

# Collaborative Research

# Reproducible Research

Ronald Thisted, Professor of Biostatistics and Chairman of the Department of Health Studies at the University of Chicago, works with colleagues to develop and implement practices that support research practices, including reproducible research.

**“R**esearch is not a single activity that takes place at a discrete point in time—it comprises a continuum of related activities involving a myriad of choices and decisions that evolve over a period of time. Much can happen from the point at which a research question is conceived to the point at which the results of that research have become embedded in the discipline. All too often, at the end of that process, it is impossible to answer the questions, “Where did that particular result come from? On what assumptions does this finding depend?” explains Ron Thisted, offering some of the reasons behind the use of reproducible research practices.

Thisted and colleagues at the University of Chicago, including Phil Schumm, a biostatistician in the Department of Health Studies, are working to implement these practices in their own work.

Reproducible research practices are systems that seek to maintain clarity within datasets, and between datasets and manuscripts, such that at any point in a study or review process, the results of an analysis can be replicated, corrected, or modified.

In any experimental research, and particularly in the ongoing longitudinal studies conducted by many members of the

Center for Cognitive and Social Neuroscience (CCSN), datasets change frequently. Errors are corrected, new data are added, scales are defined and refined, and results are analyzed in different ways.

“Let’s say you have done research and you have analyzed your data. You have written a manuscript, and you have submitted it for publication. Then, several months later, the manuscript is returned to you, and the reviewer suggests that you conduct a different analysis, or exclude certain subjects, or include additional variables to see whether the results change. In order to do this, as a first step, you must be able to reconstruct what you actually did in the first place. Ideally, the numbers you get today should be the same as those in the manuscript you submitted several months ago. If you understand where those numbers came from, and all of the decisions on which they depend, you can also go forward with a further analysis. But this is not always the case,” explains Thisted.

The experimental results and the data presented in tables or figures embedded within a manuscript are the end result of computations performed by analytic computer programs. The typical manuscript will be based on hundreds of lines of program instructions—computer code—that embody decisions such as the correction of coding errors, treatment of outliers, transformations of variables, coding of categories, construction of models, and sensitivity analyses. This code executes



ABOVE: Ronald Thisted, Professor of Biostatistics and Chairman of the Department of Health Studies at the University of Chicago.

the commands performed on the dataset. Therefore, maintaining records of the specific version of the code used to obtain the results, as well as the specific version of the dataset used in the original analysis, is critical.

“It is very helpful to maintain an archived copy of the exact dataset used to create a manuscript, and the exact code that was executed to obtain the submitted results, starting from the original raw dataset, so that a researcher can know exactly where every number in the manuscript comes from. For example, a researcher looking at relationships may have run the same analysis on married and single people and compared the results. In the initial analysis, he may have included couples in his sample that were not formally married in the group of married people or in the group of single people, and this initial choice must be recorded. This allows us to reconstruct subtleties of an analysis that cannot

always be included in the finished manuscript,” explains Thisted. It also makes it possible for a new collaborator to build upon earlier work.

Often, the primary person who benefits from reproducible research is the original author. Matthias Schwab and Jon Claerbout, geophysicists from Stanford University who are credited with conceptualizing much of the contemporary approach to reproducible research, stated, “One of the main tenets of reproducible research is that time turns each one of us into another person. By making an effort to communicate with strangers; we help ourselves to communicate with our future selves.” (Schwab and Claerbout, Stanford SEP, *Making Research Reproducible*, 1996).

Thisted explains, “For me, this is the most important reason to structure our research practices in a reproducible way. If I have a figure in a paper, I would like to know how to get

**ONE OF THE MAIN TENETS OF REPRODUCIBLE RESEARCH IS THAT TIME TURNS EACH ONE OF US INTO ANOTHER PERSON. BY MAKING AN EFFORT TO COMMUNICATE WITH STRANGERS; WE HELP OURSELVES TO COMMUNICATE WITH OUR FUTURE SELVES.**

SCHWAB AND CLAERBOUT

## FURTHER DISCUSSION OF REPRODUCIBLE RESEARCH



### CTSPEDIA.ORG: Clinical and Translational Science

CTSpedia was created as a national effort to collect wisdom, tools, educational materials, and other items useful for clinical and translational researchers and to provide timely and useful advice to clinical and translational researchers with specific problems.

The CTSpedia working group on Reproducible Research (RR) holds annual teleconferences to develop resources and practices. More information can be found at [www.ctspedia.org](http://www.ctspedia.org).

## IMMEDIATELY, THIS TAKES A PIECE OF REAL SCIENTIFIC WORK AND MOVES IT INTO A MORE IDEALIZED VISION OF SCIENCE.

RON THISTED

that figure again. If I want to make a figure just like it, or if I need to update the figure with additional data, I know how to do it. Just communicating with my future self is enough reason to develop these tools and ways of working.”

### *The Mechanisms of Reproducible Research*

A large part of doing reproducible research comes down to organizing work in a deliberate and systematic way. Workflow should proceed clearly from the raw data (i.e., as recorded in the lab, entered from a case report form or questionnaire, exported from a machine, or received from an archive) through the necessary data manipulations to analyses, with the raw data always preserved inviolate. The entire process should be broken down logically into distinct steps, with each step performed by a specific block of code or separate script(s). Concise but thorough documentation describing how to repeat the process (e.g., a simple README.txt file indicating which scripts to execute and in which order) should be created. Finally, all files (including the raw data) should be given informative names and stored together, using subdirectories as necessary for organization. As a quick check, a researcher could consider whether a colleague presented with the entire package could determine how to replicate the final results from the raw data, without additional assistance. These steps, properly executed, can help make research reproducible, and any researcher can use them without any additional software or expert knowledge.

Several software tools can facilitate reproducible research, especially in the collaborative setting. For example, consider multiple authors collaborating on a manuscript and making changes simultaneously. In this case, simply keeping track of which version

of the document is the most complete and up-to-date becomes difficult, even without the added difficulty of incorporating individual changes into one final manuscript. For example, consider the challenge for five co-authors revising a manuscript, with round-robin “tracked changes” in a Microsoft Word document. Fortunately, software tools known as version control systems used by software companies to manage and track their code can be used to manage this problem (Mercurial is Thisted’s current favorite, and is available for no cost). Such systems offer powerful tools for tracking and inspecting changes to files, and for sharing and merging together edits made by multiple people. In cases where the changes do not conflict, merging can often be accomplished automatically, and in cases where they do conflict, the software provides tools to facilitate resolving the conflicts. Thus, version control systems can make collaboration easier and more accurate, enhancing the reproducibility of the work in the end. Version control can also be applied to data files and analysis code, allowing a researcher to recreate the state of an entire project at any point in time, and thereby replicate intermediate results at any point in a project.

Other tools that support reproducible research make it possible to create flexible, multipurpose documents. These tools allow researchers to create multiple types of documents from one master version.

“If I write a set of exercises for class, I require several versions: one with solutions to the problems, and one without solutions,” explains Thisted. “I may also want an HTML version of the exercises to put on a website, or a PDF version for printing. Ideally, I should be able to create one master document from which I can easily generate the specific type of output I require. I should be able to use a few keystrokes

to generate an HTML version without answers, or a PDF version with the answers included, and so on. If this multipurpose document were sufficiently sophisticated, it could also be used to generate both a manuscript ready for submission and all of the computer code that was used to generate the figures, numbers, and tables in that manuscript.”

A software tool called Sweave facilitates the creation of such a document, using the R statistical programming language and LaTeX typesetting language. Thisted and colleagues are currently working on ways to extend this basic approach to make it easier to use, and to permit it to be used with various analytic packages (e.g., Stata, SPSS) and authoring environments. Researchers adopting this approach will find themselves able to create manuscripts that are immune to cutting and pasting errors, and that make it possible for independent scientists to understand and verify every result they contain.

“Baring one’s scientific soul in this way may be intimidating to an investigator,” Thisted explains, “but it should not be. If the process works, then this collection of code, data, and interpretive material—the glue that binds the pieces together in a finished manuscript—completely documents what was done and how to do it. Immediately, this takes a piece of real scientific work and moves it into our more idealized vision of science, where others can fully build on what we have done.”

### *The Importance of Reproducible Research*

“In the best of all possible worlds, this is a part of transparency in science. Transparency is important for science because scientists are human. They make mistakes. They make choices, and sometimes those choices are bad. If it is impossible to discover errors or to identify all of the choices underlying scientific reports, one simply cannot tell whether the results presented are a consequence of good science or poor choices,” says Thisted.

In the competitive environment of contemporary scientific inquiry, full datasets may not be available for independent review.

However, to make rigorous peer review possible in settings where conclusions rely on nuanced statistical analysis, many major journals, including *Science*, *Nature*, the *Journal of the American Medical Association*, and the *Annals of Internal Medicine* require some access to datasets and analysis, to ensure that the results published are sound. In doing so, these journals hope to avoid the compromising situation of publishing erroneous or even fraudulent results.

Thisted is optimistic that these practices will continue to improve over time. As he explains, “Ideally, anyone should be able to recalculate all analyses performed in a study, starting from the original source data. Some think we will get there soon, but any steps we take to support reproducible research, even just mindfully aiding our future selves through better organization, are worth taking.” ■

BELOW: Phil Schumm, Biostatistician in the Department of Health Studies at the University of Chicago.



## RECENT PUBLICATIONS FROM RON THISTED

Vanderweele, T.J., Hawkey, L.C., Thisted, R.A., & Cacioppo, J.T. (2011). A marginal structural model analysis for loneliness: Implications for intervention trials and clinical practice. *Journal of Consulting and Clinical Psychology*, 79(2):225-35.

Cacioppo, J.T., Hawkey, L.C., & Thisted, R.A. (2010). Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the Chicago Health, Aging, and Social Relations Study. *Psychology and Aging*, 25(2):453-63.

Hawkey, L.C., Thisted, R.A., Masi, C.M., & Cacioppo, J.T. (2010). Loneliness predicts increased blood pressure: Five-year cross-lagged analyses in middle-aged and older adults. *Psychology and Aging*, 25(1):132-141.

Luhrmann, T.M., Nusbaum, H., & Thisted, R. (2010). The absorption hypothesis: Learning to hear God in evangelical Christianity. *American Anthropologist*, 112(1):66-78.

Pioro, E.P., Brooks, B.R., Cummings, J., Schiffer, R., Thisted, R., Wynn, D., Hepner, A., & Kaye, R. (2010). Dextromethorphan plus ultra-low-dose quinidine reduces pseudobulbar affect. *Annals of Neurology*, 68(5):693-702.

Hawkey L.C., Thisted R.A., & Cacioppo, J.T. (2009). Loneliness predicts reduced physical activity: Cross-sectional & longitudinal analyses. *Health Psychology*, 28(3):354-63.

