Description of the survey automated process analytics system

Margo Barr and Raymond Ferguson

Centre for Epidemiology and Research, NSW Department of Health

Introduction

The New South Wales Population Health Survey provides detailed information on the health status and behaviours of people in NSW. Using computer-assisted telephone interviewing (CATI), the survey began in 1997, and became continuous in 2002 [1].

The first report of the survey—containing combined data from 1997 and 1998—required a substantial timeframe, and was not published until 2001. Firstly, all the formats were retyped into SAS [2] from the survey questionnaires. Second, any required changes to the data, as a result of the coding process were made, by typing in a series of IF statements in the SAS programs. Third, data dictionaries were retyped into the format required, and finally, the any errors that were introduced due to this process, were checked and fixed.

Indicators of interest for analysis were programmed as separate unique SAS files. This process was very labour-intensive, as it was repeated hundreds of times. The indicators were then outputted, ready to be included into the report using indicator-specific html templates. No hardcopy report was produced, due to the added cost and time required for desk-top publishing. Therefore, the electronic-only file was produced, by manually creating html pages and manually inserting the hyperlinks for each of the text, graph, and table files.

This highlighted the need for the development of an automated process analytics system [3], particularly when the survey became continuous in 2002. An automated system was then developed to clean, analyze, and report on the survey data. This system utilises the metadata from the CATI questionnaire, and minimizes the need to re-enter any information, using: an automated data management and cleaning system, an automated datamart system, an automated analysis and reporting system, and, an automated quality assurance system.

System Development

Automated Data Management and Cleaning System

Because CATI surveys are collected using programmed questionnaires, all of the metadata about the questions; including the questions themselves, response categories, legal values, and skipping patterns, are available in the questionnaire script. So, the first step in building the automated system was to develop a system to systematically output, not only the response data from the collection system, but also all of the metadata (labels, type, formats etc).

In order for the questionnaire script to compile ready-to-collect data, there are rules for how the questionnaire needs to be programmed. We used these rules and developed others, so that the metadata could be consistently extracted from the questionnaire script. These rules included; having the label name inserted after the variable, inserting codes before the question script, and the response codes.

The rules also require the need for surveys to be registered with unique allowable names.

The CATI software program, 'Sawtooth', used by the NSW Health Survey, compiles the data and outputs five separate files; DATA, TEXT; PROGRAM; LAYOUT; and LISTS for each survey job. The automated data management and cleaning system was developed in SAS to: read the PROGRAM, LAYOUT and LISTS files created by the CATI software, extract the necessary metadata, and create the SAS catalogues of formats.

The system then: reads the contents of the DATA file, merges it with the TEXT file ('open' and 'other' responses), attaches the labels and formats to each variable, creates a unique survey program identifier (using the code allocated to the registered survey name), and removes any extraneous data relating to the survey collection. The resultant RAW dataset for each survey job is outputted. This process can be run on a daily, or a weekly basis, for any survey job (survey quarter, languages, stand-alone survey), so that at any time, there is an up-to-date version of the RAW survey dataset.

Automated DataMART System

Once the survey job is completed, additional data management is required. This includes: allocation and updating of geography to telephone numbers, coding 'open' and re-coding 'other specify' responses, creation of derived variables (conversion of responses to a common parameter), scoring tools, and adjusting for skipping patterns. This also includes weighting of the sample to adjust for differences in the probabilities of selection, and, to the population benchmarks.

Although it would have been ideal to also automate coding of the 'open' and re-coding of the 'other specify' responses, it was not possible, as they require a certain amount of interpretation. The process was able to be incorporated however, by automating the downloading and uploading of the 'open' and 'other specify' responses, and then by manually coding, and/or re-coding in excel spreadsheets, with appropriate automated checks for completeness.

The automated DataMART system was developed using generic SAS programs and macros. These programs again utilised the metadata files. Thus, the MART dataset produced was a combination of the RAW dataset, the updated geographical variables, the updated recoded 'other specify' responses,' the coded 'open' responses, the newly created derived variables, and, the weighting variables.

Automated Analysis and Reporting System

In order to develop the automated analysis system, initial decisions were made on: the infrastructure within which the analysis was to be conducted, what analysis would be included in the system, and how the data would be outputted. A master setup program was developed, so that anyone using the system could use it consistently. Standard indicator and response definition files were also developed.

This master setup program pulled in a generic html template and all the infrastructural macros including: file, system setup, dataset creation, statistical, graph, table, output, reporting, analysis, management, and validation.

A system driver was also developed that defined the requirements for each type of study and/or report being undertaken in a format that could easily be edited or read, using different analysis packages. We chose to edit the driver in excel and to convert it to an xml file for storage and ease of use. This system driver listed all information required across all years and reports, including: the indicators and reporting variables—with their titles, footnotes, age bands, and graph and table footnotes, which indicators will be in which reports, and the order in which the indicators are included in reports. The system driver was designed so that new question modules, analysis methods, and reporting outputs, could easily be incorporated to meet the emerging and changing information needs of users.

The outputs from this system for each study and/or report include:

- folder structure for the particular set of analysis,
- SAS datasets containing respondent numbers, weighted numbers, prevalence estimates, standard errors and 95% confidence intervals for each of the parameters of interest (overall, by age, sex, administration area and socioeconomic status),
- graphical outputs stored as gif files,
- csv files from the SAS datasets,
- html and pdf files for each of the indicators including the gif graphical file, a table of the results, description of the indicator, number of respondents, and any other information,
- text pages for each topic area outputted as html and pdf files,
- table of contents with hyperlinks to each of the individual html pages as specified in system driver,
- FINAL dataset and associated data dictionary which included the questionnaire variables, derived variables, weighting variables and indicator variables with variable labels, variable type and name of format, question text, response codes and definitions produced from the metadata,
- a hard copy report, produced without the need for desk-top publishing, using a macro that could add all of the individual text and graphical pdf files in the order specified in the system driver, number the pages, add the page numbers to the pdf version of the table of contents, and, add the cover and imprint pages.

Automated Quality Assurance System

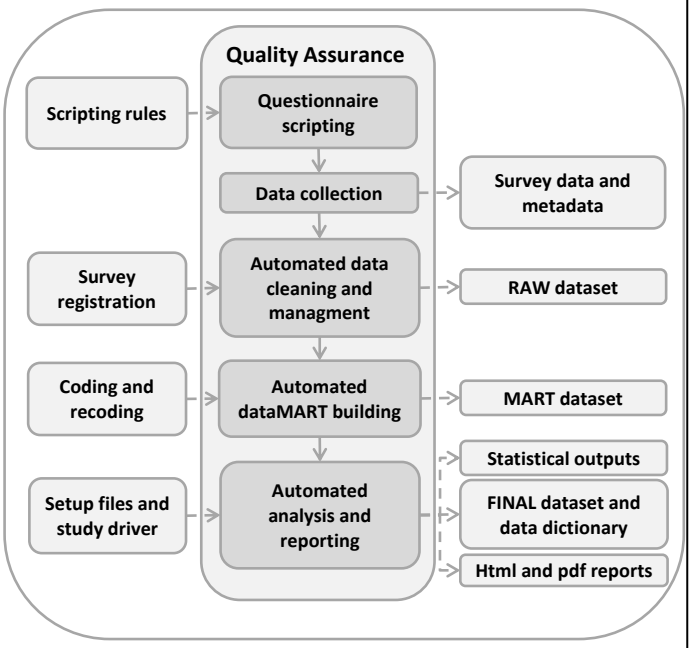
The consequence of automated processes is, that if it does not work, then it may not be obvious to the user. Therefore, it was necessary to develop a quality assurance system which draws attention to any errors during the data cleaning, analysis, and production processes.

The final automated quality assurance system developed includes: emails that are sent when SAS programs are run that include warnings and pertinent information, storage of log files when batch jobs are done, and programs that read them, automated production of data dictionaries that use the CATI metadata, analysis validation reports which produce denominators and analysis both from within and outside the automated system, comparisons of figures in related graphs, comparison of the information in the study driver to that in the produced graphs, and report production management functions such as errors listed in the report tables of contents html files.

Summary

A high level summary of the process analytics system is shown in Figure 1.

Figure 1: Overall Process



The NSWPHS surveillance system therefore has a continuous collection, analysis and reporting process that can be used across different surveys, population groups and topic areas. The surveillance system also maximizes the use of metadata, seamlessly interacts between different IT platforms and software using SAS as the driver, outputs the information in several useable file formats, and, produces the final hard copy report, without the need for costly and time consuming desk-top publishing.

References

1. NSW Population Health Surveys. http://www.health.nsw.gov.au/surveys/Pages/default.aspx
2. SAS Institute. The SAS System for Windows version 9.2.
3. Hellerstein JM. Quantitative Data Cleaning for Large Databases, University of California, Berkeley, 2008 available at db.cs.berkeley.edu