Author’s responses to editors of “An Extension of Generalized Estimating Equation Method to Model Longitudinal Medical Cost Trajectory with Medicare Claims Data Linked to SEER Cancer Registry” (AOAS2107-048)

We are grateful for the positive feedback and insightful comments from the editor. We have responded to each comment and revised the paper according to the suggestions. We use blue font to highlight the parts of the paper that contain revisions to the text. This letter includes itemized responses to the comments. The comment is quoted in italicized font, and our response follows each comment.

Response to Editor

1. “Thank you for a thoughtful revision. We all appreciate your careful responses to all comments. In particular, the responses to the AE’s comments are very thoughtful and comprehensive. We all learned a lot from your paper and responses.”

   **Response:** We are encouraged by the editors’ enthusiasm about the work, and appreciate your effort on our paper.

2. “The AE had one final suggestion that the responses to 1(d) (prediction discussion) and 4 (precision) can be added into their discussion (under reviewer 2).”

   Following the editor’s suggestion, we added the following text to the second paragraph in the Discussion section:

   As pointed out by an anonymous reviewer, the efficiency of the estimation equations may not be a critical concern when sample sizes are huge, however, the sample sizes of subpopulations or cancer subtypes data may be limited (e.g., certain minority populations), geographical locations, or rare cancer types/treatments). Thus, the model for variance could be useful to improve efficiency. We propose our method as a general methodology that works beyond the claims data application. If the sample size is large enough, and efficiency is not of concern, we may assume constant variance, which is also a special case of the proposed method (GEE1-Y).

   Our goal is not to predict future cost at the individual level, because the prediction depends on the individual’s health, behavioral, and financial conditions, which are time-varying and usually not available in an insurance claim database. In this paper, health policy makers are more interested in population-averaged cost trajectory estimates. We are not able to observe the actual survival time for censored subjects, however, the longitudinal cost data of censored subjects contribute to estimating the longitudinal trajectory conditional on survival via equation (5). While the focus of this paper is not to predict future costs for currently surviving patients at an individual level, the proposed model coupled with the estimated survival function
can predict the future cost for currently surviving patients at the population level. Developing new methods to predict an individual cost trajectory is one direction of our future work. In addition, the literature showed that the average cost trajectory of a cancer patients is U-shaped with the peak and valley of the “U” depending on survival times through descriptive analyses (Brown et al., 2002), while directly averaging across cost data from all patients cannot reveal this important feature of cost data. Last, the cumulative or total cost can be easily derived through the estimated longitudinal trajectory conditional on survival, but the reverse is not true.

3. “Also make sure all figures are not in color.”

Response: We converted Fig 1, 2, 3, 4 in the manuscript, and Fig L1 in the Web-based Supplementary Materials into gray-scale. We also updated the text without referring to colors. For your convenience, we include below the all updated figures.
Figure L1: Descriptive analysis of the monthly medical costs in SEER-Medicare data. (a) The trajectory of average monthly costs among subjects who died at selected months. A random sample of 50 uncensored subjects was selected and their individual monthly cost data are plotted in the background as gray lines. (b) The histogram of monthly costs per 1,000 US dollars. (c) The relationship between the mean and variance of the cost data, stratified by survival and by cancer stages determined at the time of diagnosis. (d) The estimated survival distributions by cancer stages. A substantial proportion of individuals with local regional stage are censored by the end of follow-up.
Figure L2: SEER-Medicare data application results for the estimated monthly cost trajectories $\hat{\mu}(t, s)$ when the survival time (month) equals $s = 24, 48, 72, 96$ and 120+ (LTS). The results are compared between local/regional stage (solid) and distant stage (dashed). The shaded areas are 95% confidence intervals.
Figure L3: SEER-Medicare data application results for the estimated monthly cost trajectories \( \hat{\mu}(t, s) \) in 2D heatmaps and 3D surfaces. The results are compared between local/regional stage (left) and distant stage (right). The estimates of the LTS group are displayed separately in each plot. White contour curves are shown on the 3D plot at $3000, $6000 and $9000 elevation.
Figure L4: Simulation results for mean cost trajectories given survival time $\mu(t, s)$ at $s = 0.4, 0.6, 0.8$ based on (a) Gamma and (b) zero-inflated Gamma cost data; and (c, d, e) the penalized spline estimation of the survival function. Sample size $n=500$. The true trajectories are plotted in black curves; on the top panel the mean estimated trajectories are plotted in dotted (EM), dashed (GEE1-N), long dashed (GEE1-Y) and lined (GEE2-Y) curves. Mean average absolute error (AE) and mean squared error (MSE) for $\mu(t, s)$ are averaged across all observations within each dataset, and then over 1000 simulation replicates. On the bottom panel the solid curve is the true survival function. The dashed curve is the estimated survival function, averaged over all 1000 Monte Carlo repetitions. The gray curves are all estimators.
Figure L5: (Figure L1) SEER-Medicare data application results for estimated monthly cost trajectories $\tilde{\mu}(t, s)$ at survival months 24, 48, 72, 96 and for LTS (120+). Estimates at local/regional stage (solid) are plotted against estimates at distant stage (dashed). The shaded areas are 95% confidence intervals.