Module 1 – lesson 01

Script

For this first lesson, I’m going to review a brief history of reproducible research and discuss why reproducible research is so important for ensuring transparency and accuracy when reporting scientific findings. I will highlight significant events that have occurred in the last decade giving momentum to the reproducible research movement. I will also list several key people to be aware of that I will refer to throughout this course and who are good to know regarding reproducibility. I also provide a list of books that are great references for learning more on your own. This lesson wraps up with a specific overview of literate programming and dynamic documentation and how the pieces fits together to form the foundation you will utilize throughout this whole course.

One of the earliest modern references to reproducible research was made by Jon Claerbout in 1992 in his book on earth soundings analysis (he is a geophysicist and now a professor emeritus at Stanford University). In his book he states that “Today, few published results are reproducible in any practical sense. To verify them requires almost as much effort as it took to create them originally. After a time, authors are often unable to reproduce their own results! For these reasons, many people ignore most of the literature.”

To address the reporting of clinical trials, the CONSORT statement was introduced in 1996. “CONSORT stands for Consolidated Standards of Reporting Trials and encompasses various initiatives developed by the CONSORT Group to alleviate the problems arising from inadequate reporting of randomized controlled trials.” Then in 2004, the International committee of Medical Journal Editors (or the ICMJE) stated that they would not publish a clinical trial that had not been registered. The ICMJE accepts registration with the WHO (world health organization) International Clinical Trials Registry Platform (ICTRP) or in ClinicalTrials.gov. This group endorses these registries because they meet several key criteria:

* they are accessible to the public at no charge,
* they are open to all prospective registrants,
* they are managed by a not-for-profit organization,
* they have a mechanism to ensure the validity of the registration data, and
* they are electronically searchable.

Another key event was the 2005 paper by John Ioannidis on “Why most published research findings are false” which is said to be “one of the most downloaded articles in the history of the Public Library of Science.”

In 2007 the FDA (food and drug administration) ‘s amendments act required more clinical trials to be registered. Then 2 years later in 2009, the Journal of Biostatistics instituted a policy for accepting and marking publications that meet certain standards of reproducibility. Their author instructions for reproducible research states that “… papers (will be) marked (with a) D if the data on which they are based are freely available, (marked with a) C if the authors’ code is freely available, and (the gold star marked with a) R if both data and code are available, and our Associate Editor for Reproducibility is able to use these to reproduce the results in the paper. (they also publish) Data and code on the journal’s website as Supplementary Materials.”

In 2011, Alsheikh-Ali, et.al., reported that “A substantial proportion of original research papers published in high-impact journals are either not subject to any data availability policies, or do not adhere to the data availability instructions in their respective journals.”

They assessed 500 research papers – of these 149 (30%) were not subject to any data availability policy. Of the remaining 351 papers that were covered by some data availability policy, 208 papers (59%) did not fully adhere to the data availability instructions of the journals they were published in, most commonly (73%) by not publicly depositing microarray data. The other 143 papers that adhered to the data availability instructions did so by publicly depositing only the specific data type as required, making a statement of willingness to share, or actually sharing all the primary data. Overall, only 47 papers (9%) deposited full primary raw data online. The rest of the papers were not subject to data availability policies and they did not make their primary data publicly available.

Also in 2010, a huge controversy erupted over a cancer clinical trial at Duke University meant to align chemotherapy treatments with patient’s gene types. However, when 2 researchers at MD Anderson (Keith Baggerly and Kevin Coombs) decided to check the work done by Duke, they found numerous spreadsheet errors leading to mis-alignment and incorrect assignment of cancer treatment therapies. Four papers published by the Duke team were retracted, the Duke lead scientist resigned, Duke shut down 3 other trials using these results, and many patients have pursued legal action. The 2010 video presentation by Keith Baggerly is very well done and details all of the issues they encountered in trying to reproduce the work done by Duke on the cancer trial.

Another very famous study which could not be reproduced – often called the “excel-error heard round the world” – was based on a paper by 2 well known economists, Kenneth Rogoff and Carmen Reinhart entitled “Growth in a Time of Debt”. In this paper, they claimed that countries whose debt exceeds 90 percent of their annual GDP (gross domestic product) experience slower growth than countries with lower debt loads — a figure that has been cited by many people (like Paul Ryan and Tim Geithner) to justify slashing government spending and implementing other austerity measures on struggling economies.

But when Thomas Herndon, a 28-year-old economics graduate student at University of Massachusetts Amherst tried to reproduce the results published by Rogoff and Reinhart he discovered a major formula error in their excel data spreadsheet. The original paper had excluded key data from the countries of Canada, New Zealand, and Australia — all countries that experienced solid growth during periods of high debt and thus undercut the conclusion that high debt forestalls growth.

From 2012 forward, even more publications have emerged noting that scientific findings cannot be confirmed. In 2013 the Center for Open Science was founded and then established the Open Science Framework. In 2014 the NIH (national institutes of health) issued their guidelines for reproducibility. Even by 2015, the Open Science Collaboration reveals that they were only able to reproduce or replicate between 30-50% of the results from over 100 studies.

Recently in 2016, Ziemann, et.al. noted that 1 in 5 papers (20%) published in leading genomics journals have supplementary data files containing erroneous gene name conversions due to Microsoft Excel default settings. This 20% is an average with some journals having higher rates (next slide).

This next set of slides list key people in the reproducible research arena - I strongly encourage you to read their publications and books. First on this list is Victoria Stodden. She is a professor at the School of Information Sciences at the University of Illinois at Urbana-Champaign. She has dual degrees in Statistics and Law. She is a leading figure in the area of reproducibility in computational science, exploring how we can better ensure the reliability and usefulness of scientific results. Her work addresses issues like standards of openness for data and code sharing, legal and policy barriers for disseminating reproducible research, robustness of replicated findings, cyberinfrastructure needed to enable reproducibility, and related scientific publishing practices.

Roger Peng is a Professor in the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health. He is an avid speaker and writer of books and other materials on the importance and how to implement reproducible research. In addition to his co-authorship of the “Implementing Reproducible Research” book, he has several other self-published books listed on the bookdown.org website. He is also the associate editor for reproducible research for the Journal of Biostatistics.

John Ioannidis (yonnadis) is also widely published on the importance of reproducibility and as already mentioned was the author of the 2005 paper entitled “Why most published research findings are false” which is said to be “one of the most downloaded articles in the history of the Public Library of Science.”

Christopher Gandrud is a research fellow at the Institute for Quantitative Social Science at Harvard University. His book on “Reproducible Research with R and RStudio” is a key test for putting all of the pieces together from start to finish keeping all data, code and documentation together and organized. In this book, he does a good job of laying out efficient workflows and recommended practices for making your work easily reproducible.

Yihui Xie (Yeewhay she) is a software engineer with RStudio and author of numerous books and R packages for publishing reproducible reports and dynamic documents including making and publishing books (with the bookdown package) and blogs (with the blogdown package). I will introduce you to both the bookdown and blogdown packages during this coursera course.

Finally, one more name to keep in mind is Federich Leisch. He is a statistics professor at the University of Natural Resources and Life Sciences in Vienna, Austria. He is the creator and lead developer of the SWEAVE package for creating dynamic reports combining code and documentation together in one seamless process. Yihui Xie’s (Yeewhay she)’s knitr package expands upon these methods and principles implemented in SWEAVE. He is also the 3rd co-author of the “Implementing Reproducible Research” Book.

The 3 books listed here have already been discussed but this slide provides the website links for each book published by CRC Press.

Additional books you may want to also explore are listed on this next slide. Specifically, you will most definitely want to read through and bookmark for future reference Jenny Bryan’s book (which is free and online) entitled “Happy Git and Github for the user (R user)”. This book will be very helpful getting Git and Github setup and working with R and RStudio – all of which you will need to complete this course. I will provide more specifics and details on all of these in the upcoming lessons and modules.

For those of you interested in utilizing templates for reproducible reports of your data analyses, I encourage you to also check out the book by Nicholas Horton and Ken Kleinman on “Using R and RStudio for Data Management” published by CRC Press; as well as checking out their Project MOSAIC, which is a community of educators working to develop a new ways to introduce mathematics, statistics, computation and modeling to students in colleges and universities.

Chester Ismay has also been instrumental in helping people get up and running using R and RStudio for reproducible reports and templates. In this course we will explore some of his rmarkdown templates and how they were constructed for use at Reed College, where he was an Instructional Technologist/Statistical Consultant. He is currently the Data Science Curriculum Lead for DataCamp (an online training platform for data science).

And finally, check out the numerous books published online and in development at bookdown.org. The books provided here are all free and readily accessible.

To get us ready for lesson 02, I want to also highlight some of the key events and history of literate programming and dynamic documentation. All of these events and developments have led up to the rmarkdown package which will be the foundation of all of our work in this coursera course going forward.

In 1992 around the same time Jon Claerbout coined the term “reproducible research”, Donald Knuth introduced the concept of “literate programming” for which “The main idea is to treat a (software/computer) program as a piece of literature, addressed to human beings rather than to a computer.”

As we’ve already mentioned, in 2002 Federich Leisch released the SWEAVE program for dynamic documentation generation.

The primary “markdown” package was introduced in 2004 by John Gruber to make it much easier to “markup” plain text files for generating HTML documents – ideally without having to learn HTML (hypertext markup language). The rmarkdown package you will use throughout this course is built upon John Gruber’s “markdown”.

Then in 2012 Yihui Xie (yeewhay she) released knitr which was inspired by SWEAVE and expanded its capabilities to work with the R language combining code and text producing dynamic documents.

In 2014, rmarkdown was released. Rmarkdown is the foundation for all of the reproducible templates you will learn in this course.

So, here is the BIG PICTURE. Everything we will do in this course happens through the RStudio IDE, integrated development environment. From here, we will be able to utilize version control and file tracking using Git linked to the cloud based Github repository. We will implement Dynamic Documentation and our reproducible templates will use the rmarkdown and knitr packages in R. We will briefly touch on a few R programming and analysis methods, but this is not the focus of this course and will be kept to a minimum.

Next in lesson 02, we will get more into the details and specifics for literate programming and dynamic documentation.