45 and Up Study Technical Note 1: Missing or Invalid Values

Like all large cohort data sets, there are some missing or invalid values in the 45 and Up Study data. While it is important that you know about the missing and invalid values, experience to date suggests that they are unlikely to have a material effect on findings obtained from 45 and Up Study data analyses. For example:

a. Where results obtained from the 45 and Up Study have been directly compared with results obtained in other studies, they have been very closely comparable (see, for example, Mealing NM et al. Investigation of relative risk estimates from studies of the same population with contrasting response rates and designs. BMC Research Methods 2010, 10: 26).
b. Journal editors and reviewers have generally been satisfied with the approaches used to address missing or invalid values when assessing papers reporting results of 45 and Up Study analyses.

So far, more than 450 researchers have worked with 45 and Up Study data on more than 120 separate projects, and more than 50 scientific papers have been published from their analyses. Researchers have rarely raised concerns with us about missing or invalid values. If you have a different experience please let us know by email to 45andUp.research@saxinstitute.org.au or by contacting the Study Infoline on 1300 45 11 45.

Information on the combined frequency of Missing or Invalid Values is available in The 45 and Up Study Baseline Questionnaire Data Book, which is available through the 45 and Up Study website - www.45andup.org.au. The following paragraphs provide some additional information about missing and invalid values and what we know about the origin of invalid values.

Missing values

Missing values are indicated by a full stop (.) or blank ( ). They occur when a response is not required because it was conditional on the response to a preceding question, and when Study participants failed to put a cross or number in a box when they should have. Only the latter represent a data deficiency. We have no way of remedying these deficiencies except by going back to participants and requesting the missing data. We have not done this and do not intend to do it systematically. It is possible, though, that some missing data will be supplied through the periodic follow-up surveys, the first of which began in 2012.

Invalid values

Invalid values occur when participants wrote in an implausible value for the response in question, perhaps because they misunderstood the question, or when data-processing operators incorrectly transcribed software-indecipherable responses. Invalid values that fail range checks, which we apply to address the problem of implausible values (whether because of participant misunderstanding or

operator error), are coded 99999 for numeric response fields and X for alphanumeric fields. The invalid ranges applied in these range checks can be found in the 45 and Up Study Data Dictionary Summary in the column headed Invalid Ranges. Please refer to http://www.45andup.org.au/helpanalysingdatafromthe45andupstudy.aspx for the most recent version of this document, which is periodically updated.

Some correction of invalid values is possible because it is known from targeted checks on data quality that a sizeable proportion of them are due to operator error. This checking has entailed selection of invalid values and consulting scanned copies of questionnaires in which they occurred. In the sample checked, the invalid values were equally distributed between transcription errors and out-of-range responses given by participants themselves.

What should you do about missing or invalid values?

To date, most researchers using 45 and Up Study data have simply excluded records with missing or invalid values for important variables from their analyses; and have used a range of approaches to deal with missing or invalid values for less important covariates. They have also declared this action in the description of their study methods, usually in a sentence or two that describes the total number of records available for the analysis and the number actually used for the analysis after excluding variables with missing or invalid values. Because of the size of the data set, this approach has generally not had a material impact on the statistical precision of the results of the analysis.

Bias due to data being systematically, not randomly, missing or invalid, however, is a possibility that researchers should consider and which could be important for some variables. If there is such a concern, it is probably best addressed by developing a multiple imputation model for the missing or the invalid data and imputing them. To do this, however, requires substantial biostatistical expertise.

If researchers are concerned about the possibility of bias in their results due to missing or invalid values of 45 and Up Study variables, they should:

- Check the frequency of missing or invalid values for variables of importance to their research by consulting the most recent 45 and Up Study Baseline Questionnaire Data Book when planning their research; and
- If this information is not sufficient, contact the 45 and Up Study through its Infoline, 1300 45 11 45, or by email at 45andUp.research@saxinstitute.org.au. Ask to be put in contact with an expert who can help work out what additional data about the missing or invalid values might help in assessing their importance to you and what action could be taken to minimise any problems they might create.