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

April 5, 2019

Professor Alan Schwartz
Department of Medical Education
University of Illinois at Chicago
Chicago, IL, 60612-7309
USA

Dear Professor Schwartz,

We are writing to you with respect to the manuscript #MDM-19-037, titled "Personalized Decision Making for Biopsies in Prostate Cancer Active Surveillance Programs" submitted to *Medical Decision Making* 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 description for the model equations in the Methods section. We have also discussed the impact of increasing use of MRI on the findings reported in our paper.

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 Study Population section of the paper.

1. Risk calculator for clinicians for bringing the concept to practice

   We thank the Referee for their suggestion of a risk calculator. We have already developed a web-application for this purpose (Figure 1). 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.

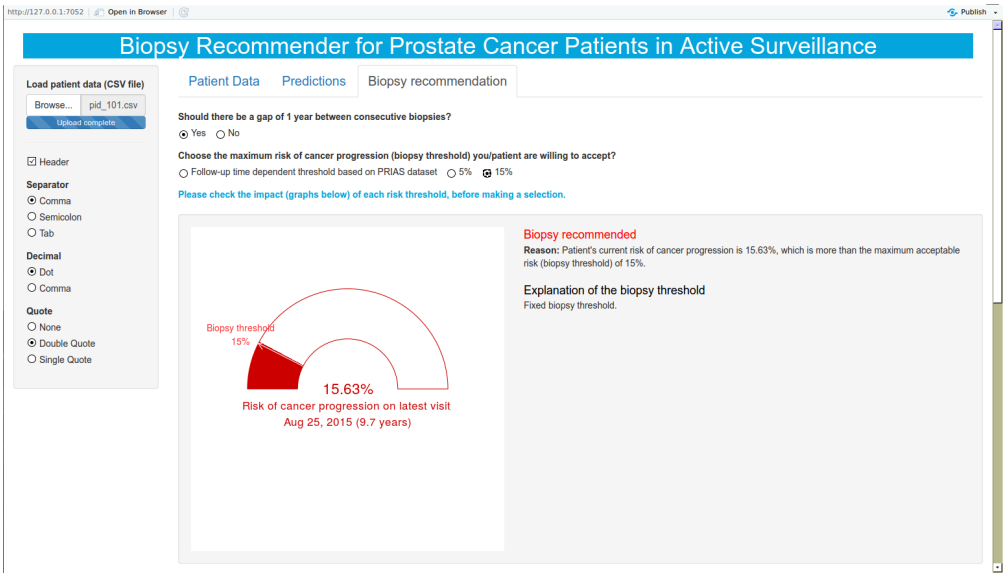

Figure 1: Web-based risk calculator for prediction of risk of prostate cancer progression.

Although the web-application is ready, our model requires further validation. To this end, we are currently validating it using active surveillance data from the GAP3 database (Bruinsma et al., 2018). The GAP3 database consists of patient data from multiple active surveillance cohorts around the world. Based on the current validation results so far, we expect the

web-application to be also useful in other cohorts which are similar to PRIAS.

2. Model robustness to Gleason score misclassification

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 than the pathological score (prostatectomy), or vice versa. Ignoring such misclassification will affect the parameter estimates as well risk predictions.

Impact of Gleason misclassification on parameter estimates: The impact of Gleason misclassification on parameter estimates depends upon the misclassification rate. Higher rates of misclassification may lead to a larger bias. Currently, there is little consensus on Gleason misclassification rates (Cookson et al., 1997; Lattouf and Saad, 2002; Melia et al., 2006; Pinthus et al., 2006; Ploussard et al., 2010), which may also vary between the cohorts. An alternative is to obtain this information from our dataset, and model pathological Gleason score as an outcome in our model (Balasubramanian and Lagakos, 2003; Coley et al., 2017). However, such information is sparsely available in the PRIAS dataset. Hence, accounting for this in our model is currently not feasible.

Impact of Gleason misclassification on risk prediction: When a Gleason score of 7 or more is misclassified as a Gleason score of 6 or less, the patient is removed from active surveillance for treatment. Hence predictions from our model may not be used afterwards. We next discuss the alternative scenario wherein a Gleason score of 7 or more is misclassified as a Gleason score of 6 or less. In this case the patient remains in active surveillance and our model may be used to schedule biopsies.

Model utilizes multiple sources of information: The personalized risk of cancer progression for a patient j depends upon four sources of information, namely the time of the latest negative biopsy t, the repeated measurements $\mathcal{Y}_{pj}(s)$ and $\mathcal{Y}_{dj}(s)$ of PSA (prostate-specific antigen) and DRE (digital rectal examinations), respectively, and the training data $\mathcal{D}_n$. That is the model does not rely completely on the observed Gleason score. This is evident by the following formula for the risk of cancer progression at a time s:

$$R_j(s \mid t) = \Pr\big\{T_j^* \leq s \mid T_j^* > t, \mathcal{Y}_{dj}(s), \mathcal{Y}_{pj}(s), \mathcal{D}_n\big\}, \quad s \geq t, \tag{1}$$

where T_j^* is true time of cancer progression (Gleason upgrade).

Impact of 100% misclassification rate on risk prediction: In order to check the robustness of predictions against Gleason misclassification, we conducted a set of simulations. The setup of

these simulations was same as that of the main simulation study reported in our manuscript. Let us assume that we schedule a biopsy for all patients at the exact time of their cancer progression T_j^*. We then assume that we incorrectly obtain a Gleason score less than 7 for all such patients, that is, we have a 100% misclassification rate. Consequently, these patients will remain in active surveillance and will be invited at future follow-up time points $s > T_j^*$. The risk of cancer progression at these future time points is given by Equation (1), with $t = T_j^*$. We then make a decision of biopsy at the future time points based on two different risk cutoffs, namely 5%, 10%. We assume that the urologist will correctly classify the Gleason score on this second biopsy. Consequently, the delay in detection of cancer progression is the difference between the time of the second biopsy and the true time of cancer progression (at which a biopsy was done, but Gleason score was misclassified).

In the aforementioned scenario, we observed that for a risk cutoff of 5% the delay in detection of cancer progression (years) is median: 0.6, IQR: 0.3–1.0. This is similar to the delay we observe in the main simulation study for this cutoff (median: 0.6, IQR: 0.3–0.9) wherein we ignore Gleason misclassification. For a risk cutoff of 10% the delay in years is median: 1.0, IQR: 0.6–1.8. In this case the third quartile of the delay is much more than the case wherein we ignore Gleason misclassification (median: 0.7, IQR: 0.3–1.0). In AS, a delay in the detection of cancer progression around 12 to 14 months is assumed to be unlikely to substantially increase the risk of adverse downstream outcomes (Carvalho, Heijnsdijk, and Koning, 2017; Inoue et al., 2018). While the third quartile of the delay with a 10% cutoff is 7 months above this limit, this result is based on an extreme scenario of 100% misclassification. In reality Gleason misclassification rates are much lower, and rates as high as 50% (random classification) are only likely when pathological Gleason score is exactly 7 (Cookson et al., 1997; Lattouf and Saad, 2002; Melia et al., 2006; Pinthus et al., 2006; Ploussard et al., 2010). In light of these arguments, a 5% to 10% cutoff may still be useful to decide biopsies.

3. Typos on page 6, line 44

We thank the Referee for pointing out the error. The new text is:

> **Data Accessibility:** The PRIAS database is not openly accessible. However, access to the database can be requested on the basis of a study proposal approved by the PRIAS steering committee. The website of the PRIAS program is `www.prias-project.org`.

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 Discussion section.

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

1. Impact of increased use of MRIs on our risk calculator

   We agree with the Referee that this is an important issue. In the recent years the use of MRIs to decide biopsies has increased. However, currently MRI data is extremely sparsely available in the PRIAS dataset. Hence, it may not be useful to include it in our model. However, in the next few years we expect to have enough MRI data, such as volume of the prostate tumor in the PRIAS dataset. This data can then be used as a third biomarker in our model, alongside PSA (prostate-specific antigen) and DRE (digital rectal examination) measurements. We expect that this extended model will lead to better risk predictions than using information from MRIs alone.

   The current model can also be used to decide the time of conducting an MRI. This is especially relevant for patients in developing countries, where an MRI scan is still expensive.

   We have added the aforementioned points briefly in the discussion section of the manuscript.

2. Use of patient characteristics, and quality of life in the simulation

   We agree with the Referee that this is an important issue. Currently, the model fitted to the PRIAS dataset, which is also used as the simulation model, accounts for patient age. The Referee also noted the importance of other factors such as having a first degree relative with cancer. However, this factor has been found to be predictive of cancer progression only in African-American patients (Goh et al., 2013; Telang et al., 2017). This is also evident by the fact that PRIAS and many other surveillance programs do not utilize this information in their biopsy protocols (Bokhorst et al., 2016; Nieboer et al., 2018). In addition, patients who have a higher risk of an aggressive form of cancer are usually not recommended active surveillance. Hence the proposed model is relevant only for low-risk prostate cancer patients eligible for active surveillance. An exception are the active surveillance patients who are old and/or have comorbid illnesses. Currently, such patients may be removed from active surveillance and are

instead offered the less intensive watchful waiting (Bokhorst et al., 2016) option. While it is possible to model watchful waiting as a competing risk in our model, but this falls outside the scope of the current work.

How the findings of the model should be utilized in the larger clinical context and what additional work will be needed before the model can be used clinically? The use of personalized risk based biopsy schedules using the proposed model requires further work. A first step will be to initiate studies to look at feasibility acceptability and usability of such an approach in daily clinical practice. Biopsy decision making may include other factors such quality of life, patient preferences etc., which are not included in this model. However having an objective estimate next to (subjective) patient and care giver considerations will aid decision making. To facilitate such shared decision making we have developed a web-application for clinicians. We expect to hand it over to clinicians after an external validation of our model in active surveillance cohorts similar to PRIAS. This latter work is currently under progress.

Response to 3rd 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 Methods section.

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

1. Base 2 logarithm transformation for PSA

   In many prostate cancer surveillance programs, and specifically in the world's largest surveillance program PRIAS (Prostate Cancer International Active Surveillance), $\log_2$ PSA values have been used for analysis. The reason is that this transformation provides an easy interpretation of the slope of the regression line that is fitted individually for each patient's $\log_2$ PSA levels. More specifically, the inverse of the slope of this regression line is known as the PSA doubling time (Roberts et al., 2001). Although we model PSA levels non-linearly using B-splines (De Boor, 1978), we decided to reuse the $\log_2$ PSA transformation for familiarity.

   Using $\log_2$ PSA transformation leads to the same model fit as with $\log_e$ PSA transformation, because logarithms with different bases are scalar multiples of each other.

2,3. Better explanation of the equations in the Methods section

   We would like to thank the Referee for motivating us to describe these in details. We have added the following text (shaded) in the Methods section.

   What are internal knots and boundary knots?

   > ... our B-spline basis function $B_k(t, \mathcal{K})$ has 3 internal knots at $\mathcal{K} = \{0.1, 0.7, 4\}$ years, and boundary knots at 0 and 5.42 years (95-th percentile of the observed follow-up times). This specification allows fitting the $\log_2(\text{PSA} + 1)$ levels in a piecewise manner for each patient separately. The internal and boundary knots specify the different time periods (analogously pieces) of this piecewise curve. ...

   Formula 6 on page 12 could also benefit casual readers with a little hand holding:

$$
\begin{aligned}
\mathrm{F}_1(t, s, \kappa) &= 2\frac{\mathrm{TPR}(t, s, \kappa)\ \mathrm{PPV}(t, s, \kappa)}{\mathrm{TPR}(t, s, \kappa) + \mathrm{PPV}(t, s, \kappa)}, \\
\mathrm{TPR}(t, s, \kappa) &= \Pr\big\{R_j(s \mid t) > \kappa \mid t < T_j^* \leq s\big\}, \\
\mathrm{PPV}(t, s, \kappa) &= \Pr\big\{t < T_j^* \leq s \mid R_j(s \mid t) > \kappa\big\},
\end{aligned}
\tag{2}
$$

where, $\mathrm{TPR}(t, s, \kappa)$ and $\mathrm{PPV}(t, s, \kappa)$ are the time dependent true positive rate and positive predictive value, respectively. These values are unique for each combination of the time period $(t, s]$ and the risk threshold κ that is used to discriminate between the patients whose cancer progresses in this time period versus the patients whose cancer does not progress. The same holds true for the resulting F_1 score $\mathrm{F}_1(t, s, \kappa)$. The F_1 score ranges between 0 and 1, where a value equal to 1 indicates perfect TPR and PPV. Thus the highest F_1 score is desired in each time period $(t, s]$. This can be achieved by choosing a risk threshold κ which maximizes $\mathrm{F}_1(t, s, \kappa)$. That is, during a patient's visit at time s, given that his latest biopsy was at time t, the visit specific risk threshold to decide a biopsy is given by $\kappa = \arg\max_\kappa \mathrm{F}_1(t, s, \kappa)$. ...

Authors should state why they chose to center age at the value of 70:

... Age_i is the age of the i-th patient at the time of inclusion in AS. We have centered the Age variable around the median age of 70 years for better convergence during parameter estimation. However, this does not change the interpretation of the parameters corresponding to the Age variable. ...

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. (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.

Bruinsma, Sophie M et al. (2018). "The Movember Foundation's GAP3 cohort: a profile of the largest global prostate cancer active surveillance database to date". In: *BJU international* 121.5, pp. 737–744.

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, Rebecca 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.

Cookson, Michael S. et al. (1997). "Correlation between Gleason Score of Needle Biopsy and Radical Prostatectomy Specimen: Accuracy and Clinical Implications". In: *Journal of Urology* 157.2, pp. 559–562.

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

Goh, Chee L et al. (2013). "Clinical implications of family history of prostate cancer and genetic risk single nucleotide polymorphism (SNP) profiles in an active surveillance cohort". In: *BJU international* 112.5, pp. 666–673.

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.

Lattouf, J.-B. and F. Saad (2002). "Gleason score on biopsy: is it reliable for predicting the final grade on pathology?" In: *BJU International* 90.7, pp. 694–698.

Melia, J et al. (2006). "A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies". In: *Histopathology* 48.6, pp. 644–654.

Nieboer, Daan et al. (2018). "Active surveillance: a review of risk-based, dynamic monitoring". In: *Translational andrology and urology* 7.1, pp. 106–115.

Pinthus, Jehonathan H. et al. (2006). "Prostate Cancers Scored as Gleason 6 on Prostate Biopsy are Frequently Gleason 7 Tumors at Radical Prostatectomy: Implication on Outcome". In: *Journal of Urology* 176.3, pp. 979–984.

Ploussard, Guillaume et al. (2010). “Pathological Findings and Prostate Specific Antigen Outcomes After Radical Prostatectomy in Men Eligible for Active Surveillance—Does the Risk of Misclassification Vary According to Biopsy Criteria?” In: *Journal of Urology* 183.2, pp. 539–545.

Roberts, Steven G et al. (2001). “PSA Doubling Time as a Predictor of Clinical Progression After Biochemical Failure Following Radical Prostatectomy for Prostate Cancer”. In: *Mayo Clinic Proceedings* 76.6, pp. 576–581.

Telang, Jaya M et al. (2017). “Prostate cancer family history and eligibility for active surveillance: a systematic review of the literature”. In: *BJU international* 120.4, pp. 464–467.