

Lessons learned by comparing “using SAS” as an epidemiological researcher and “using SAS” as a statistical programmer

Saskia van de Schans, OCS Consulting B.V., 's-Hertogenbosch, the Netherlands

ABSTRACT

In 2013, I made a career switch. From being an epidemiological researcher I switched to a career as a statistical programmer for OCS Consulting, a consultancy firm. Although there are many similarities between the two, the major one being the use of SAS software for research analyses, I soon noticed that there are many differences between the two. Several aspects of my new job can be really useful for my former colleagues, epidemiologists, and my background in epidemiology helps me do my work as a statistical programmer. In this paper I will share some background information on both jobs, then zooming in on the differences and the lessons that can be learned.

INTRODUCTION

After studying Biomedical Sciences, with a major in epidemiology, I started my career as a junior researcher in 2006 at Comprehensive Cancer Centre South (CCCS) in Eindhoven, the Netherlands. This job developed into a full PhD training, with the epidemiology of non-Hodgkin's lymphoma as primary research area. Afterwards, I transferred to the Comprehensive Cancer Centre East (CCCE) in Nijmegen, the Netherlands, as a post-doc researcher specialised in the epidemiology of haematological and urological cancer. In 2013, I started at OCS Consulting, and since then I work as a consultant statistical programmer for one of our clients in the life sciences industry.

COMMONALITIES

The most important commonality between my current job and the previous is the use of SAS for data manipulation and (statistical) analyses. Both epidemiologists and statistical programmers use data steps for the selection of rows and/or columns, categorisation of variables, and derivation of information. This is followed by a series of procedures to display the characteristics of our study population, distribution of important elements in the study groups, and proof of the difference or equivalence between those groups, with mostly correction for several confounders or effect modifiers. Both jobs offer challenging programming puzzles which I enjoy solving in order to get exactly the output you want or need. Furthermore, the results of all these efforts solve my curiosity by answering the study question that was raised at the start of the research.

DIFFERENCES

COMPANIES

In order to understand the differences between two jobs, let me start with giving some background information on both companies, because several of the differences have their origin in the company structure or main goals of the company.

The Comprehensive Cancer Centres were nine regional organisations that are funded by the national government to support specialists and patients involved in oncological and palliative care. In 2014, these organisations merged into one organisation namely the Comprehensive Cancer Centre the Netherlands. The goals of this organisation are to provide advice and information to medical doctors, nurses and hospital boards, to support in the creation of healthcare guidelines, and (for me as researcher the most interesting part) by the registrations of information about all cancer patients in the Netherlands. Since 1989, a compact set of information like sex and age of the patient, and stage and treatment for the tumour is collected in the Netherlands Cancer Registry (NCR).

OCS Consulting is a consultancy specialising in SAS software and is providing its clients with professional and quality Information Technology services since 1998. We offer a wide range of services in Business Intelligence,

PhUSE 2014

Data Management and SAS consultancy and have a strong focus on the life sciences industry. Our goal is to provide our clients with experienced and skilled professionals that add value to clinical departments and study or project teams.

Soon after joining OCS Consulting I started my first assignment as a statistical programmer for one of our clients in the life sciences industry. This client conducts various research activities to improve their products. I am located at the Clinical Trials Department, where the differences between new and existing products are investigated by collecting detailed information on the effect of the products on a range of body systems. These clinical trials should give information on which products results in the best quality of life of the customers of our client.

SUBJECT OF INVESTIGATION

The subject of investigation at the CCC was of course cancer. Cancer research always impresses people. Everybody knows somebody who has this terrible disease. It is really fulfilling to contribute to a higher quality of care or better survival rate of this group of patients. At OCS Consulting, the subject of investigation is dependent of the client you are working for, so this can vary from time to time. For me, and for most epidemiologists, it is important that the subject of investigation is something that is close to the heart. Fortunately, I am active in the life science industry, so I still will be able to contribute to the quality of life of people. I think that you can take an extra step if you care about your subject of investigation, so I would recommend companies to hire consultants that are really interested in the subject, and also to involve your employers (consultants or internal) with the subject of investigation.

ORGANISATION WITHIN COMPANY

At the CCC I was responsible for the research project from the beginning until the very last moment and would personally carry out all activities within the project. Working for major life sciences companies now, all parts of the research project are done by the experts in that field. This difference has a major impact on several aspects. First of all, the tasks that you do are completely different. As a programmer my main task is programming the data manipulation and statistical analysis and 90% of my time is used for this. As epidemiologists I had to do everything; planning, organisation, (extra) data collection, data manipulation, statistical analysis, creation and publication of articles, and informing the medical specialists in my region. Furthermore, in my post doc period I even had to get funding for my research projects, because the government does fund the CCC for her main tasks, but for large research projects support from organisations such as the Dutch Cancer Society is necessary.

The result of these differences in tasks is that your expertise will differ. As a statistical programmer you really (need to) become an expert in programming. As a PhD student you are not going to be an expert on programming, but you will be the expert on your research project. So my former colleagues could really learn from my new programming skills, but the skills of knowing what is important in a research project, resistance to stress, communication and time management are really useful in my new job, and training in this would be useful for my new colleagues.

The dedication of having your own project is major; so to say it is your “baby”. It is your responsible and you want to fulfil your dream of becoming a doctor. I think it is really important for the clients of a programmer to try to get this dedication by involving all team members in other aspects of the project, for example: “Why do we want to study this?” and “What will be the end result of our investigation?”.

Another advantage of doing everything in the project is that you have the freedom to do it like you think is best, but this can be very lonely. So, I think that it is really important for epidemiologist to make sure that they are connected to other researchers, statisticians, data managers, communication experts et cetera, because you need their help and expertise. If you want to do everything yourself, you will make the same mistakes that the PhD before you also made, and this would be a waste time and money.

At our client we do the research as a team. Every task is done by someone who is an expert in that field; data collection is done by data management, researchers develop the protocol and sales will promote the results to our customers. By using all these experts the quality of work will be very high, but only if the communication in this team is optimal! Although communication costs time, it is really important; if the programmer is not aware of some important aspects of the data, or about the outcome that should be created, this is prone to errors. You do not have this risk if you collected the data yourself and if you formulated the question your research is supposed to

PhUSE 2014

answer yourself. So the lesson to learn here is that as a programmer you need to get connected and communicate to the others in your team, this is a really important aspect of the job.

Finally, if you work in a team you can lean on each other, but you are also responsible for each other. For example, planning should involve all team members. If somebody is sick or on holiday your team members are affected by this. On the other hand, in my current assignment we are working with a team of programmers, and we all have a similar expertise, so we can support each other. At the CCC my projects stood still during sickness or holiday. So, the social skills of a programmer used to make a connection with the people in your team are really important.

WAY OF WORKING

Most epidemiologists start their project with nothing more than a research question. From there they collect their data, analyse this data and write an article about it. "Nothing wrong with this", I always thought, and then I was welcomed in the world of clinical trials. My eyes were opened by the professionalism that is used in this industry. Everything is planned and prepared before data is even collected. Researchers are busy with writing study reports and statistical analysis plans (SAP), with detailed mock tables, which is an empty shell with the format of a table of figure. All details are discussed and programs are already programmed before all data is in. As this significantly improves the quality of the research, I am sure that it would be really beneficial for my former colleagues to take the time to prepare their research in a similar way.

Another thing that really can improve the quality of the work and saves time is standardisation, like the use of CDISC standards such as SDTM and ADaM. If data is gathered and analysed in a similar manner between different studies, SAS programs could be re-used. In my current assignment, macros are created that construct the standardised tables that display for example subject characteristics and adverse events. We can do this because data in these domains is always structured similarly. In my previous job we would have to program such tables over and over again, from scratch for every new research project. Not only is this very time consuming, it's also likely to have a negative impact on the quality of the output. Although, it would not always be possible to construct standardised tables for epidemiologists, because research projects are not always structured similar.

That brings me to another point, which is the Quality Control (QC) that is performed by a second programmer. QC is obviously meant to identify programming errors. Having this process adds some interesting aspects to the programmer's jobs; you need to be very critical towards the work of your colleagues, and on the other side you need to open yourself to the criticism of others. This will improve your work, because you can learn to prevent your errors, and looking at the programs of a more experienced programmer can teach you new tips and tricks.

The advantage of being an epidemiologist is that you can really dive into your data. It has to be stated that "playing" with your data, to really get grip on the data and find some unexpected but very interesting study results is a nice thing to do and is worth doing. Maybe life sciences companies could also invest more time to dig deeper into their study results. When the primary parameter is investigated, other gathered information could be used for exploratory analysis to get new and unexpected ideas.

STATISTICAL METHODS

At the CCC the statistical methods that we used were different from the ones we are using in my current assignment. At CCC we used SAS macros to calculate the incidence and mortality of cancer by the number of cancer cases/deaths at different time points (for example between 1989 and 2012) and corrected this by the number of persons in the population at that moment, by age and sex. Another important outcome is survival; this can be calculated with different methods. You can calculate observed survival (the percentage of people that is still alive after a certain moment), disease specific survival, relative survival (observed survival in cancer patients divided by the expected survival in general population), or conditional survival. If you want to show survival trends in time there are also different options; in many cases cohort analysis is used. This displays the real survival of a group of patients, but you disregard the information of patients that could not be followed during the total period, and the information is outdated, especially after the introduction of new and successful treatments. Complete analysis is already a little more up to date, because you also use the information of patients diagnosed in more recent years. Period analysis estimates the survival of the newest patients, by censoring the left side of the information (see figure 1). Paul Dickman created macros in SAS to calculate all types of survival analysis. The advantage of my epidemiology training is that I have a strong basis to connect with the statistical team in my new job. Because of this I can quickly understand their needs and the programs and macros that are already created

PhUSE 2014

to do advanced statistical analysis about the biological effect of their products. Furthermore, clients of OCS Consulting could benefit of the experience I have with the statistical analysis described above.

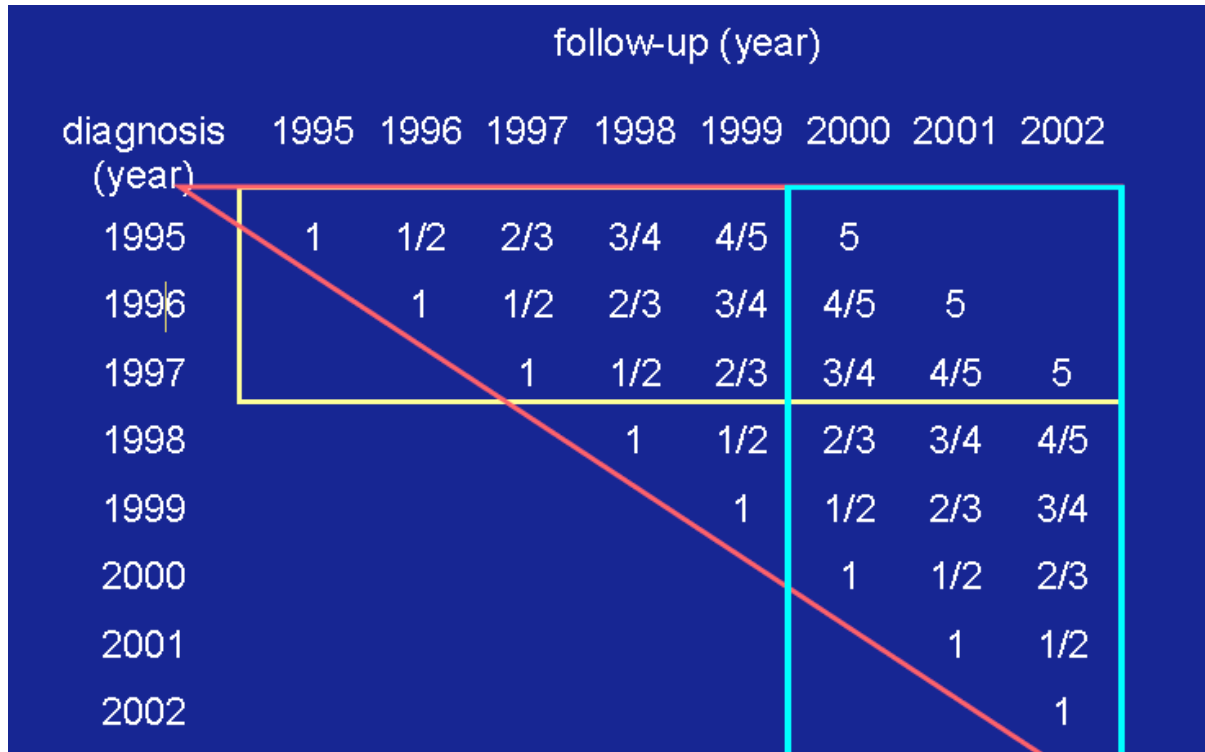


Figure 1: Different methods of survival analysis.

Legend: within the yellow lines = information used for cohort survival analysis, within the red lines = information used for complete survival analysis, within the blue lines = information used for period survival analysis.

DATASETS

Not only the types of analyses are different, also the structure of the datasets is totally different. In the Netherlands Cancer Registry all information is gathered in one SAS dataset. Every tumour has one row in that database and all information about that tumour, the treatment, the patient, et cetera is recorded in that row. You can imagine that this is a very wide dataset, with many empty fields when data is not relevant for that tumour. As a statistical programmer in the life sciences industry I really needed to get rid of my fear of merging files, because so many details are collected in clinical trials it is not possible or desirable to capture this in one row. So for my former colleagues I will give some real simple, but important examples of SAS code that I never used before, but do use on a daily basis now.

Example of a merge of two files:

```
DATA newfile;
  MERGE oldfile1
        oldfile2;
  BY subject_id; *and other variables that are present in both files, and
                  make it possible to merge the right rows together;
RUN;
```

Example of how to shift your data from a vertical to a horizontal structure, or vice versa; It changes rows into columns, and columns into rows:

```
PROC TRANSPOSE DATA=oldfile
  OUT=newfile;
  VAR = variable_that_are_transposed;
  BY = variables_that_are_not_transposed; *use proc sort procedure before
                                          this;
  PREFIX = name_of_new_variables; *instead of COL1 COL2 etc.;
RUN;
```

PhUSE 2014

USE OF SAS

Another useful tool that I came across now but never used in my previous job is the creation of reports with SAS. As an epidemiologist I copied and pasted my outcome into my article or into a graphical program, but now I use ODS. (Watch out: ODS is a tree-letter-abbreviation (TLA), people in the world of clinical trials love to use TLA's, and you do not pronounce this as "odds" ;-)).

Example of a creation of a report in PDF and RTF at the same time:

```
ODS _all_ close;
ODS PDF FILE = location\program_name.PDF style=journal;
ODS RTF FILE = location\program_name.RTF style=journal;

PROC REPORT DATA=name_file;
  COLUMN variable_1 variable_2 variable_3;
  DEFINE variable_1 / ORDER "Title of variable 1"
    STYLE(COLUMN){cellwidth=3.5cm just=right};
  DEFINE variable_2 / "Title of variable 2"
    STYLE(COLUMN){cellwidth=3.5cm just=right};
  DEFINE variable_3 / "Title of variable 3"
    STYLE(COLUMN){cellwidth=3.5cm just=right};

  * ORDER makes sure that the order of variable 1 is kept in the created table;
  * CELLWIDTH makes it possible to set the cell width;
  * JUST makes it possible to place the text left/right/centre of the table;
  * Many other options are possible;

  COMPUTE after variable_1 / style=[font_size=0.5pt BORDERBOTTOMWIDTH=1
    BORDERBOTTOMCOLOR=black];
    line ' ';
ENDCOMP;

  * This creates a black line between every new option of variable_1;

RUN;
ODS _all_ close;
```

CONCLUSION

Both an epidemiological researcher and a statistical programmer use SAS to data manipulation and analysis. The skills necessary for both jobs are not only programming, but also planning, communication and teamwork are very important. The difference between both worlds was very surprising to me, and I think we can learn from each other. I do not think there is a wrong or right way to do research in health care, but these differences show us the strengths and threats that we face, which we can use to be even better in our own fields.

REFERENCES

www.ocs-consulting.nl
www.IKNL.nl
www.cdisc.org
www.pauldickman.com for relative survival analysis programs

ACKNOWLEDGMENTS

I would like to thank OCS Consulting for granting me the opportunity to transform myself from an epidemiologists into a statistical programmer, and the PhUSE conference for letting me share my experience.

RECOMMENDED READING

PhUSE 2014

Diverging trends in incidence and mortality, and improved survival of non-Hodgkin's lymphoma, in the Netherlands, 1989-2007. Van de Schans SA, Issa DE, Visser O, Nooijen P, Huijgens PC, Karim-Kos HE, Janssen-Heijnen ML, Coebergh JW. Ann Oncol. 2012 Jan;23(1):171-82.

Actual prognosis during follow-up of survivors of B-cell non-Hodgkin lymphoma in the Netherlands. Van de Schans SA, van Steenberghe LN, Coebergh JW, Janssen-Heijnen ML, van Spronsen DJ. Haematologica. 2014 Feb;99(2):339-45

CONTACT INFORMATION

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

Saskia van de Schans
OCS Consulting BV
P.O. Box 3434
5203 DK 's-Hertogenbosch
+31 (0)73 523 6000
sasquestions@ocs-consulting.com
<http://www.ocs-consulting.nl>

Brand and product names are trademarks of their respective companies.