Reproducible Medical Research with R

Peter D.R. Higgins, MD, PhD, MSc

2020-04-15

Contents

Chapter 1

Preface

Welcome to Reproducible Medical Research with R (RMWR). I hope that this book meets your needs.

1.1 Who This Book is For

This is a book for anyone in the medical field interested in analyzing the data available to them to better understand health, disease, or delivery of care. This could include nurses, dieticians, psychologists, and PhDs in related fields, as well as medical students, residents, fellows, or doctors in practice.
I expect that most learners will be using this book in their spare time at night and on weekends, as the medical school curriculum is already packed full, and there is no room to add skills in reproducible research to the standard curriculum. This book is designed for self-teaching, and many hints and solutions will be provided to avoid roadblocks and frustration. Many learners find themselves wanting to develop reproducible research skills after they have finished their training, and after they have become comfortable with their clinical role. This is the time when they identify and want to address problems in their practice with the data they have before them. This book is for you.

1.2 The Spiral of Success Structure

This book is structured on the concept of a "spiral of success", with readers learning about topics like data visualization, data wrangling, data modeling, reproducible research, and communication of results in repeated passes. These will initially be at a superficial level, and at each pass of the spiral, will provide increasing depth and complexity. This means that the chapters on data wrangling will not all be together, nor the chapters on data visualization. Our goal is

to build skills gradually, and return to (and remind students of) their previously built skills in one area and to add to them. The eventual goal is for learners to be able to produce, document, and communicate reproducible research to their community.

1.3 Motivation for this Book

Most medical people who learn R to do their own data analysis do it on their own time. They rarely have time for a semester-long course, and their clinical schedules usually will not allow it. Fortunately, a lot of people learn R on their own, and there is a strong and supportive R Community to help new learners. A 2019 Twitter survey conducted by `@RLadies` found that more than half of respondents were largely self-taught, from books and online resources.

There are a lot of good resources for learning R, so why one more? In part, because the needs of a medical audience are often different. There are distinct needs for protecting health information, generating a descriptive Table One, using secure data tools like REDCap, and creating standard medical journal and meeting output in Word, Powerpoint, and poster formats.

More and more, all science is becoming data science. We are able to track patients, their test results, and even the individual pixels (voxels) of their CT scans electronically, and use those data points to develop new knowledge. While one could argue that health care workers should collect data and bring it to trained statisticians, this does not work nearly as well as you might expect. Most academic statisticians are incentivized to develop new statistical methods, and are not very interested (or incentivized) to do the hand-holding required to wrangle messy clinical data into a manuscript.

There also are simply not enough statisticians to meet the needs of medical science. Having clinicians on the front lines with some data science training makes a big difference, whether in 1854 in London (John Snow) or in 2014 in Flint, Michigan (Mona Hanna-Atisha). Having more clinicians with some training will impact medical care, as they will identify local problems that would have otherwise never reached a statistician, and probably never been addressed with data otherwise.

1.4 The Scientific Reproducibility Crisis

Beginning as far back as 1989, with the David Baltimore case, and increasingly publicly through the 2010s, there has been a rising tide of realization that a lot of taxpayer-funded science is done sloppily, and that our standards as scientists need to be higher. The line between carelessly-done science and outright fraud is a thin one, and the case can be made that doing science in a sloppy fashion

defrauds the funders, as it leads to results that can not be reproduced nor replicated. Particularly in medicine, where incorrect findings can cause great harm, we should take special care to do scientific research which is well-documented, reproducible, and replicable. This topic as a motivating force for doing careful medical research will be expanded upon in Chapter 1.

1.5 What this Book is Not

1.5.1 This Book is Not A Statistics Text

This is not an introduction to statistics. I am assuming that you have learned some statistics somewhere in secondary school, undergraduate studies, graduate school, or even medical school. There are lots of statisticians with Ph.D.s who can certainly teach statistics much more effectively than I can. While I have a master's degree in Clinical Research Design and Statistical Analysis (isn't that a mouthful!) from the University of Michigan, I will leave formal teaching of statistics to the pros.

If you need to brush up on your statistics, no worries. There are several excellent (and free!) e-books on that very topic, using R. Some good examples include (go ahead and click through the blue links to explore):

1. Learning Statistics with R (LSR)
2. Modern Dive
3. Teacup Giraffes

We will cover a lot of the same materials as these books, but with a less theoretical and more applied approach. I will focus on specific medical examples, and emphasize issues (like Protected Health Information) that are particularly important for medical data. I am assuming that you are here because you want to analyze your own data in your probably very limited free time.

1.5.2 This Book Does Not Provide Comprehensive Coverage of the R Universe

This book is also far from comprehensive in teaching what is available in the R ecosystem. This book should be considered a launch pad. Many of the later chapters will give you a taste of what is available in certain areas, and guide you to resources (and links) that you can explore to learn more and do more beyond the scope of this book.

1.6 Some Guideposts

Keep an eye out for Guideposts, which look like this:

Warnings

This is a common gotcha. Watch out for this.

Tips

This is a helpful tip for debugging.

Try It Out

Take what you have learned and try it yourself in the code box below.

Challenge - take the next step and try a more challenging example.

Try this more complicated example.

Explore More - resources for learning more about a particular topic.

If you want to learn more about Shiny apps, go to https://mastering-shiny.org to see an entire book on the topic.

Chapter 2

Getting Started and Installing Your Tools

One of the most intimidating parts of getting started with something new is the actual getting started part. Don't worry, we will walk you through this step-by step.

2.1 Goals for this Chapter

- Install R on your Computer
- Install RStudio on your Computer
- Install Git on your Computer
- Get Acquainted with the RStudio IDE

2.2 Website links needed for this Chapter

While in many chapters, we will list the R packages you need, in this chapter, you will be downloading and installing new software, so we will list the links here for your reference

- https://www.r-project.org
- https://rstudio.com/products/rstudio/download/
- https://git-scm.com/downloads

2.3 Pathway for this Chapter

This Chapter is part of the **TOOLS** pathway. Chapters in this pathway include

- Getting Started and Installing Your Tools
- Updating R, RStudio, and Your Packages
- Advanced Use of the RStudio IDE
- When You Don't Want to Update Packages (Using *renv*)
- Major R Updates (Where Are My Packages?)

2.4 Installing R on your Computer

2.5 Installing RStudio on your Computer

2.6 Installing Git on your Computer

2.7 Getting Acquainted with the RStudio IDE