# Discussion on the Comments Received on the MDM Paper Draft

## 1 Definition of Harm/Benefits and their Link with the Biopsy Threshold $\kappa$

We recently had a meeting with Monique and her department where we discussed various measures of harm/benefits. A few points from that meeting:

1. There is a lack of data for measures of harm/benefit such as risk of complications, probability of death due to too much delay in detection of progression, and quality of life, etc.

2. #biposies (more is harm, less is benefit), and delay in detection of cancer progression (more is harm, less is benefit) are the best quantifiable measures of harm/benefit.

3. There is no easy and/or correct way to relate the #biopsies and delay. That is, 1 biopsy can not be equivalent to 'X' years of delay. Thus the total harm cannot be obtained as a single utility function, but it remains as two separate measures.

4. In a *surveillance setting* there is little expert knowledge available on consequences of using a biopsy risk threshold $\kappa$ on the #biopsies and delay. Only after a simulation like ours (or after a real trial for risk based schedules), can we understand the true utility of these thresholds.

Because of point number 3, getting the odds of harm/benefit is impractical. Consequently, linking (un-quantifiable) odds to threshold $\kappa$ is not possible either. Methods such as Net-benefit, avoid the issue in point 3 by assuming that there is a latent equivalence between #biopsies and delay at some particular threshold on which a doctor does not know if a biopsy should be conducted. Conditional on that being true, the utility of the threshold can directly be based on the threshold (example: 10% means 1:9 ratio of harm to benefit). However, as is evident by our simulation results, for most thresholds between 5% and 20% the basic assumption of harm being equal to benefit will not hold true. Hence, the Net-benefit is not the true utility of a threshold under such a scenario. The true utility of the threshold is simply the #biopsies and delay, because eventually that is what really matters to patients and doctors. An alternative way to evaluate the 2 measures together can be this: We can put a constraint on #biopsies (e.g., not more than 3 biopsies) and minimize for the delay. And that may be an indirect utility of the threshold.

**Accounting for Variance:** The utility of each threshold cannot be a single number for #biopsy / harm, but is rather an entire distribution. Some schedules have a very large variance for #biopsies (annual schedule), and others for delay.

Hence, any method which links threshold with #biopsies and delay should account for the large variances. Existing approaches focus on mean, or median and avoid variance completely, despite the fact that it is large.

### 1.1 Risk ‘automatic’

We agree that the word automatic in our paper does not correctly describe what we actually do. It should actually be *Risk ‘ROC’*. The motivation of an idea where an expert does not select a threshold but rather it is data driven is, point number 4 of Section 1.

When the assumption of harm being equal to benefit does not hold true, what is left is simple true positive rate, false positive rate, false negative rate, and true negative rate. We suggest that in absence of expert knowledge a threshold can also be chosen by looking into prediction accuracy, in the observed data set. That is, what does a threshold predict about who will obtain cancer versus others? We use $\kappa$ which maximizes the F1-score given by:

$$\frac{2TP}{2TP + FP + FN} \tag{1}$$

However, it is important to note that F1-score is NOT a utility function. We do not say, that F1-score gives the best net-benefit or is best for the patient. Is $\kappa$ from F1-score is better than a 5% threshold chosen by a urologist?, can only be evaluated by looking at the **true utility**: the entire distribution of the #biopsies and delay. The latter two are only available after simulation and hence utility of an approach can only be discussed retrospectively (supported by point number 3, point number 4 and the rest of discussion in Section 1).

## 2 Fixed Budget of Biopsies Approach

Thank you for proposing this idea. Indeed, a patient might also propose a fixed budget on number of biopsies that they can handle during their AS follow-up period. The two variants that you proposed are:

(A) Fixed schedule for all, with budget equal to number of biopsies due to PRIAS and annual schedule, but optimized in terms of timing, for example shorter intervals in early phase of follow-up.

(B) Personalized schedule for all, with budget equal to number of biopsies due to PRIAS and annual schedule.

The motivation for this approach: The difference between (A) and (B) will show how much we gain by personalizing.

This is an interesting idea. We can check this by trying various fixed heuristic schedules. Or, another way to do this could be what we have already done. We put a fixed budget on delay and compare the #biopsies (compliment of what you propose). As shown in Figure 1 the personalized approach does better than PRIAS in both #biopsies and the delay for all types of patients. The difference in the number of biopsies is the gain by personalizing biopsies. It is important to note that the reason we match the budget for delay but not for biopsies is that because patients do not want as many #biopsies as are scheduled in PRIAS. Otherwise the whole issue of non-compliance would not exist.

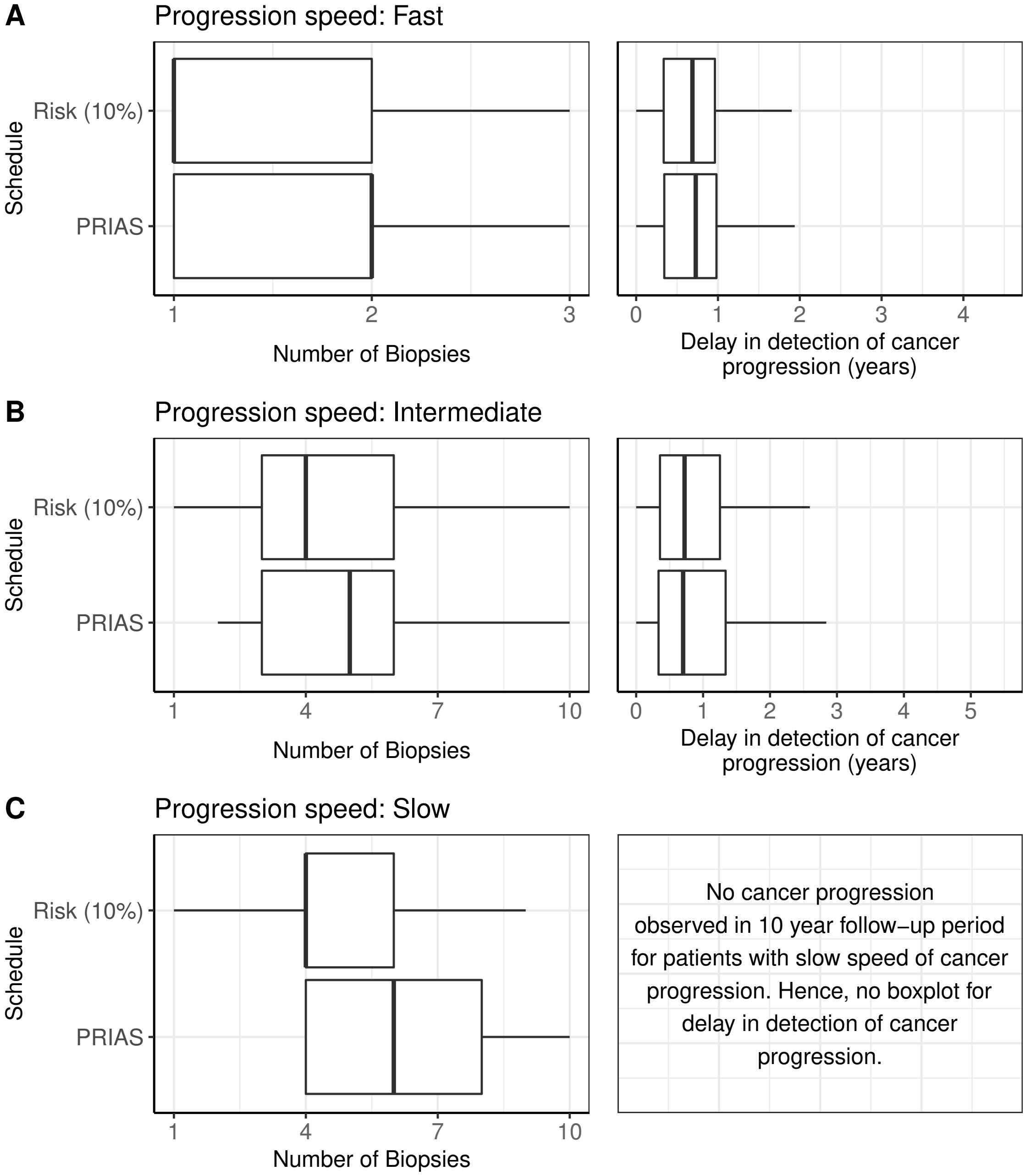

Figure 1: Boxplot showing variation in number of biopsies, and the delay in detection of cancer progression, in years (time of last biopsy - true time of cancer progression) for two biopsy schedules: personalized biopsies with 10% risk and PRIAS schedule. Biopsies are conducted until cancer progression is detected. **Panel A:** results for simulated patients who had a faster speed of cancer progression, with progression times between 0 and 3.5 years. **Panel B:** results for simulated patients who had an intermediate speed of cancer progression, with progression times between 3.5 and 10 years. **Panel C:** results for simulated patients who did not have cancer progression in the 10 years of follow-up.

# 3 Statistical Modeling Issues

1. Age 70 centering point, and age quadratic term

   Yes, as you mentioned, 70 is the centering point. We chose 70 as the centering point because it is the mean age at the time of inclusion in AS, in PRIAS. Transformation of age around a centering point provides better MCMC convergence while fitting the model. This is especially relevant because of the quadratic effect of age. Since we have a large dataset with 58911 repeated measurements of 5270 patients, and 866 events, we tried to model age in more sophisticated manner.

2. PSA velocity strong effect

   Yes, this is indeed an important finding. There are 3 important points in this regard: Firstly, the effect of velocity looks large compared to the effect of DRE log odds because they are measured on different scales. A better comparison, as you suggested previously can be done this way: Increase in velocity from 1st to 3rd quartile (of fitted velocities) leads to 1.92 fold increase in hazard of cancer progression. Similar increase in DRE log odds leads to 1.40 fold increase in hazard. Thus, we can say that the impact of both is similar.

   Secondly, typically in a *screening setting* (Andrew Vickers' paper on velocity) velocity and/or PSA-DT have been calculated by either fitting a linear regression on observed PSA values, or by just calculating slope directly from the first and last observed PSA values. However, almost all PSA profiles in the *surveillance setting* are highly non linear. In comparison to the existing work, the velocity that we fit is given by the derivative of the underlying non-linear PSA trajectory (we account for measurement error). It is thus a more sophisticated function, given by ($i$-th patient):

   $$\text{Velocity}_i(t) = \frac{\mathrm{d} \sum_{k=1}^{4} (\beta_k + b_{ki}) B_k(t, \mathcal{K})}{\mathrm{d}t},$$

   where, $B_k(t, \mathcal{K})$ is the basis function of a B-spline, $\beta_k$ are fixed effects, and $b_{ki}$ are the random effects. The equation, $\sum_{k=1}^{4} (\beta_k + b_{ki}) B_k(t, \mathcal{K})$ gives us an individualized non-linear PSA trajectory over time. So the joint model PSA velocity is somewhat different in my opinion. A recent paper used a joint model and found PSA velocity to be important [Cooperberg et al., 2018].

   Lastly, as you mentioned it may be that we are missing on certain covariates (confounding from unobserved sources). It is possible, however that is a general issue in all statistical models. In our model an extra step I did was I also added rate of change of log odds of DRE in the hazard function. Even in its presence, PSA velocity was almost the same effect size in our model. Changing positions of knots in the B-spline non-linear evolution of PSA also lead to similar results. Thus, I am confidant that given the data that we have, these results hold.

3. AUC is a bit abstract, and AUC around 0.6

   Thank you for pointing this out. I also agree that we have not made the time depedent AUC concept clear. I will try to explain it better in the next draft, especially its importance in context of personalization.

Regarding the low AUC: the previous AUC calculations were not complete, and were also not correct. The new AUC numbers are presenting in Figure 2. A short summary of this plot is: at any given follow-up time point the AUC numbers depict the model's ability to discriminate between patients who obtained cancer progression and patients who did not obtain cancer progression, in the last 1 year. The AUC at year 1 is 0.65 and then at year 3 it reaches a maximum of 0.75, and eventually decreases to 0.6 at year 5 (95-th percentile of observed progression times).

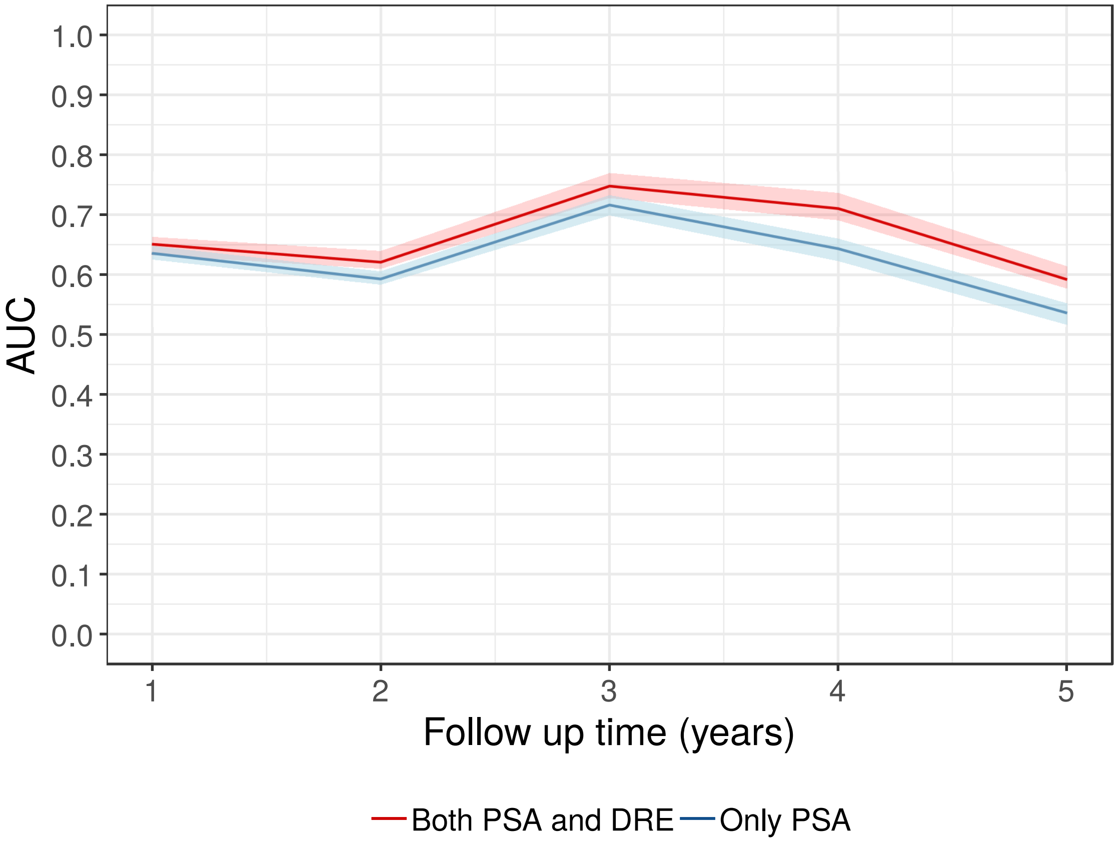

Figure 2: Area under the receiver operating characteristic curve, calculated at 5 different follow-up time points, for two different joint models (one model ignores DRE, and another does not).

4. 95% credible intervals are quite small at individual level

   The plot in question, is probably the plot of probability of DRE > T1c over time. To confirm that the numbers are correct we calculated the 95% credible intervals again and found the results to be the same. However, these are not individual patient level results. Instead, these are marginal probabilities obtained by integrating out the subject specific random effects from the DRE model equation. In Figure 3 we present both the marginal log odds and marginal probability of obtaining a DRE > T1c.

5. t-distribution QQ-plot not perfect

   The choice of t-distribution (df=3) originated from the feedback of reviewers on our previous Biometrics paper. For the model that we propose in this

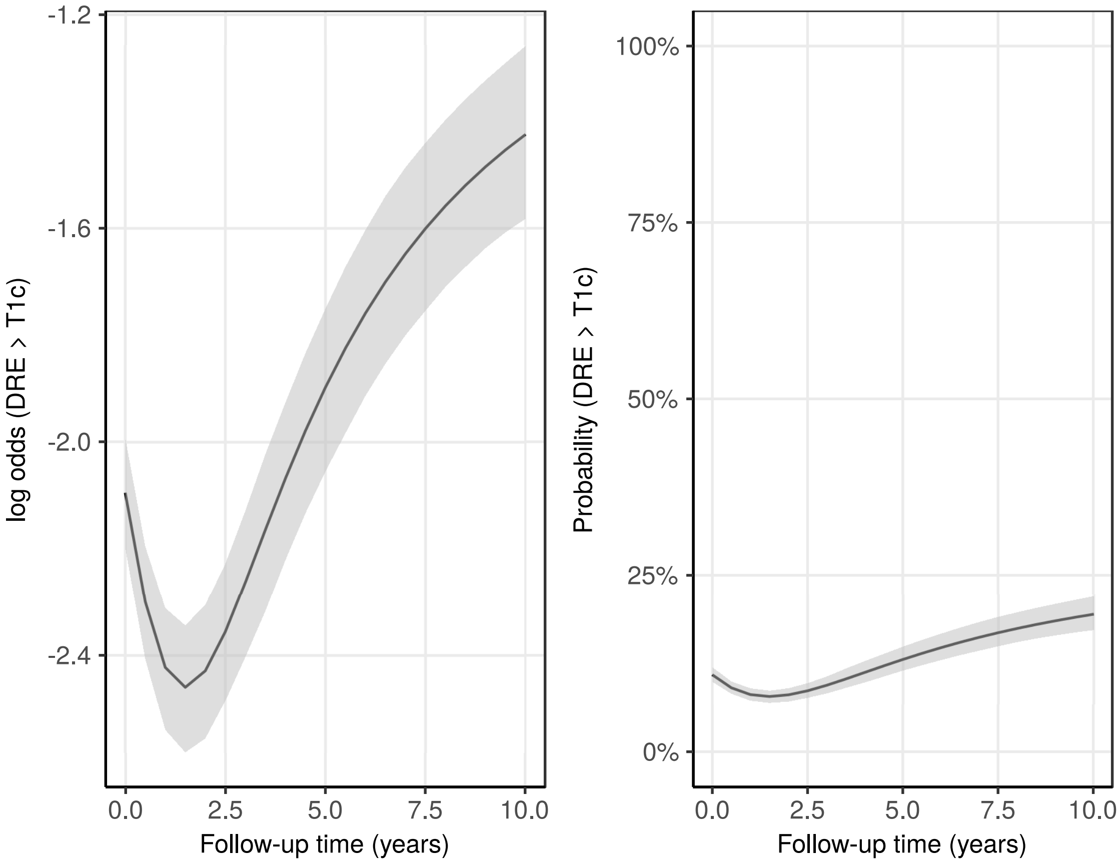

Figure 3: Marginal log odds (left panel) and marginal probability (right panel) of obtaining DRE more than level T1c, with 95% credible interval for a hypothetical patient who was included in AS at the age of 70 years.

paper, I tried normal distribution for error term in the PSA model, and I also tried a t-distribution (df=4). However, the best QQ-plot was obtained with t-distribution (df=3). Any less degrees of freedom will lead to a distribution with infinite variance, in which extreme values of PSA will not be outliers. To make it more clear to our readers that this is the best we can obtain, we will provide a improve the explanation for this plot.

# References

[Cooperberg et al., 2018] Cooperberg, M. R., Brooks, J. D., Faino, A. V., Newcomb, L. F., Kearns, J. T., Carroll, P. R., Dash, A., Etzioni, R., Fabrizio, M. D., Gleave, M. E., Morgan, T. M., Nelson, P. S., Thompson, I. M., Wagner, A. A., Lin, D. W., and Zheng, Y. (2018). Refined analysis of prostate-specific antigen kinetics to predict prostate cancer active surveillance outcomes. *European Urology*, 74(2):211 – 217.