Module 1 – lesson 01

So why is reproducible research important and incredibly beneficial? Let's take a look at its beginnings to find the answer.

In the early '90s, a geophysicist named Jon Claerbout revised his book *Earth Soundings Analysis* with a valid complaint. He claimed that few published results are reproducible in any practical sense. To verify them requires almost as much effort as it took to create them in the first place. After a time, even the authors are often unable to reproduce their own results! For these reasons, many people ignore most of the literature.

Then, in 1996, the Consolidated Standards of Reporting Trials (or CONSORT) published a set of guidelines to fix problems that developed from inadequate reporting of randomized controlled trials. Following their lead, in 2004, the International Committee of Medical Journal Editors stated they wouldn't publish a clinical trial that had not been registered, and that they would only endorse registries meeting several key criteria, including that they must be:

* free and publicly accessible,
* open to all prospective registrants,
* managed by a non-profit organization, and
* electronically searchable and validated

As a result of this turmoil in validity and research, the Food and Drug Administration got on board and required the registration of even more clinical trials. The *Journal of Biostatistics* also encouraged reproducible practices of author submissions. They began marking accepted papers based on the standards of reproducibility that were followed. For example, papers marked with a 'D' marking meant the data on which the study is based is freely available. A 'C' marking means the authors' code is freely available. And an 'R' marking is the gold standard, meaning that not only are the data and code freely available, but the associate editor for reproducibility was also able to reproduce the same results as the paper.

An example of an article given the highest designation for a fully reproducible article is one published in 2009 entitled “Air pollution and health in Scotland: a multicity study.” To see the article's marking, you have to download the PDF and look for the marking letter in a bold box at the top right.

Most compelling, in the early 2000s, John Ioannidis published an article with the highest downloads in the history of the Public Library of Science. It was entitled "Why most published research findings are false." Despite multiple organizations attempting to fix this issue, the Open Science Collaboration revealed in 2015 that they were only able to reproduce or replicate between 30-50% of the results from more than 100 studies. Ziemann, et.al. in 2016 also found that 20% of papers 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, and some journals have even higher rates.

But this isn't new. In 2011, Alsheikh-Ali [Al shake], et.al. assessed 500 research papers with unsettling results. Of these 500 papers sent to high-impact research journals, 30% were not subject to any data availability policy. The 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 (that's only 9%!) deposited full primary raw data online.

In the last several years, reproducibility errors have been at the center of some major controversies. In 2010, a published cancer clinical trial at Duke University was tested by two MD Anderson researchers, Keith Baggerly and Kevin Coombs. They found numerous spreadsheet errors leading to misalignment and incorrect assignment of cancer treatment therapies. Because of this, four papers published by the Duke team were retracted, the Duke lead scientist resigned, and Duke shut down three other trials using these results, and many patients have pursued legal action.

Another famous study, often called the “excel-error heard round the world” – was based on a paper by two well-known economists at Harvard, Kenneth Rogoff and Carmen Reinhart. In their paper “Growth in a Time of Debt”, the authors claimed that countries whose debt exceeds 90 percent of their annual gross domestic product experience slower growth than countries with lower debt — a figure that's been cited by many people in order to justify slashing government spending. But when Thomas Herndon, a 28-year-old economics graduate student at the University of Massachusetts tried to reproduce the results, he discovered a major formula error in the 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.

Due to these critical moments in the last few decades, reproducibility continues to grow as new policies are adopted and the practices are applied. But part of the benefit of being in this course is that you get to be part of the reproducible research movement!