Dear all,

I wanted to follow-up on our discussion this past Sunday about the upcoming analyses of linked datasets.

I would like to suggest the following plan.

Before anything, we need to make a table describing study/interview outcomes. Starting from the sample file, we should describe what happened to each individual selected. Then, we should compare the characteristics of those who were interviewed vs. those who were not interviewed. These are characteristics according to the HDSS, for example, their residence, gender and age differences.

Then, turning to the analysis of the survey datasets, I suggest that we start with the analysis of data on parents, since these are a little easier to handle than data on siblings. This will make analyzing the siblings data easier later.

Here are some tasks/analysis I would suggest re: parental data:

1) making the dataset long, i.e., with two rows per respondent (1 for the mother, 1 for the father)

2) renaming variables about parents with similar names but different suffixes, depending on their source (survey or HDSS). For example, the variable describing the vital status of each parent according to the survey could be vitalstat\_ssh, whereas the same variable according to the HDSS could be vitalstat\_hdss

3) comparing the characteristics of parents reported during the survey who could be linked to the HDSS to those who could not be linked. These are the characteristics reported in the survey, e.g., vital status, reported age of live parents, reported age at death/time since death.

4) measuring the concordance in vital status between the survey and HDSS datasets. This means cross-tabulating vitalstat\_ssh and vitalstat\_hdss if we use the names I mentioned earlier. Then we can describe the characteristics of those with discordant vital status (e.g., survey says father is alive while HDSS says father is deceased).

5) measuring the % of parents with missing data in the survey on age (if alive) or age at death/time since death (if deceased), and testing if this varies between mothers and fathers, and linked and unlinked to the HDSS.

6) for the subset of parents who could be linked to the HDSS, measuring the concordance in reported ages and dates. This could be done with scatterplots showing ages/dates according to HDSS on x-axis and age/dates according to survey on y-axis.

7) for the subset of parents linked to the HDSS, investigating the determinants of differences in reported ages/dates using regression models.

That is quite extensive, but this will take us to a nice set of analyses about parental reporting. We can then adapt our scripts to the analysis of data on siblings.

Please let me know what you think, and we can discuss more on a call, and divide tasks as well.