Re: Shared Decision Making of Burdensome Surveillance Tests Using Personalized Schedules and Their Burden and Benefit.

Summary

This manuscript built joint models for multivariate longitudinal outcomes and time-to-event data and create personalized schedules by planning tests on future visits. I have the following comments for the authors.

Major Comments

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?

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.

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

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.

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.

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

7. In Section 4.1, the automatic chosen threshold $k^*(v)$ should be listed. The details of how $k^*(v)$ 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.

8. In Figure 6, what do the red dots represent? We can see that $k^*(v)$ did not outperform the simple PRIAS schedule. Among the progressing patients, $k^*(v)$ and PRIAS have similar number of biopsies, but $k^*(v)$ has higher median time delay. Actually, $k^*\{v|E(D) \leq 0.75\}$ and $k = 10\%$ have either higher number of biopsies or longer time delay, as compared to PRIAS. The advantage of the proposed method is questionable.

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?

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.

Minor Comments

1. Supplement P23, l34, 9 should be Table 9.