Consolidated data analysis and presentation using an open-source add-in for the Microsoft Excel® spreadsheet software

Daniel Kraus

First Department of Medicine, Division of Nephrology, Würzburg University Hospital, Würzburg, Germany

Correspondence to:

Daniel Kraus
First Department of Medicine
Division of Nephrology
Würzburg University Hospital
Oberdürrbacher Strasse 6
97080 Würzburg, Germany
daniel.kraus@uni-wuerzburg.de

Abstract

A free and open-source software tool is presented that facilitates the analysis and the visualisation of data in basic life science. Daniel's XL Toolbox is an add-in for the Microsoft Excel® spreadsheet software. It enables scientists to store their data in one place and obviates the need to use separate tools for the analysis and presentation of the data. The Toolbox offers analysis of variance with *post hoc* multiple comparison testing and linear correlation and regression analyses. It can apply error bars to the graphs automatically, and style the graphs in a way suitable for publication. The graphs can be exported in high resolution to TIFF, PNG, as well as to EMF file types. Finally, the Toolbox offers several work flow and productivity features such as automatic time-stamped backups, transposition assistant, worksheet management, and much more. The Toolbox is available at http://xltoolbox.sourceforge.net.

Keywords: Analysis, Visualization, Work flow, Software, Open source

For the storage and the exchange of small- to medium-sized datasets, Microsoft Excel® is the *de facto* standard software tool. However, while the data can be grouped and labelled and graphs can be created very easily by using Microsoft Excel, the software lacks the ability to perform certain statistical analyses that are commonly used in biomedical research, and it has only limited capabilities to create publication-ready figures. Therefore, scientists often resort to specialised separate software tools to accomplish these tasks. However, the duplication of data by using 'copy and paste' between applications may leave the data in a disparate state. This often makes it difficult to comprehend the entire analysis and work flow unless the scientist takes extra time for meticulous documentation and labelling of every copy of the data. Experience tells that this is rarely the case.

To remedy the shortcomings of Microsoft Excel with regard to the needs of scientific data analysis, an add-in was developed by using Excel's built-in extensibility feature. This project was started in 2008 as a set of macros intended for personal use. Over time, more and more features were added, turning the add-in into a versatile tool for all Microsoft Excel versions from 2003 through 2013. The add-in was named 'Daniel's XL Toolbox' and made freely available via an open-source software platform under the GNU General Public License; it can be downloaded at http://xltoolbox.sourceforge.net. Users can investigate the source code either online or by summoning Microsoft Excel's built-in Visual Basic Editor by pressing the ALT and F11 keys while Excel is running. An overview of the features is given in Table 1 and Figure 1. A few of them are described in this article.

Planning an experiment

Stratification of study groups is a norm in large-scale clinical trials. It is, however, uncommon for basic research involving only a few dozen study subjects, e.g. mice. Yet, random differences between the individual characteristics may have a large effect on study groups with few subjects. To ensure equality between the group characteristics, the XL Toolbox offers two alternative algorithms to allocate the study subjects. The first is derived from an algorithm described by Endo *et al.*[1] It uses Kullback–Leibler divergences to assess differences between the study groups, and assigns the study subjects to a group so that the groups diverge as little as possible. The second algorithm performs analysis of variance (ANOVA) and allocates the

DOI: 10.1179/2047480613Z.000000000181

Table 1: Overview of the XL Toolbox features

Data analysis	Data visualisation	Work flow
● One-way ANOVA ● Two-way ANOVA ● Formula builder ● Frequency analysis (histogram) ● Group allocation (stratification) ● Linear correlation ● Linear regression ● *Post hoc* testing after one-way ANOVA ● Transpose wizard	● Automatic error bars ● Graph annotation (asterisks, labels, colours) ● Simple, reusable graph design ● Graph watermarks ● Graph export to high-resolution files ● Copy graph properties to other graphs	● Automatic, time-stamped backups ● Copy worksheet properties to other worksheets ● Open file from file name in worksheet cell ● Selection assistant ● Simplified worksheet management (add, delete, move, rename) ● Special paste functions

study groups to the subjects so that the ANOVA's *P*-value is maximised. Both the methods allow for a certain degree of randomisation to make the allocation unpredictable. While the algorithms offered by the Toolbox are not nearly as sophisticated as those of dedicated software packages, they are well suited for small- to medium-sized experiments or clinical trials if only a few characteristics are to be stratified.

Arranging and describing the data

Many laboratory machines such as polymerase chain reaction cyclers and microplate readers produce output in Microsoft Excel files. A common task, then, is to re-arrange the layout of the data on a worksheet. Getting these data into an order that makes sense for the experiment at hand, is a cumbersome and error-prone process of repeatedly copying and pasting. The XL Toolbox's Transpose Wizard automates this process and includes functions to summarise the transposed data and calculate standard curves.

The Formula Builder command assists with the analysis of the unordered data of several experimental groups. Thus, the raw data can be written in the order in which it is obtained, e.g. from a cohort of study subjects, where the adjacent subjects may belong to different groups.

Performing statistical analyses

The Analysis ToolPak add-in that is shipped with Microsoft Excel, offers a one-way ANOVA that can be used to test for statistically significant differences between three or more groups of data. However, its usage is not straightforward. It forces the users to arrange the data in a specific way, and it does not support *post hoc* multiple comparison testing. The improved ANOVA that is offered by XL Toolbox is able to perform analyses regardless of how the data are arranged, and it includes three different *post hoc* testing algorithms (Bonferroni–Holm, Holm–Šidák, and Tukey). The data may be laid out in horizontally- or vertically oriented tables or lists, as long as they are labelled consistently. The one-way ANOVA checks the assumption of homoscedasticity (homogeneity of variance) by using a modified version of Levine's test as described by Glantz and Slinker.[2] If the assumption is not met, non-parametric methods must be used, yet currently the XL Toolbox does not offer one.

If there are two independent variables rather than one, the Toolbox's two-way ANOVA can be used. This function is also very flexible with regard to the layout of the data. Even lists of data with combined group labels in the form 'genotype/treatment' or 'gender, age' and similar are accepted. The XL Toolbox can analyse the data from repeated intra-individual measurements. *Post hoc* testing is currently not offered for the two-way ANOVA.

Creating publication-ready graphs with error bars

Scientists and journal editors often prefer graphs with a clear, minimalistic design. By default, Microsoft Excel produces rather colourful graphs that are not suitable for publication in scientific journals without further modification. The most basic and arguably most efficient graph design uses only black and white, clearly distinguishable symbols, and readable labels. The XL Toolbox

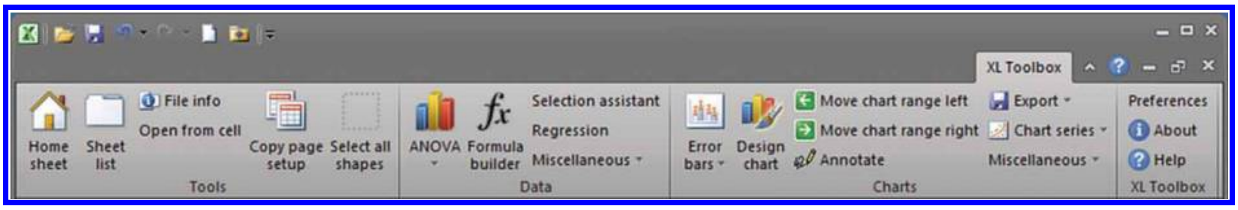
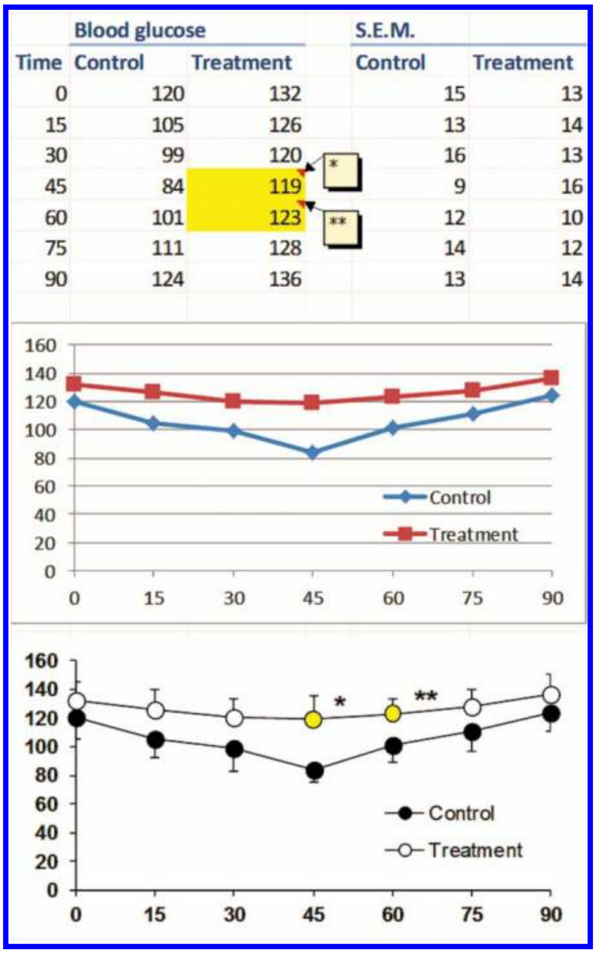


Figure 1: The XL Toolbox 'ribbon' user interface. The application shown is Microsoft Excel 2010.



	Blood glucose		S.E.M.	
Time	Control	Treatment	Control	Treatment
0	120	132	15	13
15	105	126	13	14
30	99	120	16	13
45	84	119	9	16
60	101	123	12	10
75	111	128	14	12
90	124	136	13	14

Figure 2: A representative graph of made-up data before (top) and after (bottom) using the ‘Chart design’, ‘Annotate chart’, and ‘Error bars’ commands of the XL Toolbox. After inserting the graph, the size was adjusted, the legend was moved over the chart area, and the *X*-axis scale and interval were adjusted. The subsequent modifications were performed with just a few mouse clicks by using the XL Toolbox.

enables scientists to define simple, professional designs that meet these criteria and apply them to their graphs with a few clicks.

Adding error bars to scientific graphs by using Microsoft Excel’s built-in commands is very time-consuming. Therefore, a user-friendly and fast way to add custom error bars was implemented for the XL Toolbox (Figure 2). Users can either rely on fully automatic detection of the error data, or take advantage of an interactive mode. If requested, the error bars for a line graph can be automatically pointed in the positive or the negative direction so as to minimise the overlap.

Exporting to high-resolution graphic files

Many scientific journals require high-resolution graphic files in the TIFF format; some journals accept TIFF files exclusively. The XL Toolbox can produce such files from one or many graphs or a combination of graphs and other drawing objects or even spreadsheets. The ‘export for publication’ command offers the following options:

- *File formats*: TIFF, PNG, and EMF
- *Colour spaces*: Black and white, greyscale, RGB, and CMYK
- *Colour management*: Available for chart export (however, Microsoft Excel itself does not support display colour management for graphs)
- Any resolution within reasonable limits

Spreadsheets can be used as layout tables to produce figures that contain several panels. If the ALT key is held while a graph is clicked and dragged with the mouse, the graph will snap to the grid of the spreadsheet. To add a panel letter to a graph, first select the graph, and then insert a text box. The text box will be linked to the graph and stay with it when the graph is moved. The ‘select all shapes’ command of the Toolbox followed by Microsoft Excel’s ‘group’ command can be used to obtain a single multi-panel figure. Alternatively, choose ‘all graphic objects on the current worksheet’ when exporting for publication.

Technical limitations

Daniel’s XL Toolbox requires Microsoft Excel running on the Windows® platform. Owing to major differences between the Windows and the Macintosh® editions of Microsoft Excel, the add-in does not run on the Macintosh. The Toolbox is not compatible with OpenOffice and LibreOffice, since these office suites employ an entirely different extensibility technology.

Perspective

Currently, a major rewrite of the XL Toolbox is under way that implements modern programming concepts and is based on the Microsoft® .NET framework. This will facilitate maintenance of the code as well as the addition of new features. For example, it is planned to add SVG and PDF export capability and expand the set of methods for statistical analyses.

Conclusion

Daniel’s XL Toolbox offers many tools for the analysis, visualisation, and management of scientific data. Only a few of them could be described in this article. Interested medical writers are invited to explore the extensive online documentation at http://xltoolbox.sf.net. The author welcomes feedback and suggestions.

Acknowledgements

The author has drawn on the programming experience and statistical expertise of many people. An exhaustive list of acknowledgements can be found on the Toolbox's website at http://xltoolbox.sf.net/about. The author is especially grateful to all users who acknowledge the Toolbox in their articles.

Conflicts of interest and disclaimers

Although the XL Toolbox is available free of charge, the author accepts voluntary payments ('donations') for this software on the website. This software project is maintained in the author's own time; it is not officially endorsed by the author's employer. Daniel's XL Toolbox is an independent software and is not affiliated with, nor has it been authorised, sponsored, or otherwise approved by Microsoft Corporation. Microsoft, Windows, and Excel are either registered trademarks or trademarks of Microsoft Corporation in the USA and/or other countries. Macintosh is a trademark of Apple Inc., registered in the USA and other countries.

Author information

Daniel Kraus is a physician-scientist with a long-standing personal interest in computer programming. He works as a staff physician at Würzburg University's main hospital. In his laboratory, his team study basic mechanisms of energy metabolism in adipose tissue.

References

1. Endo A, Nagatani F, Hamada C, Yoshimura I. Minimization method for balancing continuous prognostic variables between treatment and control groups using Kullback-Leibler divergence. Contemp Clin Trials 2006;27(5):420–31.
2. Glantz SA, Slinker BK. Primer of applied regression & analysis of variance. 2nd ed. New York: McGraw-Hill; 2000.