Statistical software for medical professionals

Olivera Djordjevic
Rehabilitation Clinic "dr Miroslav Zotović", Belgrade, Serbia and
Medical Faculty, University of Belgrade, Belgrade, Serbia
odordev@eunet.rs

Abstract: Statistical data analysis is inherent part of medical research method. The choice of certain statistical software in the medical research depends on many factors. Although the popularity of free statistical software is rising, an ease of use and inertia of the professional surrounding might not be less important factors in the decision-making process and choosing the "right" statistical package. However, new free software that focuses on non-programmers' looking to point-and-click their way through analyses may be the "free software answer" to both efficient, free, and easy to use.

Keywords: statistical analysis; medical research; free software.

I. Introduction

Medical statistics has been introduced in medical research by sir Austin Bradford Hill through in the late 1930s [1][2][3]. During the 1940s, 1950s, and 1960s, the use of formal statistical methods in medical research grew and statistical data analysis subsequently became an inherent part of medical research method.

The optimal situation in medical research environment is that medical researchers have at least critical level of statistical understanding, while a statistical researcher (if there is any) should sufficiently understand the nature of data. Unfortunately, this is not a common situation, and sometimes, statistical software tools may damage statistical practice if they distract attention from statistical goals and tasks, onto the tools themselves [4]. Considering this circumstance, the engineers of statistical software should create tools that should prevent this misuse; general statistical software has, however, been criticised for failing to trap misuse in medical research [4] [5]. Statistical computer software has been criticised for giving its users false confidence when performing analyses and thereby distancing relevant statistical appreciation from common research practice [4][5].

The question of choice of the statistical software for handling medical data is wide and depends on many rational and some not so rational and yet, potentially much influential factors.

Our main interest in this paper is to make an overview of the most popular statistical software in medical research. The statistical software that is used for handling medical data in the segment of health service such as pharmaceutic, economics, management and administration, policy analysis and policy implications of health care research, public health focus, and financing/insurance focus remained beyond the scope of our overview.

Figure 1: Sir Austin Bradford-Hill has introduced statistics in medical science with the series of articles in the journal The Lancet in the late 1930s and dedicated most of his career to work in this field [1][2][3]. Photo: By Unknown, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=33258846

II. How do we choose data analysis tool?

The choice of a particular software package for a particular medical research depends on the study's specific analytical needs, the nature of our data, the suitability of a particular software application for a specific analysis, on the choice of the packages that are available to us, on the investigators' skills and experience, on the budget for statistical software if we have to buy it, and on the time that has to be spent on learning the software. Among the less rational but not less important and decisive factors of selecting the "right" statistical software is the what is used in local professional milieu (albeit not knowing if that one is the most adequate one). Sometimes this inertia may not be significant issue, since there are many similarities between statistical software packages (SPSS, SAS, R, Stata, JMP, …) [6] in the logic and wording they use even if the interface is different. But, if we are limited with the budget and the skills for command-line interfaces than these similarities are not especially helpful.

DOI: 10.5281/zenodo.3472316

The comprehensive approach on the key factors that are important for the choice of the right statistical software in the form of "matrix of factor classification" has been presented by Cavaliere [7]. It consists of the system of subjective and objective, endogenous and exogenous factors that has an impact on the decision process. An interesting model of the decision process related to the choice and acquisition of the statistical software has also been presented. As mentioned earlier, this is strongly context dependent process that is marked by the functional requirements, the level of statistical and programming needs, knowledge of the researchers and IT infrastructure.

III. Overview of prevalence of statistical software in medical research

In an interesting and regularly updated review on the frequency of the statistical software, prof. Muenchen [8] lists the most usual ways of analysing the frequency of statistical tools: job advertisements, scholarly articles, survey of use, books, blogs, discussion forum activities, programming popularity measures, sales and downloads, and competition use growth in capability.

The frequency of requirements for certain data analysis software in *job advertisements*, give us an estimation of how much is that software demanded. Job advertisements are comprehensively designed with the detailed description of the needed software skills and are backed by money. Therefore, they should be a trustful source of the need and popularity of certain data analysis software. However, medical research in our local community are seldom sponsored on a regular basis. So, my feeling is that the more comprehensive way of estimating the popularity of statistical software for medical data handling would be analysing the data science tools used in *scientific articles*.

Although expensive proprietary software, SPSS is still the most dominant package. After analysing 80000 articles for 2018 found on Google Scholar, Prof. Muenchen [8] shows that SPSS is by far most dominant package (about 50% in all academic articles). It has been so for over 20 years. He suspects that this might be due to its balance between power and ease of use. R is in the second place with around half as many articles. It offers big power, but less ease of use. However, his extensive analysis also shows some trends of change: decline of the traditional software packages and increase of use of data software associated with AI/ML; cheap or free software is in increase in demand, expensive is down.

Our subjective impression of the still unperturbed dominance of the SPSS in medical research articles is backed up with the finding in Prof. Muenchen's analysis [8]: despite consistent decline during previous 10 years, SPSS is still extremely dominant for scholarly use. This analysis shows that R and SAS are still the right behind it. Similar pattern, but in much smaller figures followed SAS and GraphPad Prism.

Surveys of use give us different estimations on the popularity of statistical software, may be a bit too strongly dependent of the undertaker of the survey itself. According to the Rexer Analytics survey [8][9] of data scientists R has a more than 2-to-1 lead over the next most popular packages, SPSS Statistics and SAS. Microsoft Excel Data Mining software is slightly less popular.

According to the results of Lavastorm Analytics Community Group, Data Science Central and KDnuggets. from 2013 [10], Excel comes out as the top self-service analytic tool and R comes out as the top advanced analytics tool with 35.3% of respondents, followed closely by SAS. MS Access position in 4th place is a bit of an outlier as no other surveys include it at all. A review of statistical analysis of software programs used in biomedical research, based on a Labome survey [8][11] of randomly selected, formal publications that has been updated in August 2019, the list of the main brands of data analysis and graphing software and the number of articles among the articles is as follows: Prism, SMP, StatView, Excel, Origin, while SPSS and R are found to be in the lower half of the 10-item list of statistical software most frequently used in biomedical research [11] [12][13][14][15][15][16].

Books. The frequency of showing up a software name in the book title may give us an estimation on the software popularity. Considering the effort that accompanies the risk-taking process of writing and publishing that preceded by research of the market demand, the number of published books mirrors the requirements for a specific software. According to the number of books that include a software name in its title, the most popular software packages are SAS, SPSS Statistics, R, JMP…) [8].

The rise of software that uses the workflow (or flowchart) style of control has been observed as a trend, recently. Software that uses this approach includes: KNIME, Microsoft Azure Machine Learning, RapidMiner, SPSS Modeler, SAS Enterprise Miner, SAS Studio, Dotplot Designer and Microsoft Azure Machine Learning. Workflow-driven software is almost as easy to learn as menu driven software and they are also time-saving since we can save and re-use the work. The wide use of this interface is allowing non-programmers to make use of advanced analytics, but we have not observed significant rise in popularity of these potentially useful packages in medical research.

As mentioned before, successful data analysis is based on both statistical knowledge, mustering of the statistical software and the ability to interpret results.

Although medical professional seek a comprehensive statistical software, if the research team do not have fully designated statistical engineer or statistically educated team member, most of "lay" statistician seek software which is intuitive (obvious; second nature; mustered by instinct), cross-platform (available for more than one operating system, such as Windows, Mac OS X, Linux), menu-driven (instead of having to issue commands, most of the procedures can be accessed via pull-down menus in the graphical user interface (GUI)

and do not have steep learning curve (or takes considerable time and effort to learn).

On the opposite side from the user friendly on the spectrum of the statistical data tools are text-based user interfaces, typed command labels or text navigation, steep learning curve of command-lines interfaces which require commands to be typed on a computer keyboard are not user friendly toward medical professionals. So far, and despite of sharp ten years drop, SPSS is still the most frequently used statistical software. Its use was still 66% higher than R in 2018 [8]. It seems that (lay) statisticians are motivated to find the way to pay for ease of use. Recently arrived free software that uses graphic user interface that is menu driven and is similar in style to SPSS such as JASP, jamovi, and BlueSky Statistics might meet the needs of medical researchers who are not willing to invest their time in learning command line- based software. For example, BlueSky Statistics is a free and open source, cross platform, graphical user interface for the R software that focuses on beginners looking to point-and-click their way through analyses [17]. It remains to be seen whether these easy to use, free, and open source software will chip away at SPSS dominance.

IV. Conclusion

Although the popularity ranking of each package varies depends on the criteria used, we can still see major trends: SPSS, R, SAS, and Stata tend to always be in the top. Medical professionals who are doing statistics on their own in the research strive to the ease-of-use in a software and that is how quickly and effortlessly they can find out how to do what they want without time consuming prior instruction, consultation of manuals or third-party help. Trustful free software which fulfil these requirements would probably meet the needs of the most of them.

References

[1] A. B. Hill, *Principles of medical statistics*, (The Lancet Postgraduate Series), London: The Lancet, 1937.

[2] A. B. Hill, *A Short Textbook of Medical Statistics*, (11th edition), London: Hodder and Stoughton, 1984.

[3] A. B. Hill, I. D. Hill, *Bradford Hill's Principles of Medical Statistics*, London: Edward Arnold, 1991.

[4] I. E. Buchan, *The Development of a Statistical Computer Software Resource for Medical Research*, University of Liverpool, Liverpool, England, 2000.

[5] D. G. Altman, "The scandal of poor medical research," British Medical Journal, Vol. 308, pp. 283-4, 1994.

[6] https://en.wikipedia.org/wiki/List_of_statistical_software [Accessed Sep. 1, 2019]

[7] R. Cavaliere, "How to choose the right statistical software? A method increasing the post-purchase satisfaction," J Thorac Dis, Vol. 7, No. 12, pp. E585-E598 , 2015.

[8] R. Muenchen, *The Popularity of Data Science Software*, Available at http://r4stats.com/articles/popularity/

[9] https://www.rexeranalytics.com/Data-Miner-Survey-2013-Intro.html [Accessed Sep. 1, 2019]

[10] https://www.infogix.com/ [Accessed Sep. 1, 2019]

[11] M. Johnson, "Statistical Analysis Software Programs in Biomedical Research," Mater Methods, Vol. 4, pp. 1282, 2014.

[12] J. Holth, S. Fritschi, *et al*., "The sleep-wake cycle regulates brain 7. interstitial fluid tau in mice and CSF tau in humans," Science, Vol. 363, pp. 880-884, 2019.

[13] M. Lee, B. Siddoway, *et al*., "Somatic APP gene recombination in Alzheimer's disease and normal neurons," Nature, Vol. 563, pp. 639-645, 2018.

[14] T. Clairfeuille, A. Cloake, *et al*., "Structural basis of α-scorpion toxin action on Nav channels," Science, Vol. 363, 2019.

[15] A. N. Bolaños, F. A. Sempere, *et al*., "Prenatal activity from thalamic neurons governs the emergence of functional cortical maps in mice," Science, Vol. 364, pp. 987-990, 2019.

[16] J. Yang, J. Chen, *et al*., "PAC, an evolutionarily conserved membrane protein, is a proton-activated chloride channel," Science, Vol. 364, pp. 395-399, 2019.

[17] R. Muenchen, *A Comparative Review of the Blue Sky Statistics*, GUI for R, Available at http://r4stats.com/2018/06/21/bluesky/.