The Health Survey for England -- Preserving consistency over multiple data sources

Sven Sjödin,
SCPR (UK)

1. About the Survey

The Health Survey for England is sponsored by the Department of Health. It is a continuous survey conducted every year throughout the year. A new sample is issued every month. The monthly sample size for the 1998 survey is 1140 addresses. The survey started in 1991 and has been running continuously since 1993. The SCPR has conducted the survey since 1994 in collaboration with the University College of London.

The Health Survey for England is a household survey in the sense that most of the household members are eligible for interview. The current rules for eligibility allow for ten adults and two children aged two to fifteen. Each interview starts with a household section to collect household level information. The interviewer will then select up to four respondents for each session of **concurrent interviewing** to gather person level information. The person level data is partly collected through self-completion booklets.

Concurrent interviewing is a method for increasing the efficiency of household surveys. The sessions are built up by a succession of tables that enables the interviewer to progress with more than one respondent at a time. The tables are short sequences of questions which are repeated for each respondent in the session.

On agreement, each respondent will be visited by a nurse. The nurse records various measurements and collects samples. The samples are then sent to a laboratory for analysis. Each sample and measurement requires a signed consent by the respondent. Because of the nature of the survey, there is an obligation to report the medical results to the respondents, who are also asked for their consent for the results to be given to their GP (physician).

Apart from the laptops and any paper documents, the interviewers and the nurses carry a fair amount of equipment. The interviewers measure height and weight using stadiometers and electronic scales. The nurses have Dinamaps for blood pressure readings; various tape measures; equipment for taking blood samples; and straws for saliva samples.

2. Problems in the Old Days

It is of vital importance for the quality of the survey that the **person identifiers** in each part of the data are consistent and correct. The Health Survey for England has an unusual number of data sources that contribute to the final data set. In addition to the main and nurse interviews, there are self-completion booklets, consent booklets and various laboratory results.

This has always generated a vast effort to detect and solve inconsistencies, even once the main questionnaires were converted to CAPI in 1995. All the information passed between the

interviewer and the nurse was on paper. An error could occur each time a serial number or a person number was transcribed. The only measure used to counter this problem was the use of the respondent's date of birth to supplement these internal identifiers. For this reason the date of birth was typically recorded seven or eight times per respondent. Every time an inconsistency was discovered a very complex set of rules was applied to resolve it.

The data collected with the nurse paper questionnaire also required a substantial amount of editing. As every measurement is repeated at least twice there was an obvious risk of error from misreading the equipment, recording the wrong value or recording a value in the wrong place.

3. Nurses going CAPI

The 1998 survey faced the need to minimise the risk of such inconsistencies. By converting the Nurse Questionnaire to CAPI (Blaise III), we expected a significant reduction in the editing work. Most of the edit checks on consistency can be implemented in the data model. There are also the usual advantages of CAPI compared to paper such as enforcing the route; control over value ranges; and the possibility of recording the time spent interviewing. All the nurses are now equipped with laptops and modems and are given the appropriate training.

Converting the Nurse Questionnaire to CAPI reduces the risk of internal inconsistencies. But it does not actually address the problem of inconsistencies between these different sources of data. A mechanism is required to pass data from the interviewer to the nurse in a way that is more reliable than paper forms. The situation is very similar to that of a panel survey. To control for consistency, data from previous waves are sent out for each new wave.

Ideally, the nurse starts each case with all the relevant information already present. The nurse interview is then coupled to the main interview so that, by definition, the identity of each person is consistent between the two.

We set out to achieve this by amending our in-office system. From the incoming main interviews, we generate data records which are made available to the nurses to download on to their laptops. The Survey Management system writes out ASCII
format data files and the appropriate laptop identifier. The ASCII files are then converted to Blaise files for the nurse data model to use as external information.

This **transfer information** is a household level record with a set of person level sub-records. It stores the name, age, sex and date of birth of each eligible household member. For children aged two to fifteen, it also holds information about the parents in the household.

4. Time Is Of The Essence

Our initial approach was to let the transfer mechanism drive the allocation of cases to nurses. That is, the nurses were only given access to the cases that had passed through the in-office system. This hard coupling guaranteed the consistency between the main and the nurse interview.

However, there were are three main time factors that ruled out this approach:

- First, unlike a panel survey wave, the nurse appointment can come very soon after the main interview. An interviewer can make an appointment for the nurse for the very same day. The transfer system needs at least one day to turn over the data and it is not operative over weekends. Therefore, the nurses have to be able to conduct the interview before they receive the transfer information. In such cases the nurse has to enter the corresponding data from a paper form. This means that the nurses have to be issued with the same set of sampled addresses as the interviewers. A further implication is the fact that the nurse data model has to be a household level, rather than a person level, questionnaire.

- Second, the interviewer may not finish the whole household in one single appointment. Some households can take weeks to complete. Our standard procedure is to book in cases only if they have a final outcome code. For the Health Survey for England we also need to be able to process incomplete households in case there is information to pass on to the nurse. The nurse data model will then have information about the composition of eligible respondents, but will not know whether or not each respondent agrees to see the nurse.

- Third, the same time factor applies to these residual respondents as apply to the whole household. The nurse has to be able to start the interview before the final **nurse agreement** data are available. In fact, we also have to allow the nurse to interview respondents who explicitly rejected seeing her at the time of the main interview. They may well have since changed their minds.

The nurse agreement data are sent out as a separate data file which is shared by all the nurses. The reason for this is the fact that, while the transfer information is produced just once per household and requires some manual intervention, the agreement data may arrive at the office at a later date. The initial nurse agreement data, usually blank, are created for each eligible person at the same time as the transfer information. It is then updated automatically each time new respondents have completed the main interview. This information is needed to warn the nurse if she tries to interview a person who has rejected, or yet agreed, to see her.

Whenever a nurse opens a household or selects a person within a household and the data model cannot find the corresponding transfer record or agreement data, the nurse has to override a warning to continue. These warnings are implemented as conditional fields. On the household level, the field also gives the nurse a way of screening out addresses that do not require a nurse visit.

5. The Nurse Data Model

All these complications make the nurse data model very elaborate. It has to cater for several different scenarios. It has to record whether or not it is started in a manual or automatic mode, i.e. if the transfer information is present on opening the case. It also has to recognise when this information subsequently arrives. If the household was opened in a manual mode and the transfer information has since become available it has to cross check the information entered by the nurse against the transfer information.

The table below explains the different scenarios and their implications for the nurse data model. Each scenario starts with the state of the household when the transfer information is downloaded to the laptop.

Table 1. Scenarios and Nurse Data Model Behaviour

1	The household is not yet opened.	The ideal situation. The transfer information is read and stored. No checking is required.
2	The nurse has entered all the household information	The transfer information is read and checked against the nurse entries. These entries are not altered, but any inconsistent person level records are marked out. If so, the nurse is prompted to correct inconsistent person numbers.
3	The nurse has entered some of the household information.	The transfer information is read and checked against the nurse entries as above. Any residual person level records are added to the household composition data.

The nurse data model consists of a short household level section and up to twelve Nurse Schedules for person level data, declared as parallel blocks. The household section contains a household grid which is either filled in by the nurse (manual mode) or automatically from the external data file. It also handles the hidden mode control fields. Each Nurse Schedule is made active in the main RULES section if the nurse has manually entered a nurse agreement code or the person record is found in the nurse agreement external file.

There are also up to twelve parallel drug coding schedules. This is where the nurses code the recorded drug names according to the British National Formulary (BNF) coding scheme. Declared as parallel, they are independent of the RULES section of the Nurse Schedule and can be completed at any point during the interview.

6. Challenges Along The Way

Any Blaise III data model that depends on heavy initial computations and checking has to be programmed with great caution. For the nurse data model it is essential that each step of calculations is completed before the next step can start. The Blaise III **dynamic checking** mechanism, however, will try to execute all the RULES sections in what seems like a simultaneous manner unless it is stopped by appropriate conditions. It has taken a great deal of experimentation to find these conditions.

The most time consuming parts of the Blaise programming were:

- The section in the main questionnaire in which the interviewer selects respondents for sessions of concurrent interviewing; and

- The initial control structures of the nurse data model.

In both cases the difficulties arose out of the need to find ways of stopping the dynamic checking mechanism from setting vital control variables before the previous calculations and interviewer inputs are properly validated.

Both interviewers and nurses can be expected to need more than one appointment to complete each household. This has special implications for the design of the data model. For example, the key routing fields have to be safeguarded so that keying errors on re-entry will not take whole sections of the questionnaire off the route.

A different kind of technical challenge was posed by the nurse agreement data. If it fails to update on the nurse laptops there is a risk that nurses cannot proceed with some person level interviews. The corresponding file on the network is almost constantly in use as it is downloaded each time a laptop connects to the office. This sometimes leaves very short time slots during which it can be updated. It has taken some time both to recognise and to solve this problem.

7. The Future

Although the data has not yet been edited, we believe that the current setup will greatly reduce inconsistencies within the nurse interview data and between the main data and the nurse data, thereby reducing the editing effort.

Looking into the future, are there any ways to further improve the consistency between the multiple sources of data in the survey?

One obvious possibility relates to the self-completion section of the main questionnaire. By turning this into a Computer Assisted Self Interviewing (CASI) section we could eliminate the risk of attaching an incorrect person identifier to the data. This has been considered, but rejected, as it would defy the principle of concurrent interviewing if the laptop has to be passed between respondents and other respondents have to wait for their turn.

It is possible that we can improve the in-office system or use other electronic channels, e.g. the Internet or email, to speed up the data transfer from interviewers to nurses. If this can be achieved we can also return to the ideal situation of only issuing nurses with the appropriate cases and at the more suitable person level.

At the more futuristic end of the scale, we could minimise the risk of inconsistently labelled paper components once laptops have built in label printers or printers are small enough to be portable. This would also work for nurse sample labelling.

In an ever growing market of home medical kits, e.g. pregnancy tests and DNA tests, it is not impossible that one day we will be able to supply the nurses with blood and saliva analysis equipment, thereby cutting out the laboratories altogether.

For the 1999 survey there will be yet more challenges. Part of the sample will be aimed at ethnic minorities with interviews conducted in languages other than English. We have opted for a script based system, rather than the Blaise languages facility, as six extra languages using non-Latin characters would pose too much of a maintenance problem.

To improve the probability of finding ethnic minority households we will apply a method called **focussed enumeration**. For selected sample points the interviewers will call at the three addresses to the right of and the three addresses to the left of the sampled address in search of ethnic minority households.

A related issue is that, for multi-ethnic households, we will need to be able to re-allocate half finished interviews to interviewers with the appropriate language skills.

8. Summary

The Health Survey for England is a survey with multiple sources of data. Apart from the main and the nurse questionnaires, there are self-completion booklets, consent booklets and various laboratory results. There is a clear risk of inconsistent case and person identifiers between the data sources. The task of identifying and solving these inconsistencies is very time consuming and prone to error.

For the survey year 1998 we have tried to reduce the risk of such inconsistencies with special attention paid to the coupling of the main and the nurse interviews. By converting the nurse questionnaire to CAPI and implementing an automatic transfer of data from the main interview to the nurse laptop, we have sought to reduce the risk of inconsistencies both within the nurse interview and between the main interview and the nurse interview.

Some time related factors complicate the system. First, the first nurse visit can follow very shortly after the main household interview, before the household composition data can be transferred. Second, the nurse may want to interview respondents before downloading any information on whether or not they agree to see her.

These factors make the nurse data model very complex. It has to cope with different scenarios depending on the state of the interview when the external data is transferred. It also has to cater for the Blaise III dynamic checking mechanism as this will prematurely try to assign values to vital control fields unless stopped by appropriate conditions.

The way forward is to quicken the transfer of data from the main interview to the nurse. Ideally, this should also drive the allocation of person level cases to the nurses. The risk of inconsistencies can never be completely eliminated, but there are some possible technical developments that may further reduce them.