capacity than the PRIAS based model may lead to an even better balance between the number of tests and the time delay in detecting progression. As for the practical usability of the PRIAS based model in prostate cancer surveillance, the model still needs external validation and improvements in its predictive perfor-

mance. Despite that, we expect this model's overall impact to be positive. There are two reasons for this. First, the risk of adverse outcomes because of personalized schedules is quite low because of the low rate of metastases and prostate cancer specific mortality in prostate cancer patients (Bokhorst et al., 2015). Second, studies (Carvalho, Heijnsdijk, and Koning, 2017; Inoue et al., 2018) have suggested that after the confirmatory biopsy at year one of follow-up, biopsies may be done as infrequently as every two to three years, with limited adverse consequences. In other words, longer delays in detecting progression may be acceptable after the first negative biopsy. ...

3. **The functional form $f_k(\cdot)$ on P9 does not reflect the multivariate nature of the longitudinal outcomes.**

The referee has correctly noted that the functional form $f_k(\cdot)$ on page 9 does not reflect the multivariate nature of the longitudinal outcomes. However, this was done on purpose to keep the notation simple. Although we do notify the readers in the original manuscript that "(subscripts k dropped for brevity)". The corresponding text in question from the page 9 of the original manuscript is below.

...Some examples, motivated by the literature (subscripts k dropped for brevity), are,

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

4. **Figure 2 is presented without explanation in the main text, besides what appears in the caption. Panel C has no mention even in the caption. It is unclear why it is presented and what message it delivers, if any.**

We thank the Referee for motivating us to improve the explanation for Figure 2. The purpose of Figure 2 was to illustrate to the readers that the cumulative-risk of progression updates over time automatically with new clinical data. This is important because it allows us to develop schedules and estimate their burden (number of tests required) and benefit (time delay in detecting progression) is such a manner that they also update automatically with new data over time. To overcome the shortcomings of our earlier explanation, we have a now

a more comprehensive explanation and figure caption in Section 3.1 of the new manuscript. It is as follows:

> ... A key property of this cumulative-risk function is that it is time-dynamic (illustrated in Figure 2). That is, it automatically updates over time as more longitudinal and invasive test result data becomes available. We next exploit this property to first develop schedules that are also personalized and time-dynamic, and subsequently to estimate the burden (number of tests required) and benefit (time delay in detecting progression) of the resulting schedules in a time-dynamic manner.
>
> .....Figure 2 here (caption below) .....
>
> **Figure 2 The cumulative-risk function (1) is time-dynamic** because it automatically updates over time as more longitudinal and invasive test result data becomes available. We illustrate this using a single longitudinal outcome, namely, a continuous biomarker of disease progression (All values are illustrative). **Panels A, B and C** are ordered by the time of the current visit v (dashed vertical black line) of a new patient. At each of these visits, we combine the accumulated longitudinal data (shown in blue circles), and time of the last negative invasive test t (solid vertical green line) to obtain the updated cumulative-risk profile $R_j(u \mid t, v)$ (dotted red line with 95% credible interval shaded) of the patient. The benefit of this time-dynamic property is that the resulting schedules in Section 3.2 and their estimated burden and benefit in Section 3.3 are also time-dynamic.

5. **Starting from Section 3.2, the notation gets heavy. The authors should give example to illustrate the meaning of some notation (e.g., t_l, $N_j(S_j^k)$) using Figure 3. The notation of t_l is hard to understand as it is defined on P13.**

We thank the Referee for motivating us to improve the explanation for the notation t_l of the 'time of the last invasive test', and $N_j(S_j^k)$ of the 'number of tests conducted'. Following the referee's suggestion we have added explanation for t_l and $N_j(S_j^k)$ using the example of Figure 3. The added explanation of t_l in Section 3.2 reads as:

> ... We further illustrate the test scheduling process using Figure 3. In the figure, at the current visit (a real physical visit of a patient) denoted by $l = 1$ the corresponding time of last test t_1 is set to $t_1 = t = 1.5$. Here, t is the time of the last known test, likely extracted from the medical records of the patient. At the current visit $l = 1$ the cumulative-risk is lower than the set threshold of 12%. Thus, a decision of not conducting a test is taken at current time u_1, denoted by $Q_j^{\kappa}(u_1 \mid t_1, v) = 0$. It is important to note at this point all visits with $l > 1$ are future visits that have not yet occurred. At the next visit $l = 2$ (the first future visit), the corresponding time of last test t_2 is still set to $t = 1.5$ because t is still the time of the last test. However, at this visit $l = 2$ the cumulative-risk is more than the set threshold and it is decided to plan a test at this visit, denoted by $Q_j^{\kappa}(u_2 \mid t_2, v) = 1$. Consequently, at the third visit $l = 3$ (the second future visit), the time of the last test t_3 switches from t to $t_3 = t_2$, and t_2 remains the time of last test until at any future test a new test is planned again. The process is continued until the last planned visit $l = L$. We should note that in all future test decisions (visits with $l > 1$), we use only the observed longitudinal data up to the current (real visit) visit time $u_1 = v$, i.e., $\{\mathcal{Y}_{1j}(v), \ldots, Y_{Kj}(v)\}$. ...

The added explanation of $N_j(S_j^k)$ in Section 3.3 reads as:

> ... To understand $\mathcal{N}_j(S_j^{\kappa})$, consider Figure 3 wherein the schedule contains two planned future tests at future visit times $u_2 = 3.5$ and $u_4 = 5.5$ years. Suppose that when the patient undergoes a real test at $u_2 = 3.5$ years, progression is detected and the patient is removed from surveillance. Then, the total tests performed will be $\mathcal{N}_j(S_j^{\kappa}) = 1$. On the other hand, if progression is detected on a real test at u_4 then total tests performed will be $\mathcal{N}_j(S_j^{\kappa}) = 2$. In a real world situation it is not known when a patient will progress and how many of the planned tests will be really conducted. However, we can obtain a personalized estimate of the number of future tests that will get conducted, denoted by the expected value $E\{\mathcal{N}_j(S_j^{\kappa})\}$, and defined as, ...

6. **Figure 4 is presented without much explanation. It is difficult to understand.**

We thank the Referee for motivating us to improve the explanation for Figure 4. To assist readers in understanding it, we have now added the following explanation for it in Section 3.4:

> ... To illustrate this Euclidean space, we use the example patient shown in Figure 3. For this patient, using (2) we obtained 200 schedules corresponding to 200 risk thresholds between 0% and 100% separated by every 0.5%. For each such schedule, we obtained the personalized expected number of tests and personalized expected delay using (4) and (5), respectively, and plotted them in two dimensions in Figure 4. ...

In the subsequent paragraph we have added references (colors and shape both have been mentioned to assist readers) to Figure 4. They are underlined in the following text that has been taken from the revised manuscript.

> The ideal schedule (blue rectangle in Figure 4) for j-th patient is the one in which only one test is conducted, at exactly the true time of progression T_j^*. In other words, the time delay will be zero. If we weigh the expected number of tests and time delay as equally important, then we can select as the optimal threshold at current visit time v, the threshold $\kappa^*(v)$ which minimizes the Euclidean distance (dashed gray lines connecting the black circles and blue rectangles in Figure 4) between the ideal schedule, i.e., point (1, 0) and the set of points representing the different personalized schedules S_j^κ corresponding to various $\kappa \in [0, 1]$, i.e.,

7. **In Section 4.1, the automatic chosen threshold $\kappa^*(v)$ should be listed. The details of how $\kappa^*(v)$ is estimated in the real data and how the testing schedule is determined as in Figure 5 should be given. Also, in Figure 5, when k = 10%, how the testing schedule is determined should be also given as there is a large time gap between the two tests after year 6.**

   We thank the Referee for motivating us to check the robustness of our model against Gleason score misclassification. A biopsy Gleason score can be misclassified to be less (or more) than the pathological score that is obtained after prostatectomy. Ignoring such misclassification will affect the parameter estimates as well risk predictions. However, since joint models utilize a relative risk sub-model for modeling time-to-event data, their robustness to misclassification is similar to relative risk models (e.g., Cox proportional hazards model). We next discuss the challenges in accounting for Gleason misclassification.

8. **In Figure 6, what do the red dots represent? We can see that $\kappa^*(v)$ did not outperform the simple PRIAS schedule. Among the progressing patients, $\kappa^*(v)$**

and PRIAS have similar number of biopsies, but $\kappa^*(v)$ has higher median time delay. Actually, $\kappa^*\{v \mid E(\mathcal{D}) \leq 0.75\}$ and $\kappa = 10\%$ have either higher number of biopsies or longer time delay, as compared to PRIAS. The advantage of the proposed method is questionable.

We thank the Referee for motivating us to check the robustness of our model against Gleason score misclassification. A biopsy Gleason score can be misclassified to be less (or more) than the pathological score that is obtained after prostatectomy. Ignoring such misclassification will affect the parameter estimates as well risk predictions. However, since joint models utilize a relative risk sub-model for modeling time-to-event data, their robustness to misclassification is similar to relative risk models (e.g., Cox proportional hazards model). We next discuss the challenges in accounting for Gleason misclassification.

9. **In the supplement B.2, Eq(6), it is unclear why this particular function form is selected. Did the authors do any model selection? Any rationale for this model?**

   We thank the referee for giving us the opportunity to provide our motivation with regards to the choice of the model. However, since this question is overlaps with Question 2, our response to this question is same as that for Question 2.

10. **In the supplement B.4, the Q-Q plot of df=4 should also be provided so that the reviewer can see why df=3 is selected.**

    We thank the referee for giving us the opportunity to defend the choice of the error distribution for PSA. The QQ plot of residuals in the supplementary file now contains an extra panel for a model with t-distribution having four degrees of freedom. The model assumption for the error term was best met by the model with t-distribution having three degrees of freedom. The updated QQ plot is also shown in Figure 1 in this letter.

11. **Supplement P23, l34, 9 should be Table 9.** We thank the referee for noting our error. We have fixed it and now it reads as:

    > ... The corresponding results, using 500×250 test patients are presented in Table 9.
    > ....

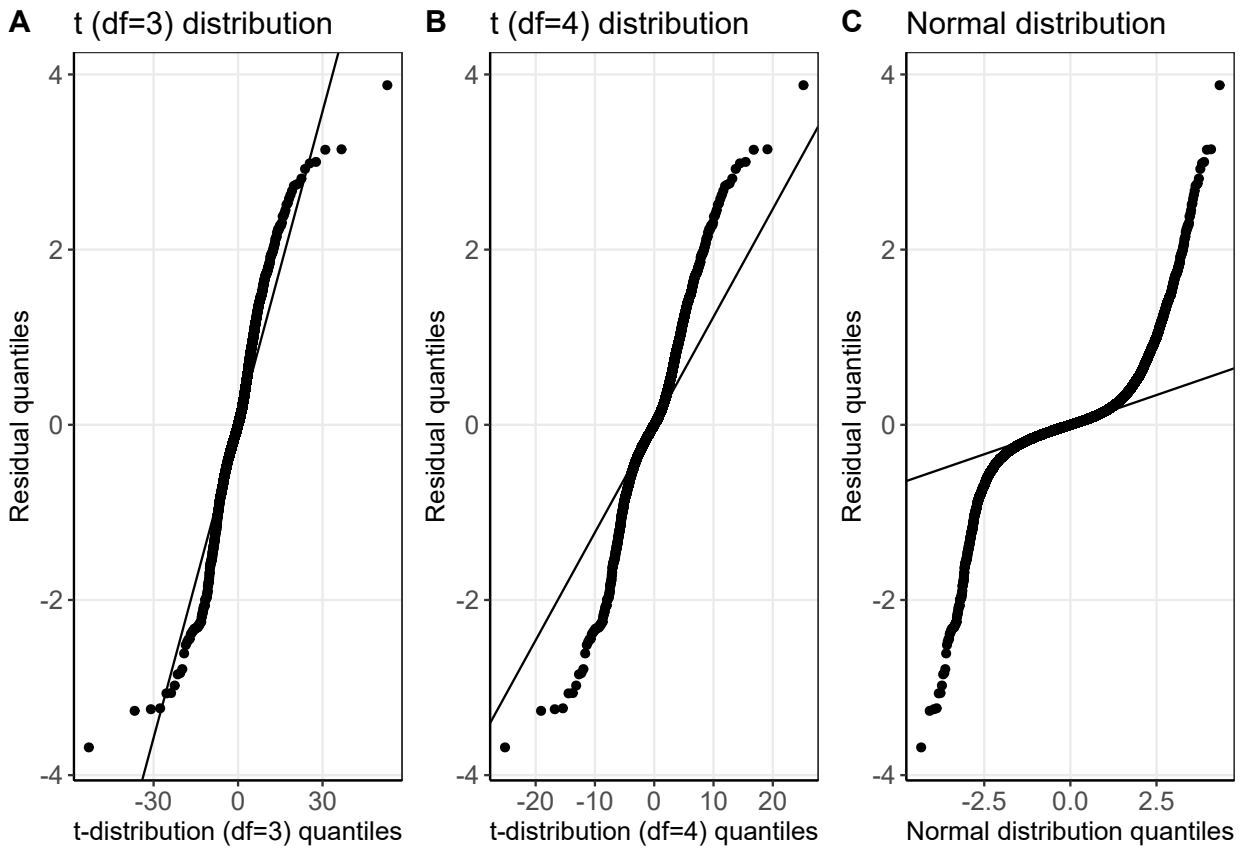

Figure 1: Quantile-quantile plot of subject-specific residuals from the joint models fitted to the PRIAS dataset. **Panel A**: model assuming a t-distribution (df=3) for the error term ε_p. **Panel B**: model assuming a t-distribution (df=4) for the error term ε_p. **Panel C**: model assuming a normal distribution for the error term ε_p.

Response to 2nd Referee's Comments

We would like to thank the Referee for their constructive comments, which have allowed us to considerably improve our paper. The main differences of the new version of the manuscript compared to the previous one can be found in the section titled "Study Population".

You may find below our responses to the specific issues raised.

1. **A key point of the paper is to determine subsequent schedules of multiple tests over the entire follow-up period at a specific clinical visit time v, whereas the existing literature mostly focused on the timing of the subsequent test following the time of visit. It is not clear why it is useful in this clinical setting to figure out all subsequent test schedules. The clinical decision at the visit is if a test should be scheduled earlier, on time, or late compared with the standard fixed schedule, please provide some justification why knowing a hypothetical schedule (also see #2) would help with that clinical decision.**

   We thank the Referee for their suggestion of a risk calculator. We have already developed a web-application for this purpose (Figure **??**). In this web-application doctors can load patient data via CSV files (and other formats such as SPSS files). To aid in shared decision making, the web-application not only estimates the cumulative risk of cancer progression at the current visit, but also on future visits. These estimates are time-dynamic, that is, they get updated as additional patient data is collected over time. In addition, patient-specific fitted PSA and DRE profiles and their future predictions are also provided.

2. **Related to 1, the long-term schedule developed based on the information up to the visit day $v, \mathcal{Y}(v)$ without updating based on newer information is limited. As the decision on later tests after the subsequent one would be dependent on the additional finding from that test to be scheduled and newer clinical information collected at the next visit prior to the future scheduled visit. That is, predicting later tests based on information collected at the current clinical visit time may not be meaningful, as it is not updated by new information Y cumulated up to a more recent visit $\mathcal{Y}(v+s)$ for $s > 0$).**

   We thank the Referee for motivating us to check the robustness of our model against Gleason score misclassification. A biopsy Gleason score can be misclassified to be less (or more) than

the pathological score that is obtained after prostatectomy. Ignoring such misclassification will affect the parameter estimates as well risk predictions. However, since joint models utilize a relative risk sub-model for modeling time-to-event data, their robustness to misclassification is similar to relative risk models (e.g., Cox proportional hazards model). We next discuss the challenges in accounting for Gleason misclassification.

3. **Please provide details of how the cumulative-risk $R(T > u \mid T > t, Y(v))$ can be calculated from a time-varying covariate model of the form $P(T > t|Y(t))$ from a joint modeling framework, given $Y(v)$ with $v > t$.**

   We thank the Referee for motivating us to check the robustness of our model against Gleason score misclassification. A biopsy Gleason score can be misclassified to be less (or more) than the pathological score that is obtained after prostatectomy. Ignoring such misclassification will affect the parameter estimates as well risk predictions. However, since joint models utilize a relative risk sub-model for modeling time-to-event data, their robustness to misclassification is similar to relative risk models (e.g., Cox proportional hazards model). We next discuss the challenges in accounting for Gleason misclassification.

4. **The schedule of planned future tests S_j^k is determined by $R_j(u \mid t, v)$ with u and t vary but v is constant in determining the set of S_j^k, therefore S_j^k should be dependent on v. Should the calculation of $EN_j(S_j^k)$ and $D_j(S_j^k)$ be additionally conditioning on $T_j^* > v$?**

   We thank the Referee for motivating us to check the robustness of our model against Gleason score misclassification. A biopsy Gleason score can be misclassified to be less (or more) than the pathological score that is obtained after prostatectomy. Ignoring such misclassification will affect the parameter estimates as well risk predictions. However, since joint models utilize a relative risk sub-model for modeling time-to-event data, their robustness to misclassification is similar to relative risk models (e.g., Cox proportional hazards model). We next discuss the challenges in accounting for Gleason misclassification.

5. **In PRIAS study, is the dataset split into a training set for joint modeling, and a testing set for determining the personalized schedules and calculating the validation summaries? Also if $\kappa = 10\%$ was decided as in Figure 4, should one consider cross-validation?**

   We thank the Referee for motivating us to check the robustness of our model against Gleason score misclassification. A biopsy Gleason score can be misclassified to be less (or more) than

the pathological score that is obtained after prostatectomy. Ignoring such misclassification will affect the parameter estimates as well risk predictions. However, since joint models utilize a relative risk sub-model for modeling time-to-event data, their robustness to misclassification is similar to relative risk models (e.g., Cox proportional hazards model). We next discuss the challenges in accounting for Gleason misclassification.

6. **Page 14, last line, $t_5 = 4.5$ year, should it be $t_5 = 5.5$ year?**

   We thank the referee for noting our error. We have fixed it in the revised manuscript.

References

Balasubramanian, Raji and Stephen W Lagakos (2003). "Estimation of a failure time distribution based on imperfect diagnostic tests". In: *Biometrika* 90.1, pp. 171–182.

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.

Carvalho, Tiago M de, Eveline AM Heijnsdijk, and Harry J de Koning (2017). "Estimating the risks and benefits of active surveillance protocols for prostate cancer: a microsimulation study". In: *BJU International* 119.4, pp. 560–566.

Coley, R Yates et al. (2017). "Prediction of the pathologic Gleason score to inform a personalized management program for prostate cancer". In: *European Urology* 72.1, pp. 135–141.

Inoue, Lurdes YT et al. (2018). "Comparative Analysis of Biopsy Upgrading in Four Prostate Cancer Active Surveillance Cohorts". In: *Annals of Internal Medicine* 168.1, pp. 1–9.