The Journal Of The International Federation Of Clinical Chemistry And Laboratory Medicine
eJIFCC

STATISTICAL MANAGEMENT OF AUTOIMMUNE DISEASES DATA

Prof. Mladen Petrovečki, Ph.D.
Assistant Minister for Science, Ministry of Science, Education and Sports, Republic of Croatia

Olga Gabela, B.Sc.; Tea Marcelić, B.Sc.
Junior researchers, University School of Pharmacy and Biochemistry, Zagreb, Croatia

11.1 Handling the data - statistical management

From the statistical point of view, basic purpose of research of autoimmune diseases, as it is mainly with all others, is to collect the data to obtain information about particular disorders. Usually, data are obtained from one or more samples of patients that represent the population, the whole group of individuals of our interest. Just as an example, if the topic of investigation is expression of CD45 lymphocyte isoforms in newborns shortly after they have been delivered (Juretić et al., Ref.), then the sample of newborns has to be representative of the same population. We collect and analyze the data on a sample of some (randomly chosen) N individuals and use them to draw conclusions about the population. That process - making inferences - is the meaning of scientific research. After reading the paper and learning something new on lymphocytes in early newborns, we expect to have the same finding in all newborns in the world (generalization), i.e., that facts from the paper, related to individuals in that research, become general facts considering all individuals with same characteristics (and that is why those characteristics of individuals have to be briefly but clearly explained in the research paper). Of course, this is a simplified explanation of generalization and we also have to recognize that information from the sample does not fully indicate what is true in the population. There is also a sampling error that has to be considered.

If the sample is not representative, if it just enumerates a group of individuals or subjects that were examined, measured, studied and analyzed, then the process of making inferences about the population does not exist at all - there is no population for which facts from the study might be true. But so-called convenience sample might be useful just to test something before it will be utilized in the real scientific research, for example, to test if new database on autoimmune diseases is user friendly by entering the data of first ten patients that come in the office. Of course, there is no statistics to be reported about these patients, only about filling the forms.

Correct sampling is only the first step. Statistical management of data obtained from biomedical research and from clinical trials is a complex knowledge of choosing an appropriate statistical method for the data analysis and data presentation. Beside statistical analysis itself, i.e., utilizing predefined mathematical calculations with collected data to compute tests' specific values with their probability values, handling the data also implicitly includes that researcher has knowledge on sampling, data and errors types, outliers, distributions, measures of average and data spread, hypothesis theory, study design, data transformations, etc. Readers of scientific papers must also have basic knowledge on statistical data handling; otherwise, reading can not be critical, as expected.

As presented recently by Tom Lang in Croatian Medical Journal, inadequate statistical reporting in scientific literature is mostly due to authors' poor knowledge about research design and statistics, statisticians' inability to communicate statistics to authors, editors and readers, lack of involvement of statisticians at the beginning of research, and not applying statistical reporting guidelines.

11.2 Statistical errors

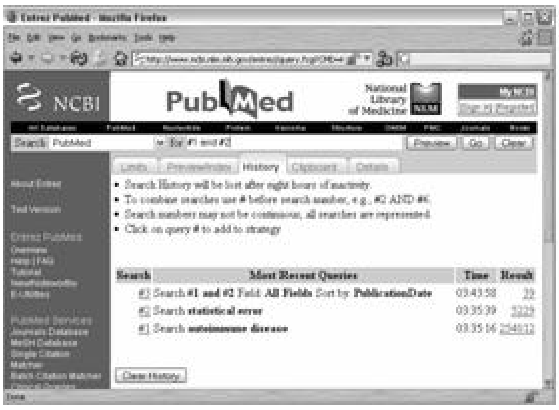
Statistical errors in scientific reporting are not rare, as someone might expect, and some critical reviewers of biomedical literature found that about half of the articles that used statistical methods did so incorrectly. Even big, high-impact and prestigious journals are not immune on statistical errors. Most errors concern basic statistical concepts and can be easily avoided by following guidelines. Typically, guidelines are published as regular printed handbooks (for example: Lang and Secic, Ref.), they can be prepared as educational electronic manuscripts and published on Internet, or can be published in medical journals as a part of guidelines for authors (for example: Editorial Policy of Croatian Medical Journal, available in the 1st issue of each volume, but also at www.cmj.hr).

One good example of electronic manuscript is "Guides to Good Statistical Practice" from the Statistical Service Centre of The University of Reading (UK), intended primarily to give help to research and support staff in development projects. The guides are available to read online (http://www.rdg.ac.uk/ssc/publications/guides.html) or to download for printing and reading offline from the same Internet address. They were also rewritten and published in 2004 as a book (Stern et al., Ref).

Some biomedical journals introduced statistical editors to confront the problem, but still without finding the perfect solution to the problem, suggesting that some other measures are necessary, such as strict editorial policy on statistical review, monitoring of revised manuscript version and enrollment of formally trained biostatistician (Lukić & Marušić, Ref).

11.3 Are autoimmune diseases data immune on statistical errors?

Probably not, but no brief and straightforward answer was found after keywords "autoimmune disease" and "statistical error" were run through Medline search: 39 papers published between 1981 and 2005, three of them, review articles, were found in database but none of them considering the topic of this paper (Fig. 1). In most publications "statistical error" was considering new or corrected approach to the subject, analysis of statistical errors in procedures performed through the study, or comparative analysis of two or more techniques (only abstracts were analyzed).



- Search History will be lost after eight hours of inactivity.
- To combine searches use # before search number, e.g., #2 AND #6.
- Search numbers may not be continuous; all searches are represented.
- Click on query # to add to strategy

Search	Most Recent Queries	Time	Result
#3	Search **#1 and #2** Field: **All Fields** Sort by: **PublicationDate**	03:43:58	22
#2	Search **statistical error**	03:35:39	5229
#1	Search **autoimmune disease**	03:35:16	254012

Figure 1. Medline search, using PubMed service on Internet

To find an answer to the question, a small observational study on non-representative sample was performed. In between 15.208 articles published in numerous biomedical journals in years 2004 and 2005, listed by keyword "autoimmune disease" from Medline, only seven journals with high impact factors were considered in selection and only those that can be found in a printed version at the Central Medical Library of the Zagreb University School of Medicine: American Journal of Hematology, Arthritis Research and Therapy, Arthritis & Rheumatism, Autoimmunity, Diabetes, Journal of Autoimmune Diseases, and Lupus.

From 1.075 articles published in these seven journals, fifty papers were chosen by chance, but at least one per journal, to form simple convenience sample. Two young biomedical scientists, O.G. & T.M., attentively read articles and in 18 of them found at least one error related to the statistical analysis or statistical data presentation. In total, more than thirty errors were found that can be presented through six groups (mistakes, errors, doubts and ambiguities will be presented individually during the lecture, here only a summary of groups is given).

11.3.1 Data types, presentations and mismatch

This group of errors consists mostly of mistakes, typing errors and inadequate data presentations. Some are unintentional, but sometimes authors are unaware of possibility to present the data using texts, tables or graphs. Also, some problems concerning data types in making the difference between categorical and numerical data occur, resulting in wrong statistics and possible mistakes in conclusions. Some "historical tables" were still noticed, listing all data from the experiment, and with no consequent data analysis.

Numerical data might appear in two distinct types, discrete and continuous, and each of them has their own characteristics of presentation. Also, if data have measuring unit, it has to be presented.

11.3.2 Average and dispersion

Problem with summarizing the data appears when authors do not have knowledge on their own data (How big is the sample? Is distribution normal? How sampling was conducted?) or when authors do not understand types of summarizing techniques. Numerous errors were found, from small to important ones. After reading results of presentations, readers definitely can not infere from sample to population.

11.3.3 Problems with "Subjects and Methods"

Unclear sampling methods and appearance of questionable control subjects were detected. Information about statistics authors frequently write together with results, instead of putting statistics theory into the Methods section of the paper. Surprisingly, in some papers more methods of statistical analysis were discussed than afterwards was presented in the paper.

11.3.4 Statistical errors

This is probably the most important group of errors, covering all kinds of inappropriate statistics that might lead to completely wrong conclusions. Unfortunately, some were found, even one dealing with low correlation coefficient, close to zero but significant, that was considered important!

11.3.5 P-values

All kinds of errors in presenting probability values (no p-values, only statements what is significant and what is not with no numbers, improper decimals, etc.) and wrong explanations on significance were frequently noticed.

11.3.6 Results

Statistical explanation, if used as the only explanation while presenting results is wrong, indicating that author probably does not understand output of statistical analysis. Results should be presented in a way that everybody from the same scientific field can read, understand and have no doubts about them.

11.4 Instead of the conclusion

Although some authors reported that big scientific journals might differ from small ones in a preparation of manuscript for publishing, considering statistical management of the manuscript and statistical data reporting and presentation, this small ad hoc study proved that field of autoimmunity research still suffers from statistical unclearness. And that could bring up the next question - how much can we rely on the results we are reading about?

Literature

1. Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics. 3rd ed. London: Prentice hall Int. Inc 2001.
2. George SL. Statistics in medical journals: a survey of current policies and proposals for editors. Med Pediatric Oncol 1985;13:109-112.
3. Glanz SA. Biostatistics: how to detect, correct and prevent errors in the medical

literature. Circulation 1980;61:1-7.
4. Lukić IK, Marušić M. Appointment of statistical editor and quality of statistics in a small medical journal. Croat Med J 2001;42:500-3.
5. Juretić E, Gagro A, Vukelić V, Petrovečki M. Maternal and neonatal lymphocyte subpopulations at delivery and 3 days postpartum: increased coexpression of CD45 isoforms. Am J Reproductive Immunol 2004;52:1-7.
6. Lang T. Twenty statistical errors even YOU can find in biomedical research articles. Croat Med J 2004;45:361-370.
7. Lang T, Secic M. How to report statistics in medicine: annotated guidlines for authors, editors, and reviewers. Philadelphia: Am. College of Physicians 1997.
8. Petrie A, Sabin C. Medical statistics at a glance. London: Blackwell Science 2000.
9. Stern RD, Coe R, Allan EF, Dale IC, ed. Good Statistical Practice for Natural Resources Research. Statistical Service Center. Published May 2004. Paperback, 416 pages, ISBN 0851997228.