Department of Biostatistics

Erasmus University Medical Center

PO Box 2040, 3000 CA Rotterdam

the Netherlands

February 16, 2018

Professor Anastasios A. Tsiatis

Department of Statistics

North Carolina State University

Raleigh, NC, 27695-8203

USA

Dear Professor Tsiatis, Dear Butch,

We are writing to you with respect to manuscript BIOSTS-15055 submitted to *Biostatistics* 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 several changes in the revised version of the manuscript. In particular, we have invested more in the new version in evaluating the performance of our proposed personalized screening strategy. To this end we have performed a simulation study, presented in Section 5, in which we compared the performance of the utility function $U(u \mid t)$ with the classical strategy that assumes the same screening intervals for all patients. The results suggest that the personalized strategy requires fewer screenings and reaches the optimal intervention time point with less absolute error. Moreover, we now present in a clearer manner the intuition behind the two terms of $U(u \mid t)$, and we refer to the fact that we end up with an informative visiting process requiring an appropriate analysis method. Finally, with regard to the length of the paper, we have attempted to reduce its length as much as possible, without losing the coherence of the main points we are trying to make. The previous version was 25 pages long and you have asked us to reduce that to 20 pages (including title, authorship, abstract, body of manuscript, acknowledgments, references). While according to the suggestions of the Reviewers

we have included new pieces of information, we have managed to reduce the length of the paper to 21.5 pages. We hope that this is acceptable (the number of figures and tables has been reduced to the requested total of five).

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

Yours sincerely,

the Authors

Response to AE's Comments

We would like to thank the Associate Editor for his/her 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 Sections 2, 3 and 5. In addition, changes regarding the specific comments have been made throughout the text.

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

- Evaluation of the proposed method & Cost-Effectiveness:

  1. Following the advice of the AE and the Reviewers, we have invested more in the new version of the paper in evaluating the performance of our proposed personalized screening strategy. In particular, the analysis of the Aortic Valve dataset, presented in Section 4, illustrated that the utility function $U(u \mid t)$ effectively adapts screening intervals depending on the characteristics of the patients' longitudinal trajectories. However, due to the fact in this dataset the data have been already collected under a fixed-planning design, we cannot directly assess the advantages of planning measurements using the framework of Section 3. To this end we have performed a simulation study, presented in Section 5, to evaluate the performance of $U(u \mid t)$.

  2. The setting we considered was motivated by the Aortic Valve dataset. In particular, we assume that patients undergo a valve transplantation and we follow their progression after this operation using the longitudinal outcome (aortic gradient). The aim is to effectively plan re-operations (i.e., re-operations can be considered the 'treatment' in this case). We assume that at any particular follow-up time t, the cardiologist will request a re-operation if the conditional survival probability $\pi_j(t + \Delta t \mid t)$ falls under a specific threshold κ. In our simulation we took $\Delta t = 5$ and $\kappa = 0.8$. The goal is to select the screening strategy that follows each individual patient both efficiently and effectively.

  3. Cost: Assuming that the monetary cost of each examination remains (roughly) the same during follow-up, we would prefer a procedure that requires the fewest possible screenings.

3

4. <u>Effectiveness:</u> On the other hand, we would also like to get as close as possible to the optimal time point at which a medical action is required. Following the advice of the AE, we assess effectiveness based on the accuracy of survival risk prediction. In particular, we first find for each patient the *true* optimal intervention time point t_j^{opt} as the time point for which the *true* conditional survival probability $\pi_j^{true}(t_j^{opt} + \Delta t \mid t_j^{opt})$ equals κ. Following, for each patient we compare two screening strategies, namely, the standard fixed-intervals screening, in which all patients come back at the same fixed intervals (namely every two years as in the Aortic Valve study), and our personalized screening strategy. For each strategy we also find the optimal re-operation time point and compare it to the true one. More details on the exact manner in which this study was performed can be found in Section 5 of the paper and Section 3 of the supplementary material. The results, presented in Figure 2, showed that the personalized strategy requires fewer screenings and reaches the optimal intervention time point with less absolute error.

- <u>Informative new measurement:</u> The AE is right that under our proposed personalized screening strategy, the timing of the next measurement depends on all previous measurements. This informative visiting process would require that the recorded data should be analysed under an approach that accounts for this feature. For example, if the data would be analysed under a full likelihood-based method, such as the joint modeling framework we have used, then the visiting process can be ignored provided that the model is correctly specified. On the other hand, if a method of moments analysis method would be chosen (e.g., a GEE approach), then the visiting process needs to be taken into account. We now mention this issue in the discussion.

  That being said, we do not see any problem planning multiple screenings per subject using our proposed approach because in doing so it is not required to refit the model. That is, the planning of any subsequent longitudinal measurement for subject j is based only on the posterior $[\boldsymbol{\theta} \mid \mathcal{D}_n]$ (with $\mathcal{D}_n$ denoting the original dataset in which the joint model was fitted), that does not update with any new measurements for this or any other future subjects.

Response to 1st Referee's Comments

We would like to thank the Referee for his/her 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 Sections 2, 3 and 5. In addition, changes regarding the specific comments have been made throughout the text.

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

1. Indeed, in the standard context of screening in which Markov and multistate models are typically applied, often the same screening intervals are assumed for all patients. What is investigated in this case is, for instance, what are the extra costs of reducing the screening interval from five years to two years for everyone, and what would the benefits of doing so (i.e., how many more patients we treated correctly). In this context microsimulation is often used. As we discuss in the following point, with our method we can do something similar, but with personalized screening intervals. In addition, these models are based on the Markov assumption that does not allow to utilize the whole available longitudinal history.

2,3. Evaluation of the proposed method & Cost-Effectiveness:

   (a) Following the advice of the AE and the Reviewers, we have invested more in the new version of the paper in evaluating the performance of our proposed personalized screening strategy. In particular, the analysis of the Aortic Valve dataset, presented in Section 4, illustrated that the utility function $U(u \mid t)$ effectively adapts screening intervals depending on the characteristics of the patients' longitudinal trajectories. However, due to the fact in this dataset the data have been already collected under a fixed-planning design, we cannot directly assess the advantages of planning measurements using the framework of Section 3. To this end we have performed a simulation study, presented in Section 5, to evaluate the performance of $U(u \mid t)$.

   (b) The setting we considered was motivated by the Aortic Valve dataset. In particular, we assume that patients undergo a valve transplantation and we follow their progression after this operation using the longitudinal outcome (aortic gradient). The aim is to effectively plan re-operations (i.e., re-operations can be considered the 'treatment' in

5

this case). We assume that at any particular follow-up time t, the cardiologist will request a re-operation if the conditional survival probability $\pi_j(t + \Delta t \mid t)$ falls under a specific threshold κ. In our simulation we took $\Delta t = 5$ and $\kappa = 0.8$. The goal is to select the screening strategy that follows each individual patient both efficiently and effectively.

(c) <u>Cost:</u> Assuming that the monetary cost of each examination remains (roughly) the same during follow-up, we would prefer a procedure that requires the fewest possible screenings.

(d) <u>Effectiveness:</u> On the other hand, we would also like to get as close as possible to the optimal time point at which a medical action is required. Following the advice of the AE, we assess effectiveness based on the level of accuracy the of the estimated survival probabilities. In particular, we first find for each patient the *true* optimal intervention time point t_j^{opt} as the time point for which the *true* conditional survival probability $\pi_j^{true}(t_j^{opt} + \Delta t \mid t_j^{opt})$ equals κ. Following, for each patient we compare two screening strategies, namely, the standard fixed-intervals screening, in which all patients come back at the same fixed intervals (namely every two years as in the Aortic Valve study), and our personalized screening strategy. For each strategy we also find the optimal re-operation time point and compare it to the true one. More details on the exact manner in which this study was performed can be found in Section 5 of the paper and Section 3 of the supplementary material. The results, presented in Figure 2, showed that the personalized strategy requires fewer screenings and reaches the optimal intervention time point with less absolute error.

4. Due to space limitations and because we have been requested by the Editor to reduce the length of the paper, we have decided to illustrate our proposed method for personalized screening only for the continuous biomarker aortic gradient. The results from the discrete marker were similar to the continuous one, but due to its nature (i.e., dichotomous data contain less information than continuous data) the personalized screening strategy for the dichotomous marker was less adaptive than for the continuous one. Nonetheless, the Reviewer is right that a relevant question is how our proposed method can be extended to handle multiple markers at the same time. We are currently investigating how this could be achieved.

We have opted not to include such an extension in the current paper, because it falls outside its scope, which is to introduce our basic ideas for personalizing screening strategies.

Response to 2nd Referee's Comments

We would like to thank the Referee for his/her 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 Sections 2, 3 and 5. In addition, changes regarding the specific comments have been made throughout the text.

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

- Sections 2 & 3: Following the advice of the Reviewer we have combined Sections 2 and 3 of the previous version into one shorter section. The omitted information (i.e., the details of the joint modeling framework) have been put in Section 1 of the supplementary material. In addition, in Section 4 that describes the analysis of the Aortic Valve dataset we have removed Tables 1 and 2 from the main paper and placed them in the supplementary material, and we have accordingly reduced the text.

- Evaluation of the proposed method & Cost-Effectiveness:

  1. Following the advice of the AE and the Reviewers, we have invested more in the new version of the paper in evaluating the performance of our proposed personalized screening strategy. In particular, the analysis of the Aortic Valve dataset, presented in Section 4, illustrated that the utility function $U(u \mid t)$ effectively adapts screening intervals depending on the characteristics of the patients' longitudinal trajectories. However, due to the fact in this dataset the data have been already collected under a fixed-planning design, we cannot directly assess the advantages of planning measurements using the framework of Section 3. To this end we have performed a simulation study, presented in Section 5, to evaluate the performance of $U(u \mid t)$.

  2. The setting we considered was motivated by the Aortic Valve dataset. In particular, we assume that patients undergo a valve transplantation and we follow their progression after this operation using the longitudinal outcome (aortic gradient). The aim is to effectively plan re-operations (i.e., re-operations can be considered the 'treatment' in this case). We assume that at any particular follow-up time t, the cardiologist will request a re-operation if the conditional survival probability $\pi_j(t + \Delta t \mid t)$ falls under

8

a specific threshold κ. In our simulation we took $\Delta t = 5$ and $\kappa = 0.8$. The goal is to select the screening strategy that follows each individual patient both efficiently and effectively.

3. <u>Cost:</u> Assuming that the monetary cost of each examination remains (roughly) the same during follow-up, we would prefer a procedure that requires the fewest possible screenings.

4. <u>Effectiveness:</u> On the other hand, we would also like to get as close as possible to the optimal time point at which a medical action is required. Following the advice of the AE, we assess effectiveness based on the level of accuracy the of the estimated survival probabilities. In particular, we first find for each patient the *true* optimal intervention time point t_j^{opt} as the time point for which the *true* conditional survival probability $\pi_j^{true}(t_j^{opt} + \Delta t \mid t_j^{opt})$ equals κ. Following, for each patient we compare two screening strategies, namely, the standard fixed-intervals screening, in which all patients come back at the same fixed intervals (namely every two years as in the Aortic Valve study), and our personalized screening strategy. For each strategy we also find the optimal re-operation time point and compare it to the true one. More details on the exact manner in which this study was performed can be found in Section 5 of the paper and Section 3 of the supplementary material. The results, presented in Figure 2, showed that the personalized strategy requires fewer screenings and reaches the optimal intervention time point with less absolute error.

- <u>Intuition behind the utility function:</u> We acknowledge the comment of the Reviewer that in the previous version of our paper the intuition behind our proposed utility function $U(u \mid t)$ was not totally clear. Following his/her advice we have now included more details in Section 3.1 to explain how this function works. In particular, our aim is to select the time u to plan the next longitudinal measurement for a biomarker. To achieve this we need to balance two aspects, namely, we would like to maximize the information we gain by measuring $y_j(u)$ at this time point, provided that the patient was still event-free up to u. The two terms of the utility function $U(u \mid t)$ aim at these two aspects. More specifically, the first term is the expected Kullback-Leibler divergence between the posterior predictive conditional distributions with and without this extra measurement. The larger this divergence is, the

more information we expect to gain by measuring the longitudinal outcome at u. However, when the true event actually occurs before u, i.e. $T_j^* < u$, then $\text{EKL}(u \mid t) = 0$, because we did not have the chance to measure $y_j(u)$. On the other hand, when $T_j^* > u$, then, heuristically, as u is set further away from the last time point a longitudinal measurement was taken, the more information we would expect to gain regarding the shape of the longitudinal profile of subject j. The value of u that maximizes this information gain will depend on the characteristics of the subject's profile. As we just mentioned, $\text{EKL}(u \mid t)$ penalizes against selecting $u > T_j^*$; however, the fact that we still need to wait up to time u when $u < T_j^*$ also means that we inevitably increase the risk that the patient will experience the event. Because we would not like to wait up to a point that it would be too late for the physician to intervene, we need to further penalize for waiting up to u. This explains the inclusion of the second term in the utility function.

- Information by measuring the biomarker / first term of the utility function: The Reviewer asked whether the information we are gaining is from measuring the longitudinal outcome or simply from the fact that the patient was alive up to time point u. To answer this we can rewrite first term of the utility function as:

$$\text{EKL}(u \mid t) = E\left\{ \log \frac{p\big(T_j^* \mid T_j^* > u, \big\{\mathcal{Y}_j(t), y_j(u)\big\}, \mathcal{D}_n\big)}{p\{T_j^* \mid T_j^* > u, \mathcal{Y}_j(t), \mathcal{D}_n\}} \right\},$$

where the expectation is taken with respect to the joint predictive distribution $[T_j^*, y_j(u) \mid T_j^* > t, \mathcal{Y}_j(t), \mathcal{D}_n]$. In this formulation and fixing the time point u, $y_j(u)$ can be regarded as a latent variable which is integrated out. Hence, we compare two marginal distributions, the numerator in which we integrate over $y_j(u)$, and the denominator, which is defined already marginally defined. This has similar flavor as what we do with random effects for longitudinal data, i.e., $p(\boldsymbol{y}_j) = \int p(\boldsymbol{y}_j \mid \boldsymbol{b}_j) p(\boldsymbol{b}_j) d\boldsymbol{b}_j$. Due to the correlation in the repeated measurements, the marginal distribution obtained by integrating over the random effects would indicate better fit than the marginal distribution based on a simple linear regression which ignores the correlation.

- Results from the application: As we mentioned above, the purpose in using our proposed personalized screening strategy in the Aortic Valve dataset was in order to show that the utility function $U(u \mid t)$ effectively adapts screening intervals depending on the characteristics of the patients' longitudinal trajectories. In addition, actually the suggestion of the Reviewer

regarding planned versus unplanned re-operations motivated us to focus on the true optimal re-operation in the simulation study. As discussed above and in Section 5, the personalized screening strategy got much closer in absolute error to the true optimal time than the fixed-intervals strategy.

Regarding the gold standard of screening for this type of patients, there is no general rule or guideline. Usually in the first 10 years echocardiographic screening is done every year or every other year. In the second post-operative decade it is generally recommended to have echocardiographic follow-up every year, since the median time to a re-operation for valve failure for these valve substitutes is 15-20 years, and the failure hazard increases with time. Clinical practice is however very variable, with many patients told by their cardiologist after 10 years of follow-up that they no longer needed to be followed because they were doing so well. This ambiguity actually motivated our research.