

PhUSE 2012

Paper HE01

Skills for SAS® programmers in Epidemiology

Philip R Holland, Holland Numerics Limited, Royston, Herts, UK

ABSTRACT

SAS® programmers in Epidemiology departments require very different skills to those working on clinical trials, because of the sheer size, complexity and irregular nature of the data used. This paper will investigate the skills required, and suggest that recruiting programmers with skills from outside the pharmaceutical sector may not be such a bad idea.

SIZE REALLY IS IMPORTANT!

A major difference from clinical trials is in the amount of data that a SAS programmer has to be able to cope with. Data sets are routinely measured in gigabytes, not megabytes, so care must be taken to reduce the data volumes as much as possible before performing any processing by only keeping the columns you actually need and by using a subset of the data set records. Even then the sorting of the data should be kept to a minimum, the amount of data stored in WORK should be considered at every step, and any expectation of processing taking minutes, rather than hours, be treated as unlikely.

WORKING WITH LONGITUDINAL DATABASES

SAS programmers in Epidemiology uses longitudinal databases where the patient details, medical events and prescriptions need to be selected separately from different files, and then combined as required. This is similar to that found in clinical trials data, but the medical events and prescriptions here are all at random times. Also, as each disease and therapeutic area has different coding requirements, so there will only be a small amount of reusable SAS code, so, more often than not, a new program will have to be written for each analysis.

If SAS programmers who had previously worked with credit card data were to analyse this data, then they would probably see that:

- The patients are like credit cards.
- All medical events and prescriptions are like credit card transactions.
- Disease and therapeutic areas are like merchant types, or countries, where transactions occur.

NEED TO CLEAN THE DATA

Many of the longitudinal databases were created to collect information from clinical staff using drop-down lists for patient management, either clinical, financial, or both. This results in data that includes quite common mismatches, where very different items have very similar names, as well as data values that should be, but are not, compulsory. The lack of internal data checking can also give rise to anomalies, which are not necessarily critical for patient management, but add to the complexity of the data analysis when processed for studies in Epidemiology:

- The gender can be recorded as male, female, or unknown.
- Medical events and prescriptions can have associated dates before the practice registration date, or after the date of death.
- Prostate cancer and other male-specific conditions can be recorded for females.
- Hysterectomy and other female-specific conditions can be recorded for males.

Depending on the analysis and data selection, these anomalies can be corrected, or the patients removed altogether. However, most of the programming effort will be spent in specifying codes to identify the medical conditions, or prescription drugs, of interest.

CHOOSING A SUITABLE DATABASE

Databases have differing characteristics, e.g. geographical location, age profile, duration of patient registration, etc., so no single database includes the data required for every study in Epidemiology. Therefore the database, or databases, selected can be an important decision when designing a study:

- GPRD, ffGPRD and THIN – UK, GP records, 6m+ patients, representative age sample, average registration 7.5 years, updated monthly.
- Lifelink – US, employee claims database, 1.8m patients, good elderly coverage and poor coverage of 25-34, average registration 6.9 years, no updates after Q1 2003.
- IHCIS – US, managed care database, 20m patients, poor coverage of elderly, average registration 1.5 years, updated 6-monthly.

Database selection must also include what data items are available, as many databases do not contain all types of data:

- Outpatient and hospitalisation records
- Laboratory data
- Smoking data
- Repeat prescriptions
- Height and weight data
- Death information
- Links between family members
- Links between medical events and prescriptions

The selection decision should also take into account the therapeutic area, as some databases have better representation of certain medical conditions, due to their patient selection:

- Lifelink has better coverage of Type II diabetes than IHCIS due to the higher coverage of elderly patients.
- US databases include more data on certain drugs than UK databases due to very different prescribing patterns to that in the UK.

SUMMARY

SAS programmers in Epidemiology may need to be able to cope with the following situations:

- Processing massive amounts of data.
- Cleaning input data, so that it is consistent, which is not often done by Data Management.
- Developing coding list for medical events and prescriptions, again not done by Data Management.
- Selecting an appropriate database for each protocol.
- Adapting SAS code to cope with differing database data content and structures.

SOME THOUGHTS FOR MANAGERS...

Epidemiology data processing is like:

- Detecting fraud in credit card transactions.
- Selecting customers for direct marketing.
- Analysing web logs.

Epidemiology data processing is not like:

- Clinical trials.

Who should, at least, be interviewed for Epidemiology data processing jobs?

PhUSE 2012

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name:	Philip R Holland
Company:	Holland Numerics Ltd
Address:	94 Green Drift, Royston, Hertfordshire SG8 5BT, UK
Phone:	+44 7714 279085
E-mail:	phil@hollandnumerics.com
Web:	www.hollandnumerics.com

This paper can be downloaded from the Holland Numerics Ltd web site at
www.hollandnumerics.com/SASPAPER.HTM

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.