Documenting the analysis of an epidemiological study

Matthias Wjst
Molekulare Epidemiologie
Institut für Epidemiologie
GSF - Forschungszentrum für Umwelt und Gesundheit
Ingolstädter Landstrasse 1
D-85758 Neuherberg / Munich
Germany
mail: wjst@gsf.de
voice: +49 89 3187 4565
fax: +49 89 3187 3353

11Nov2005

In a widely noticed paper John Ioannidis proposed that most published research findings are wrong [1]. Unfortunately there is no systematic research, how errors can be prevented but I believe there are some conditions that could increase current performance. A first step would be the release of datasets together with a careful description of all variables.

Even in cases where a full release is not possible, the documentation of the analysis could be made available. The analysis of a large epidemiological dataset is a complex task that may involve internal and external work over a long time period. It seems demanding to keep up with different datasets, command files, output tables and figures, as well as with different manuscript versions. Unfortunately even Good Epidemiological Practice Guidelines do not address this issue [2],[3]. I am explaining therefore here concept that I developed during the past decade.

Methods

All files are stored in a single project directory that consists of a root directory and at least two subdirectories, one for backup and one for supplemental files (that may not be backuped). Each new project gets a new directory.

The project root directory holds also an electronic diary file. It consists of a spreadsheet with ~ 10 single sheets. The first sheet contains the project outline (see Supplement), a title, date, owner, location, a short narrative description as well as few literature links. The second sheet contains a blog with three columns: date of entry, a general descriptor, while the last column includes any free text that is associated to this action. The descriptor column may contain a keyword, for example history, todo, link, comment, question, conclusion, etc.. In addition the header line is marked as “auto-filter” allowing the selection of descriptors where a a single mouse may show for example all comments. The third sheet holds a file inventory of the projects´ root directory. The fourth sheet contains the statistical program code or command files (where I usually copy and paste program code directly to the R program editor). The remaining sheets contain result listings, tables and figures. By selecting a tabulator at the bottom a quick switch between blog, command files, figures and tables is possible. I am using no spreadsheet functions at all although it may sometimes be helpful to sum up rows and columns for tables. For writing up the final paper, tables and figures are simply copied from this electronic notebook.

As I usually access large databases across the local network (where data may be always subject to change), it is still important to have local snapshots of the data in the project directory. Retrieving raw data, merging datasets, recoding variables and saving them locally is usually performed by a program script that finally stores the retrieved data in the root directory. For a quick restart I am also storing the whole R statistics environment.

A major problem occurs when the original data are modified or expanded, or when local variables need to be redefined. Before any of these changes are taking place, I am creating a checkpoint by running a backup that saves all files in the root directory into a backup file. This is done by a short batch job (Box 1) that compresses all files into a single file. In addition this batch job adds a signature with an encryption key to preserve the integrity of the backup file. This procedure seems to be better suited than my previous approach of renumbering files with increasing order number as those saved analysis jobs would no more work with the current dataset. After having created this backup I continue with data modifications and replace all result files with the newer versions. All files in the root directory therefore represent the most recent state of the analysis.

The supplement directory is not included in the backup. It contains reference files, for example frequently used PDF files, symbolic links to local files or outside links to internet sites. Data and description in paper based laboratory notebooks are referenced by \\labbook\owner\booktitle:bookpage while references to the computer file also use UNC convention \\server\volume\path\filename:tablename At the end of the project, all backup files may be permanently saved on a DVD or even shared over the interned. By using only public domain software, all files can be accessed by any collaborator or external auditor.

Results and discussion

Documentation of analytic procedures is a critical step of an epidemiological study. Although there are many guidelines and SOPs for conducting clinical studies [4] the literature on documentation is scarce. Only very recently, data cleaning has been addressed [5] highlighting several procedures from a conceptual, logistical and statistical view. The authors discriminate a screening, diagnostic and treatment phase and recommend proper and transparent documentation of all phases without going into further details.

A recent state-of-the-art paper on essentials of good clinical practice [6] recommends keeping a study diary in which all major steps and events of the study are catalogued. Data should be analysed according to a prior "protocol, step by step, beginning with descriptive and proceeding to inferential statistics. Note any necessary modifications of the analytical plan in the study diary and give the reason for modifications. Look for qualified advice when needed."

Although a more rational approach to documentation is being suggested here this may not be used as an argument for further increasing paperwork in epidemiological studies. Many of studies are already overdocumented consuming too much time by filling in useless time sheets. Documentation needs of funding agencies, co-workers, review boards need to be balanced against freedom of academic research.

Analysis of an epidemiological study can not be done by simple checklists and execution of SOPs [7] as there always remains a subjective aspects. However, even rather quick as described here will allow to follow the main analysis track and get an idea about the quality and quantity of tests performed on a particular dataset. This might be particular important in teaching epidemiology [8].

Box 1: A batch script that saves the content of the root directory into a single backup file and adds an electronic signature

```
@echo off
rem called as backup.cmd without additional parameters
rem from the working directory
rem the directory where the zip file resides:
set p=c:\Program\system\
rem the directory where the signature resides:
set g=c:\Program\GnuPG\
rem creates a zip file name with current date
set z=%date:~9,4%%date:~6,2%%date:~3,2%.zip
rem compresses root directory content into backup file
%p%zip %cd%\backup\%z% %cd%\*.*
rem signs the new backup file
%g%gpg -b -armor %cd%\backup\%z%
exit
```

Box 2: Download links for software used

www.openoffice.org

Open Office 2.0 is a multiplatform and multilingual office suite and an open-source project.

www.r-project.org

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of platforms.

www.gzip.org

Gzip is a compression utility with a high compression rate and free from patented algorithms.

www.gnupg.org

GnuPG is a complete and free encryption solution to protect confidential communication and digitally stored information.

Box 3: Critical points of study analysis

Check recent dataflow and monitoring

Consider drawing a flowchart diagram for the following analysis

Which subset may be used?

Is there any need to exclude any individuals?

Decide on main outcomes, exposures and confounders

Check for correct coding, internal and external plausibility of variables

Look for inliers and outliers, tabulate missing values

Test variance, linearity, normality, serial correlation and collinearity

Does transformation of variables help?

Tabulate and plot main outcomes, exposures and confounders

Develop standard procedures to identify additional confounders by looking at hidden or unexpected associations

Proceed to multivariate techniques

Improve the multivariate model, change variable selection

Try an internal or external validation

Identify areas where special statistical procedures may be appropriate or where simulations may be necessary

Sum up the results and discuss the approach with others

Loop until being confident on the main findings

References

1. JP Ioannidis: **Why most published research findings are false**. *PLoS Med* 2005, **2**:e124.
2. W Hoffmann, U Latza, W Ahrens, KH Greiser, A Kroke, A Nieters, MB Schulze, M Steiner, C Terschuren, M Wjst: **[Biological markers in epidemiology: concepts, applications, perspectives (part I)]**. *Gesundheitswesen* 2002, **64**:99-107.
3. JR Jackson: **Activity in 1998 in the development of standards of good epidemiological practice**. *International Archives of Occupational and Environmental Health* 1999.
4. G Fortwengel: **Guide for Clinical Trial Staff Implementing Good Clinical Practice**. *Karger* 2003.
5. J Van den Broeck, SA Cunningham, R Eeckels, K Herbst: **Data cleaning: detecting, diagnosing, and editing data abnormalities**. *PLoS Med* 2005, **2**:e267.
6. E Altpeter, B Burnand, G Capkun, R Carrel, B Cerutti, M Mausezahl-Feuz, M Gassner, C Junker, N Kunzli, C Lengeler, et al: **Essentials of good epidemiological practice**. *Soz Praventivmed* 2005, **50**:12-27.
7. CM Clive: **Handbook of SOPs for good clinical practice**. *Interpharm, CRC press* 2004.
8. CV Phillips, KJ Goodman, C Poole: **Lead editorial: The need for greater perspective and innovation in epidemiology**. *Epidemiol Perspect Innov* 2004, **1**:1.