APPENDIX E COMPUTATIONAL DETAILS

E.1 Introduction

A large amount of data was provided to us by MSD. This presented a number of computational challenges that applied to all the main stages of the project:

1. Processing the data to make it amenable for modelling;
2. Fitting models; and
3. Applying models to project future client numbers and cash flows.

The third point for the key benefit liability was particularly intensive computationally. In this appendix we give some detail of how this was done, plus some briefer comments on each of the other stages.

E.2 Projections for the key benefit liability

E.2.1 Current client liability

As discussed in Section 4.3.7 of the report there are many possible combinations of variables in the projection data set. Excluding calendar quarter, there are 10 variables in the projection data set:

- Age of client (196 levels);
- Duration on current benefit (196 levels);
- Benefit history flags (indicators for whether the client has received the benefit in question [DPB, UB, SB, IB, OTH or NOB] at some point in the past. There are 6 flags with 2 possible values leading to 64 levels in total for all flag combinations);
- Children variables (30 levels – number of children [3 levels] x age of youngest child [10 levels]).

This means that there are potentially almost 74 million possible combinations for each of the six different benefit types leading to over 440 million states in total for each projection quarter. Therefore it was necessary to find ways to reduce the computational load by reducing the number of possible combinations of the variables. The following decisions were made to reduce the computational burden:

- Children variables were retained for DPB only. Thus, the maximum number of combinations for each projection quarter for a benefit other than DPB is 2.5 million;

- Capping of duration for DPB was implemented:
    - The DPB models do not have duration effects which discriminate between durations above 80 (unlike, for example, the IB models). Therefore the number of DPB duration levels was reduced to 84 (to facilitate seasonality where present). This reduces the number of possible combinations to 31.6 million;

- When computing the projected liability, the data was broken down into data sets containing a small number of distinct ages, usually a single age cohort (apart from those close to retiring age). For DPB, this reduces the number of possible combinations in each subset to a more manageable 161,000 and even lower (maximum of 12,500) for the other benefit types.

- Various time-saving programming methods were used to speed up the calculations from the efficient use of memory and other programming techniques, to distributing the projection job over a number of different computers and running up to 8 processes simultaneously on each computer.

Even with all these measures in place, the projection is laborious, taking approximately 60 computing hours for the current client liability. Distribution of the age cohorts allowed this to be reduced to about one and a half hours real time:

- About eight computers were linked together to run different sets of age cohorts; and

- Eight age cohorts were run simultaneously on each computer;

Distribution was achieved using custom software built at Taylor Fry for distributing SAS sessions.

E.2.2 Future client liability

The future client liability projection is even more computationally intensive than the current client liability since the liability must be estimated for 20 different cohorts for each of the future quarters rather than one as is the case for the current client liability (those who have received a benefit in the last 12 months as at the valuation date).

To reduce the computational burden, each **future year** was projected rather than projecting each quarter separately. Thus, for both components of the future client liability (those who have been off benefits for between 1 and 10 years as at the valuation date and those new to the system or off benefits for more than 10 years), the numbers coming on benefits for each quarter in the future year were estimated, projected forward to the end of that year and then combined meaning that thereafter, the year was projected in aggregate.

In addition rather than allowing all possible ages to be present in the data newcomers were assigned ages at one year intervals as follows:

- September quarter: client ages are 16, 17, 18, …, 64;
- December quarter: client ages are 16.25, 17.25, 18.25, …, 64.25;
- March quarter: client ages are 16.5, 17.5, 18.5, …, 64.5;
- June quarter: client ages are 16.75, 17.75, 18.75, …, 64.75.

A consequence of this is that at the end of the future year, all clients are aged 16.75, 17.75, 18.85, …, 64.75. Therefore, rather than 196 distinct ages for projection, there are 49 which greatly reduces the computational load.

The calculation of future client liability takes approximately 6.5 hours when distributed over a number of computers. We note that there are some further time savings possible – for example, by not separating out the liability for each of the future years or by not separating the two components of future client liability.

E.3 Other computational considerations

E.3.1 Modelling the probability of receiving benefit for minor benefits

Section 4.5.2 of the report discusses the probability models used for each of the minor benefits. The full datasets generated for these models are very large – GLM fit times for them was between 1 to 2 hours, too long for our iterative modelling approach. Thus for these models:

- A representative stratified subsample was taken of about 10% of the data. This was used to test effects and determine a final model structure; then
- The model was refit on the full dataset to obtain a more accurate probability model.

This allowed computation time to be manageable.

E.3.2 Modelling transition probabilities for key benefits

The modelling datasets for some of the benefits were particularly large, notably the probability of remaining in the same state for UB, OTH and NOB. This was handled by means of stratified sampling, where the rarer response was sampled at a higher rate to the common response to minimise the corresponding decrease in accuracy. Observations were weighted to ensure the overall rates of transition remained correct.

This approach was used in cases where the available data was already very large, and so the potential impact on model performance was immaterial.

E.3.3 Data preparation

Processing the original datasets to convert them to a form amenable to modelling took a reasonable amount of computer time, perhaps around 10 hours to produce modelling

datasets for each of the benefit types. Given this needs to be run just once, this was judged acceptable and was not further optimised or distributed.

E.3.4 Minor benefit projection

The minor benefit projections ran significantly faster than the key benefit liabilities, largely due to the fact that they were modelled as independent payment streams. Total runtime on a single thread was approximately 2 hours for both current and future client liabilities for all minor benefits, which was judged acceptable.

E.3.5 GLM fitting in SAS

We use a suite of custom-built SAS macros to carry out all GLM fitting, model diagnostics and validation. These macros substantially extend the available tools within SAS as well as optimise the use of SAS's inbuilt GLM fitting capabilities.