Review

The authors have successfully addressed a majority of the comments and questions of the previous review, in particular by providing a link to cohort effects, which has significantly improved the manuscript. With that said, the manuscript still lacks a satisfactory description of the out-of-sample predictive performance assessment procedure (related to comment 7 and 8 in the previous review). At the present state of the manuscript, the details are not clear (to me). I therefore have the following comments, questions, and suggestions:

1. Regarding your clarifying paragraph in Section 3.3:

   (a) To my understanding, the dimensions of each matrix corresponds to ages (0–95) and calendar years (20 years in total, window depending on scenario). You have a matrix for each scenario, country, model, and accuracy measure. The values in e.g. Table 4.1 for each model, country, and accuracy measure, appear by averaging over all scenarios and entries in the accuracy error matrices.

   (b) I suggest you address these details in the manuscript (you have addressed this partly in the revision report).

2. Regarding the elements of the matrices of accuracy errors, these are based on forecasts of and "observed" life expectancies (as also shown in the figures, e.g. Figure 4.4). It remains unclear how you arrive at the values in e.g. Figure 4.4 based on the data set and model output; you only allude to this in your second footnote.

   (a) Your data set consists of life table death counts. How are these mapped to the 'Training Set' and 'Validation Set' of e.g. Figure 4.4?

   (b) I suggest you provide details, preferable in rigorous mathematical writing, clarifying exactly which quantities are compared and how you arrive at these based on your model output and data set.