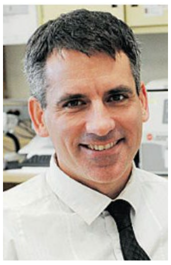

Article:
Janet E. Simons and Daniel T. Holmes.
Reproducible Research and Reports with R.
J Appl Lab Med 2019;4:471-3.
http://jalm.aaccjnls.org/content/4/3/471

Guest: Dr. Daniel Holmes is a Clinical Professor of Pathology and Laboratory Medicine at the University of British Columbia and is the Division Head of Clinical Chemistry at St. Paul's Hospital in Vancouver.

Randye Kaye: Hello, and welcome to this edition of "JALM Talk," from the *Journal of Applied Laboratory Medicine*, a publication of the American Association for Clinical Chemistry. I'm your host, Randye Kaye.

Spreadsheet programs such as Microsoft Excel have long been the mainstay of data analysis in many settings. In the clinical laboratory, Lab Directors and Managers often use spreadsheets exported from laboratory information systems to monitor quality indicators such as test volumes and result turnaround times. Spreadsheets are also used to analyze method verification studies or to conduct clinical research.

However, spreadsheet programs are limited in that they do not track data manipulation and mistakes may not be traceable. Further, spreadsheets cannot support particularly large data sets. Thus, clinical laboratorians and medical researchers are increasingly turning to R statistical programming language to analyze data. One free and open source product called R Markdown, allows for the creation of fully reproducible reports combining both data analysis code and text narrative. The November 2019 issue of *The Journal of Applied Laboratory Medicine* includes a Laboratory Reflections: Technical Tips article entitled, "Reproducible Research and Reports with R." The senior author of the letter is Dr. Daniel Holmes, who is our guest on this podcast. Dr. Holmes is a Clinical Professor of Pathology and Laboratory Medicine at the University of British Columbia and is the Division Head of Clinical Chemistry at St. Paul's Hospital in Vancouver. Welcome Dr. Holmes.

Dr. Daniel Holmes: Thank you for having me.

Randye Kaye: So, let's begin here. What is meant by the term, "Reproducible Research?"

Dr. Daniel Holmes: This is the term that is common to scientific literature to describe a study that fully exposes everything that has been done to put the research paper together or put the analysis together. So, to help you understand, let's think about what is not reproducible research. So, when I use a spreadsheet or a graphical user interface-based program to do any kind of manipulation on my data, whether it's mass spectrometry data or genomic data or just clinical trials data, I don't really know what I've done because I don't record every mouse move. I know in my head what I think I've done but six months from now, I might find that I deleted certain types of outliers in a particular way, but I don't have any definitive record of that because we don't trace mouse moves when we do analysis. So, imagine if you did your entire analysis in computer code, which sounds intimidating, but it’s not quite as bad as you think.

So, I have my initial data set, completely untouched and I apply this computer program that I've written to do all of the manipulation of the data defining the outliers and removing them and cleaning the data set and producing the images and producing the tables and producing the P-values or what have you; if everything I've done is fully exposed by that code and the code is easy to follow, then I could redo the entire analysis six months later, starting from the same data set and I would know exactly what I did if I was willing to go and read the code. So, reproducible research is research where everything that you've done is fully transparent because none of it takes place by means of a graphical user interface and mouse moves.

Randye Kaye: So, why is the concept of reproducible research so important?

Dr. Daniel Holmes: I think you can understand the value of reproducible research if you see what has happened in the absence of it. There's a wonderful YouTube video of a lecture from Keith Baggerly, who was a statistician at MD Anderson. He is now retired, but he did a forensic analysis of a ovarian cancer study based at Duke University, and discovered that the researchers involved in that study had made cutting and pasting errors between graphical user interfaces in their analysis and were randomizing patients to the wrong arm based on what amounted to frameshifted genomic data. And, as you can imagine, when you make an error like that and you’re making clinical decisions based on it, and you're treating people based on it, everything is wrong.

He had to work from the original data sets to see that they had made these errors. And so, I think the principle is that the data sets that we’re getting are so big from whether it's proteomics data, genomic data or even archived data from lab information systems. These are enormous files. We’re trying to do things like AI and sophisticated, predictive models with these data.

But, when we don't do the analysis in a completely transparent manner, errors that we have made, not intentional of course, but errors that we have made just because of our own humanness can lead to catastrophic mistakes. And so, it's important that the researcher be able to show exactly what they have done right from start to finish in order that they can be sure that they're making the right clinical inferences based on their analysis.

A simple example that everybody has experienced is very tangible, is when you're entering data in Excel and it might be LOINC codes, it might be gene names. Everybody has had Excel convert gene names or LOINC codes or something similar to a date when it isn't a date. You know, Clippy thinks he is doing you a favor when he does that but a lot of time, you just want to reach into the screen and strangle Clippy. There's a known and published pollution of gene data and public repositories from Excel converting gene names to dates. So, it's like if your gene happened to be called NOV 11 or SEPT 12 or something like that. And so, reproducible research –

Randye Kaye: Wow.

Dr. Daniel Holmes: Yeah, I know. There're whole papers about that. Reproducible research prevents that from occurring because everything is kind of locked down. It's more work to do it, it's less intuitive, it's harder intellectually, but it produces a product where you kind of understand the behavior.

Randye Kaye: Okay. So, your article talks about using R for Reproducible Research and Reports, and tell me what tools would someone in the audience need to get to get started using R for data analysis?

Dr. Daniel Holmes: Well, the nice thing about R, and I'll just throw it out there, that Python is a perfectly valid alternative, and in some ways is more flexible. R and Python are both free and open source tools. That means that you can download them and install them as long as you have administrative rights on your computer and if you don't have administrative rights on your computer, there's actually ways to interact with R entirely through a browser and in cloud environment. Obviously, you wouldn't be able to do any analysis on data with personal health identifiers and in that context, you need a computer on which you have administrative rights and you could use R or you could use other tools like Python.

In addition to the language itself, you need an editor to create the code and the most easy way in is the R Studio interactive development environment. It's also free for use for a single individual. So, you can download that from the R studio website. That will get you started with the software. Getting the software is just the first step. You have to learn how to use the software.

So, I have a few recommendations. Probably, the easiest way to get started in terms of doing the programming itself is to take an online course or to take a face-to-face course. One of the most popular online courses is from a company called DataCamp. It produces videos with a coding environment where you can practice the questions that they pose to you and you can do your learning right online there. It's very high quality, slick, pretty easy entry. You can also do a face-to-face course.

So, there are face-to-face courses offered at the American Association of Clinical Chemistry meeting and also, offered at the MSACL meeting in the United States, and if you Google around, there's usually courses that your local university intended for researchers or graduate students or people in biomedical research to get introduced how to code in R.

In terms of online resources, I was quite dependent on a website called statmethods.net when I got started. And that was about 10 years ago. The language has changed a fair bit since then, but that's still a very valuable online tool. In addition, there is good old-fashioned books. A good example is, "Introductory Statistics with R," written by John Verzani. That's focused on statistics, but you have to teach the language in order to teach how to do statistics with it.

Randye Kaye: Okay. So, you have mentioned quite a lot of resources and you may have already answered this, but are there any additional resources for learning how to use R and R Markdown?

Dr. Daniel Holmes: Well, I've mentioned resources on how to use R. R Markdown is a tool that permits you to write an entire report reproducibly. So, you write your text -- you know, your free text description of your methods or what have you, and then you can insert blocks of code. Those blocks of code are hidden in the final report or if you prefer you can have them exposed. In any case, those code blocks will generate your figures or your table for you and insert it automatically in. The output will be a PDF or a Word document or a PowerPoint presentation or HTML document, whatever you like as your final report.

The tool R Markdown is part of the suite of R tools, particularly targeted at producing finalized reports. The definitive book is called, "R Markdown, The Definitive Guide," and the author is Yihui Xie, and that book is online.

Randye Kaye: Okay, wonderful. So, let me talk about your medical practice for a moment. How would you say that the use of R has impacted that, your medical practice?

Dr. Daniel Holmes: So, the more you learn, the more you see opportunities to use R as an automation tool in your work environment. For example, we used to manually transcribe all of our mass spectrometry results from the instrument to the lab information system. But at some point, I realized, this is crazy, we're doing all this copying and pasting. We could just write a script that takes the flat file off the mass spectrometer, reprocesses it, shapes at the way it should be and pushes it over to a folder for the lab information system to absorb it. So, we did that, that was one thing that we did. Two weeks ago, I had to anonymize 500 million patient records. So, I just wrote a script that pulled out all their names and created fake names that I could match back to their real names and I hashed all the PHN, PHI et cetera. And so, those are the kinds of activities that you can do.

Believe it or not, even our requisition scanning software, which scans the outpatient requisition, finds the bar code, convert it to a PDF, stores it in a folder by today's date, is written in R. Probably not a typical application of R but if you only have a hammer, all you see is nails. So, I mean, I use R literally every day whenever a spreadsheet-like data is longer than a few pages, or the manipulations I need to make are kind of sophisticated. So, I still open spreadsheets; obviously, we all do, we receive them. But any time that we need to do something that's repetitive or involves large data sets, we do it in R. For example, our monthly key performance indicator reports.

They used to be dutifully put together by somebody cutting and pasting from data aggregated from the lab information system. Now they are completely, automatically generated with an R script based on extracts that are provided each month by the lab information system people. So, it's a way of getting rid of mundane tasks and lowering the error rate associated with them, and nobody likes doing mundane tasks, I mean, maybe some people do, but I don't like doing mundane, repetitive tasks, and so we use R to take care of anything like that.

Randye Kaye: So, you're definitely saving time and increasing efficiency and accuracy, sounds like.

Dr. Daniel Holmes: Yeah, the real concern is making errors and so, in some processes, we flushed out all of the potential for errors. But of course, when you code things, you have to test them, right? You become the sole person responsible for the accuracy. So, a lot of testing of in-house custom software is required. The other danger is if you program something, you are the go-to person if it ever needs fixing due to some change in the environment and so, that can be a bit burdensome and a person needs to be careful how much they commit to writing custom software for others to use.

One last thing that I do is I write academic manuscripts in R and an example is this article that we're discussing. It's written entirely in R and the code generates the final article. The one thing that's nice about this is, we've all been in this situation where we have to insert a table, we have to insert a figure, or we have to add some new data points, or subtract data points and then redo the statistical analysis. If you write the paper in R, even the calculations that are within a sentence can be coded so that if the original data set were modified, all of the propagated changes will occur automatically, the changes within the a sentence, the changes in the tables, the changes in the figures, the changes in the references, the changes in the cross-references, will occur automatically if you run that script again. So, even though it's harder to build and harder to write, change has become much, much easier when you've done your report, or your paper in this case, reproducibly.

Randye Kaye: Wonderful. Well, very interesting and very interesting article. Thank you so much for joining us today.

Dr. Daniel Holmes: No, you're very welcome.

Randye Kaye: That was Dr. Daniel Holmes from the University of British Columbia discussing his Laboratory Reflections: Technical Tips article entitled, "Reproducible Research and Reports with R," from the November 2019 issue of JALM. Thanks for tuning in to this episode of "JALM Talk." See you next time and don't forget to submit something for us to talk about.

Resources Provided by Dr. Holmes:

Why reproducible research is important - by looking at what happens when it is absent:

Keith Baggerly's (very illustrative) Narrative on the Forensic Analysis of Duke University Data: https://www.youtube.com/watch?v=lrm8iEMQZNw&t=338s
How Excel creates errors in the Genetic Literature:
https://www.sciencemag.org/news/2016/08/one-five-genetics-papers-contains-errors-thanks-microsoft-excel
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7?utm_content=buffer67f08&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Software:

R Language Website: https://www.r-project.org/
R Studio IDE: https://rstudio.com/
Using RStudio with no local software installation: https://rstudio.cloud/

Educational Tools:
Data Camp: https://www.datacamp.com/
Statmethods.net: https://www.statmethods.net/
Using R for Introductory Statistics by John Verzani: https://www.crcpress.com/Using-R-for-Introductory-Statistics/Verzani/p/book/9781466590731
RMarkdown: The Definitve Guide by Yihui
Xie: https://bookdown.org/yihui/rmarkdown/