Proctor Foundation Data Science Handbook

Contributors: Ben Arnold, Jade Benjamin-Chung, Kunal Mishra, Anna Nguyen, Nolan Pokpongk

2024-05-08

Contents

Welcome!

Welcome to the Francis I. Proctor Foundation at the University of California, San Francisco (https://proctor.ucsf.edu)!

This handbook summarizes some best practices for data science, drawing from our experience at the Francis I. Proctor Foundation and from that of our close colleagues in the Division of Epidemiology and Biostatistics at the University of California, Berkeley (where Prof. Ben Arnold worked for many years before joining Proctor).

We do not intend this handbook to be a comprehensive guide to data science. Instead, it focuses more on practical, "how-to" guidance for conducting data science within epidemiologic research studies. Where possible, we reference existing materials and guides.

Although many of the ideas of environment-independent, the examples draw from the R programming language. For an excellent overview of data science in R, see the book R for Data Science.

Much of the material in this handbook evolved from a version of Dr. Jade Benjamin-Chung's lab manual at the University of California, Berkeley. In addition to the Proctor team, many contributors include current and former students from UC Berkeley.

The last two chapters of the handbook cover our communication strategy and code of conduct for team members who work with Prof. Ben Arnold, who leads Proctor's Data Coordinating Center. They summarize key pieces of a functional data science team. Although the last two chapters might be of interest to a broader circle, *they are mostly relevant for people working directly with Ben.* Just because they are at the end does not make them less important.

It is a living document that we strive to update regularly. If you would like to contribute, please write Ben (ben.arnold@ucsf.edu) and/or submit a pull request.

The GitHub repository for this handbook is: https://github.com/proctor-ucsf/dcc-handbook

Chapter 1

Introduction: Work Flow and Reproducible Analyses

Contributors: Ben Arnold

This handbook collates a number of tips to help organize the workflow of epidemiologic data analyses. There are probably a dozen good ways to organize a workflow for reproducible research. This document includes recommendations that arise from our own team's experience through numerous field trials and observational data analyses. The recommendations will not work for everybody or for all applications. But, they work well for most of us most of the time, else we wouldn't put in the time to share them.

Start with two organizing concepts:

- **Workflow**. Defined here as the process required to draw scientific inference from data collected in the field or lab. I.e., the process by which we take data, and then process it, share it internally, analyze it, and communicate results to the scientific community.
- **Reproducible research**. A fundamental characteristic of the scientific method is that study findings can be reproduced beyond the original investigators. Data analyses that contribute to scientific research should be described and organized in a way that they could be reproduced by an independent person or research group. A data analysis that is not reproducible violates a core principle of the scientific method.

1.1 Workflow

Broadly speaking, a typical scientific data science work flow involves four steps to transform raw data (e.g., from the field) into summaries that communicate

results to the scientific community.

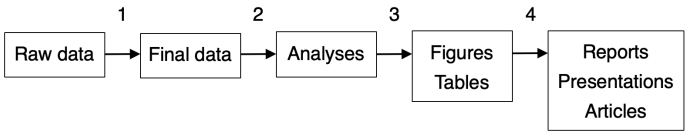


Figure 1.1: Overview of the four main steps in a typical data science workflow

When starting a new project, the work flow tends to evolve gradually and by iteration. Data cleaning, data processing, exploratory analyses, back to data cleaning, and so forth. If the work takes place in an unstructured environment with no system to organize files and work flow, it rapidly devolves into into a disorganized mess; analyses become difficult or impossible to replicate and they are anything but scientific. Projects with short deadlines (e.g., proposals, conference abstract submissions, article revisions) are particularly vulnerable to this type of organizational entropy. Putting together a directory and workflow plan from the start helps keep files organized and prevent disorder. Modifications are inevitable – as long as the system is organized, modifications are usually no problem.

Depending on the project, each step involves a different amount of work. Step 1 is by far the most time consuming, and often the most error-prone. We devote an entire chapter to it below (Data cleaning and processing)

1.2 Reproducibility

As a guiding directive, this process should be reproducible. If you are not familiar with the concept of reproducible research, start with this manifesto (Munafo et al. 2017). For a deeper dive, we highly recommend the recent book from Christensen, Freese, and Miguel (2019). Although it is framed around social science, the ideas apply generally.

Essential reading: Ten Simple Rules for Reproducible Computational Research Please read this excellent paper on computational reproducibility from Sandve et al 2013. It encapsulates many of the practices described herein (we elaborate more, and provide concrete examples).

1.3 Automation

We recommend that the workflow be as automated as possible using a programming language. Automating the workflow in a programming language, and essentially reducing it to text, is advantageous because it makes the process transparent, well documented, easily modified, and amenable to version control; these characteristics lend themselves to reproducible research.

At Proctor, we mostly use R. With the development of Rstudio, R Markdown and the tidyverse ecosystem (among others), the R language has evolved as much in the past few years as in all previous decades since its inception. This has made the conduct of automated, reproducible research considerably easier than it was 10 years ago.

If you have a step in your analysis workflow that involves point-and-click or copy/paste, then STOP, and ask yourself (and your team): *How can I automate this?*

Chapter 2

Workflows

Contributors: Ben Arnold

A data science work flow typically progresses through 4 steps that rarely evolve in a purely linear fashion, but in the end should flow in this direction:

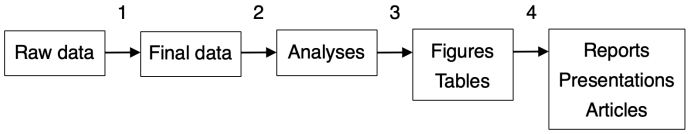


Figure 2.1: Overview of the four main steps in a typical data science workflow

Table 2.1: Workflow basics

Steps	Example activities	⇒ Inputs	⇒ Outputs
1	**Data cleaning and processing**		
.	make a plan for final datasets, fix data entry errors, create derived variables, plan for public replication files	untouched datasets	final datasets
2-3	**Analyses**		

Steps	Example activities	⇒ Inputs	⇒ Outputs
.	exploratory data analysis, study monitoring, summary statistics, statistical analyses, independent replication of analyses, make figures and tables	final datasets	saved results (.rds/.csv), tables (.html,.pdf), figures (.html/.png)
4 .	**Communication** results synthesis	saved results, figures, tables	monitoring reports, presentations, scientific articles

In many modern data science workflows, steps 2-4 can be accomplished in a single R notebook or Jupyter notebook: the statistical analysis, creation of figures and tables, and creation of reports.

However, it is still useful to think of the distinct stages in many cases. For example, a single statistical analysis might contribute to a DSMC report, a scientific conference presentation, and a scientific article. In this example, each piece of scientific communication would take the same input (stored analysis results as .csv/.rds) and then proceed along slightly different downstream workflows.

It would be more error prone to replicate the same statistical analysis in three parallel downstream work flows. This illustrates a key idea that holds more generally:

Key idea for workflows: Whenever possible, avoid repeating the same data processing or statistical a

Chapter 3

Directory Structure and Code Repositories

Contributors: Kunal Mishra, Jade Benjamin-Chung, Ben Arnold

The backbone of your project workflow is the file directory so it makes sense to spend time organizing the directory. Note that `directory` is the technical term for the system used to organize individual files. Most non-UNIX environments use a folder analogy, and directory and folder can be used interchangeably in a lot of cases. A well organized directory will make everything that follows much easier. Just like a well designed kitchen is essential to enjoy cooking (and avoid clutter), a well designed directory helps you enjoy working and stay organized in a complex project with literally thousands of related files. Just like a disorganized kitchen ("now where did I put that spatula?") a disorganized project directory creates confusion, lost time, stress, and mistakes.

Another huge advantage of maintaining a regular/predictable directory structure within a project and across projects is that it makes it more intuitive. When a directory is intuitive, it is easier to work collaboratively across a larger team; everybody can predict (at least approximately) where files should be.

Nested within your directory will be a `code repository`. Sometimes we find it useful to manage the code repository using version control, such as git/GitHub.

Other chapters will discuss coding practices, data management, and GitHub/version control that will build from the material here.

Carrying the kitchen analogy further: here, we are designing the kitchen. Then, we'll discuss approaches for how to cook in the kitchen that we designed/built.

3.1 Small and large projects

Our experience is that the overwhelming majority of projects come in two sizes: small and large. We recommend setting up your directory structure depending on how large you expect the project to be. Sometimes, small projects evolve into large projects, but only occasionally. A small project is something like a single data analysis with a single published article in mind. A large project is an epidemiologic field study, where there are multiple different types of data and different types of analyses (e.g., sample size calculations, survey data, biospecimens, substudies, etc.).

Small project: There is essentially one dataset and a single, coherent analysis. For example, a simulation study or a methodology study that will lead to a single article.

Large project: A field study that includes multiple activities, each of which generates data files. Multiple analyses are envisioned, leading to multiple scientific articles.

Large projects are more common and more complicated. Most of this chapter focuses on large project organization (small projects can be thought of as essentially one piece of a large project).

3.2 Directory Structure

In the example below, we follow a basic directory naming convention that makes working in UNIX and typing directory statements in programs much easier:

- **short names**
- **no spaces in the names** (not essential but a personal preference. Can use `_` or `-` instead)
- **lower case** (not essential, again, personal preferences vary!)

For example, Ben completed a study in Tamil Nadu, India during his dissertation to study the effect of improvements in water supply and sanitation on child health. Instead of naming the directory `Tamil Nadu` or `Tamil Nadu WASH Study`, he used `trichy` instead (a colloquial name for the city near the study, Tiruchirappalli), which was much easier to type in the terminal and in directory statements. A short name helps make directory references easier while programming.

3.2.1 First level: data and analyses

Start by dividing a project into major activities. In the example abpve, the project is named `mystudy`. There is a `data` subdirectory (more in a sec),

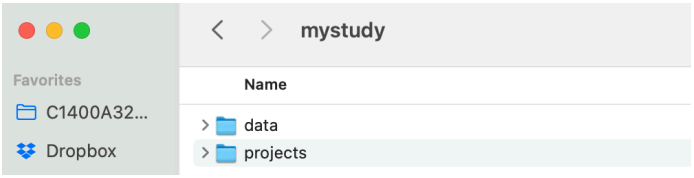
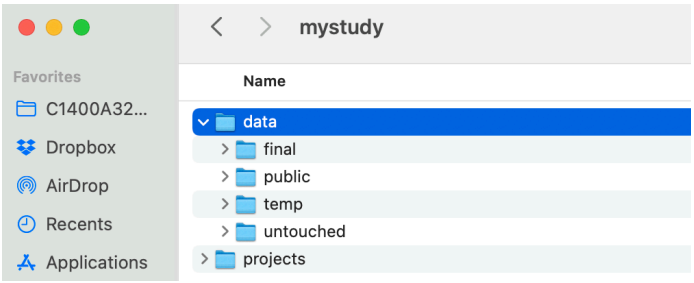


Figure 3.1: Example directory for ‘mystudy‘

and then three major activities, each corresponding to a separate analysis: `primary-analysis`,`secondary-analysis-1`, and `secondary-analysis-2`. In a real project, the names could be more informative, such as “trachoma-qpcr”. Also, a real project might also include many additional subdirectories related to administrative and logistics activities that do not relate to data science, such as irb, travel, contracts, budget, survey forms, etc.).

Dividing files into major activities helps keep things organized for really big projects. In a multi-site study, consider including a directory for each site before splitting files into major activities. Ideally, analyses will not span major activity subdirectories in a project folder, but sometimes you can’t predict/avoid that from happening.

3.2.2 Second level: data



Each project will include a `data` directory. We recommend organizing it into 3 parts: `untouched`, `temp`, and `final`. Often, it is useful to include a fourth subdirectory called `public` for sharing public versions of datasets.

The `untouched` directory includes all untouched datasets that are used for the study. Once saved in the directory never touch them; you will read them into the work flow, but **never, ever save over them**. If the study has repeated extracts from rolling data collection or electronic health records, consider subdirectories within `untouched` that are indexed by date.

The `temp` directory (optional, not essential) includes temporary files that you might generate during the data management process. This is mainly a space for experimentation. As a rule, never save anything in the temp directory that you cannot delete. Regularly delete files in the temp directory to save disk space.

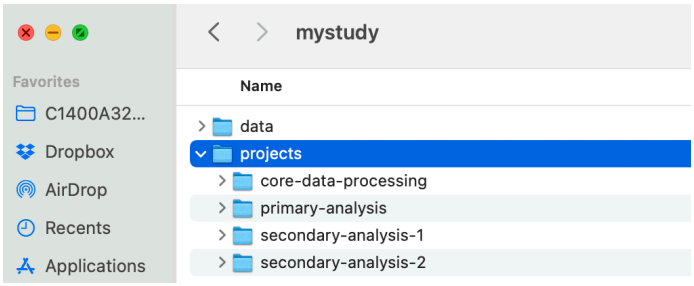
The `final` directory includes final datasets for the activity. Final datasets are de-identified and require no further processing; they are clean and ready for analysis. They should be accompanied by meta-data, which at minimum includes the data's provenance (i.e., how it was created) and what it includes (i.e., level of the data, plus variable coding/labels/descriptions). Clean/final datasets generated by one analysis might be reused in another.

NOTE: In very large projects, we occasionally need to further stratify the `data` directory into different subdirectories, according to data type. For example, in the AVENIR trial we had child mortality data measurements, and completely separate data from antimicrobial resistance (AMR) monitoring. These data were so myriad and different, that they needed to be curated in separate directories.

3.2.3 Second level: projects

We recommend maintaining a separate project subdirectory to hold each major analysis in a project. In this example, there are four workflows with generic names from the view of trial: `core-data-processing`, `primary-analysis`, `secondary-analysis-1`, `secondary-analysis-2`.

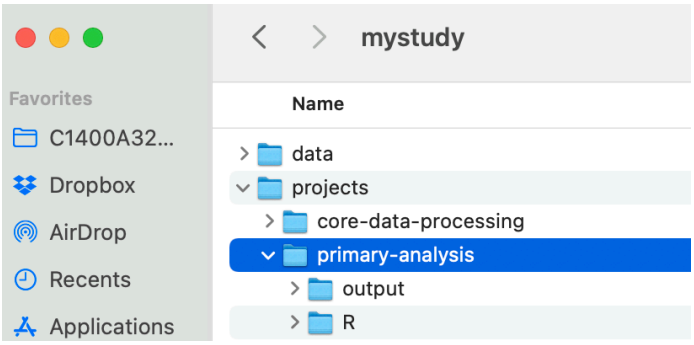
IMPORTANT: We recommend consolidating the core data processing for the study in its own project. **Core data processing** is data processing that creates a shared asset across multiple, downstream analyses. It includes the scripts for the data processing workflow that take data from `untouched` to `final` in the previous section.



Think of each analysis as the scope of all of the work for a single, published paper. We recommend dividing the analysis project into a space for computational notebooks / scripts (i.e., a `code repository`), and a second for their output. The reason for the split is to make it easier to use version control (should you choose) for the code. Version control like `git` and `GitHub` (see the Chapter on GitHub) works well for text files but isn't really designed for binary files such as images (.png), datasets (.rds), or PDF files (.pdf). It is certainly possible to use git with those file types, but since git makes a new copy of the file every time it is changed the git repo can get horribly bloated and takes up too much space on disk. Consolidating the output into a separate directory makes it more obvious that it isn't under version control. In this example, there are separate parts for code (`R`) and output (`output`). Output could include figures, tables, or saved

analysis results stored as data files (.rds or .csv). Another conventional name for the code repository is `src` as an alternative to `R` if you use other languages.

If a project creates unique datasets or relies on new data that is unique to the project and not a shared asset for the study (e.g., publicly available data), then the project itself might include a `data` subdirectory. HOWEVER, it should not include a direct copy of shared datasets stored in the `mystudy/data` repository – be sure to keep only one copy of each shared dataset in the project.

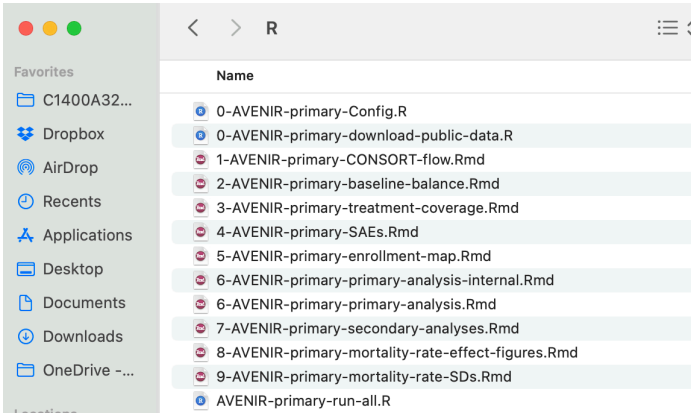


Interdependence between analyses: Sometimes a result from an analysis might be a cleaned dataset that could feed into future, distinct analyses. This is quite common, for example, in large trials where a set of baseline characteristics might be used in multiple separate papers for different endpoints, either for assessing balance of the trial population or subgroups, or used as adjustment covariates in additional analyses of the trial. In this case, the analysis-specific data processing should be *moved upstream* into the `projects/core-data-processing` workflow, and should generate a shared, final dataset in the `data/final` directory.

3.3 Code Repositories

Maintain a separate code repository for each major analysis activity (last section).

We recommend the following structure for a code repository. This is the actual repository for the AVENIR trial primary analysis.



Note that in this example, there is no actual data processing in the project. All of the data processing was consolidated upstream in a separate `core-data-processing` project for the trial. Sometimes, an analysis/project will include limited project-specific data processing that is not shared across the trial. If project-specific data processing is required, it would be one of the first scripts in the workflow. This helps ensure work conducted in step 1 of your workflow stays upstream from all analyses (see Chapter on workflows).

In this example, there is a script that downloads public datasets into a local repository instead. Also note that in this example, all of the scripts are `.Rmd` files. R scripts `.R` are equally useful.

You can glean some important takeaways from what you *do* see.

3.3.1 `.Rproj` files

An "R Project" can be created within RStudio by going to `File >> New Project`. Depending on where you are with your research, choose the most appropriate option. This will save preferences, working directories, and even the results of running code/data (though we recommend starting from scratch each time you open your project, in general). Then, ensure that whenever you are working on that specific research project, you open your created project to enable the full utility of `.Rproj` files. This also automatically sets the directory to the top level of the project.

3.3.2 Configuration ('config') File

This is the single most important file for your project. It will be responsible for a variety of common tasks, declare global variables, load functions, declare paths, and more. *Every other file in the project* will begin with `source("0-config")`, and its role is to reduce redundancy and create an abstraction layer that allows you to make changes in one place (`0-config.R`) rather than 5 different files. To this end, paths that will be referenced in multiple scripts (e.g., a

`clean_data_path`) can be declared in `0-config.R` and simply referred to by its variable name in scripts. If you ever want to change things, rename them, or even switch from a downsample to the full data, all you would then to need to do is modify the path in one place and the change will automatically update throughout your project. See the example config file for more details. The paths defined in the `0-config.R` file assume that users have opened the `.Rproj` file, which sets the directory to the top level of the project.

This GitHub repository that has replication files for this study includes an example of a streamlined `config.R` file, with all packages loaded and directory references defined.

3.3.3 Shared Functions File

If you write a custom function for an analysis and need to use it repeatedly across multiple analysis scripts, then it is better to consolidate it into a single shared functions script and source that file into the analysis scripts. The reason for this is that it enables you to edit the function in a single place and ensure that the changes are implemented across your entire workflow. In extreme cases, you might have so many shared functions that you need an entire subdirectory with separate scripts. This repository includes a very complex example (`0-project-functions`) of a large analysis published in a *Nature* series (one of three articles: https://www.nature.com/articles/s41586-023-06501-x)

3.3.4 Order Files and Subdirectories

This makes the jumble of alphabetized filenames much more coherent and places similar code and files next to one another. Although sometimes there is not a linear progression from 1 to 2 to 3, in general the structure helps reflect how data flows from start to finish.

If you take nothing else away from this guide, this is the single most helpful suggestion to make your workflow more coherent. Often the particular order of files will be in flux until an analysis is close to completion. At that time it is important to review file order and naming and reproduce everything prior to drafting a manuscript.

3.3.5 Use Bash scripts or R scripts to ensure reproducibility

Bash scripts are useful components of a reproducible workflow. If your workflow is in R scripts, bash (`.sh`) can be useful to run the entire workflow in order. See the UNIX Chapter for further details. If your workflow is mainly R markdown or a mix of `.R` and `.Rmd` files, sometimes it is easier to simply make a single R script to run the workflow. In the AVENIR example (above), that is what we used AVENIR-primary-run-all.R.

3.3.6 Alternative approach for code repos

Another approach for organizing your code repository is to name all of your scripts according to the final figure or table that they generate for a particular article. In our experience, this *only* works for small projects, with a single set of coherent analyses. Here, you might have an alternative structure such as:

```
.gitignore
primary-analysis.Rproj
0-config.R
0-shared-functions.R
0-primary-analysis-run-all.sh
1-dm /
    0-dm-run-all.sh
    1-format-enrollment-data.R
    2-format-adherence-data.R
    3-format-LAZ-measurements.R
Fig1-consort.Rmd
Fig2-adherence.Rmd
Fig3-1-laz-analysis.Rmd
Fig3-2-laz-make-figure.Rmd
```

There is still a need for a separate data management directory (e.g., `dm`) to ensure that workflow is upstream from the analysis (more below in chapter on UNIX), but then scripts are all together with clear labels. If a figure requires two stages to the analysis, then you can name them sequentially, such as `Fig3-1-laz-analysis.Rmd`, `Fig3-2-laz-make-figure.Rmd`. There is no way to divine how all of the analyses will neatly fit into files that correspond to separate figures. Instead, they will converge on these file names through the writing process, often through comsolidation or recombination.

One example of a small repo is here: https://github.com/ben-arnold/enterics-seroepi

Chapter 4

Coding Practices

Contributors: Kunal Mishra, Jade Benjamin-Chung, Ben Arnold

4.1 Organizing scripts

Just as your data "flows" through your project, data should flow naturally through a script. Very generally, you want to:

1. describe the work completed in the script in a comment header
2. source your configuration file (`0-config.R`)
3. load all of your data
4. do all your analysis/computation in order
5. save your results

Each section should be "chunked together" using comments, often with many chunks in a single section. See this file for a good example of how to cleanly organize a file in a way that follows this "flow" and functionally separate pieces of code that are doing different things. This is another example in a `.Rmd` format, where chunking is made even more obvious by the interleaving of markdown text and R code in the same notebook file.

4.2 Documenting your code

4.2.1 File headers

Every file in a project should have a header that allows it to be interpreted on its own. It should include the name of the project and a short description for what this file (among the many in your project) does specifically. You may optionally wish to include the inputs and outputs of the script as well, though the next section makes this significantly less necessary. It can be very helpful to include

your name and email address as well so others can identify who wrote the code. This is unnecessary if you are using version control (`git/GitHub`) becuase that information will be tracked by commits.

```
#------------------------------------------------------------------------------
# @Organization - Example Organization
# @Project - Example Project
# @Author - Your name, and possibly email address (if appropriate)
# @Description - This file is responsible for [...]
#------------------------------------------------------------------------------
```

Consider using RStudio's code folding feature to collapse and expand different sections of your code. Any comment line with at least four trailing dashes (-), equal signs (=), or pound signs (#) automatically creates a code section. Delimiters for chunks are a personal preference and all work equally well. For example:

```
# Section 1 ------------------
```

works equally well as

```
##############################
# Section 1 ##################
##############################
```

4.2.2 Comments in the body of your script

Commenting your code is an important part of reproducibility and helps document your code for the future. When things change or break, you'll be thankful for comments. When you revisit code you wrote two years earlier, you'll be thankful for comments. There's no need to comment excessively or unnecessarily, but a comment describing what a large or complex chunk of code does is always helpful. See this file for an example of how to comment your code and notice that comments are always in the form of:

```
# This is a comment -- first letter is capitalized and spaced
away from the pound sign
```

4.2.3 Function documentation

Every function you write must include a header to document its purpose, inputs, and outputs. For any reproducible workflows, they are essential, because R is dynamically typed. This means, you can pass a `string` into an argument that is meant to be a `data.table`, or a `list` into an argument meant for a `tibble`. It is the responsibility of a function's author to document what each argument is meant to do and its basic type. This is an example for documenting a function (inspired by JavaDocs, R's Plumber API docs, and Roxygen2):

```
#------------------------------------------------------------------------------
```

```
# Documentation: calc_fluseas_mean
# Usage: calc_fluseas_mean(data, yname)
# @description: Make a dataframe with rows for flu season and site
#               and the number of patients with an outcome, the total patients,
#               and the percent of patients with the outcome

# Arguments/Options:
# @param data: a data frame with variables flu_season, site, studyID, and yname
# @param yname: a string for the outcome name
# @param silent: a boolean specifying whether the function shouldn't output anything to the consc

# @return: the dataframe as described above
# @output: prints the data frame described above if silent is not True
#-------------------------------------------------------------------------------

calc_fluseas_mean = function(data, yname, silent = TRUE) {
 ### function code here

}
```

The header tells you what the function does, its various inputs, and how you might go about using the function to do what you want. Also notice that all optional arguments (i.e. ones with pre-specified defaults) follow arguments that require user input.

- **Note**: As someone trying to call a function, it is possible to access a function's documentation (and internal code) by `CMD-Left-Click`ing the function's name in RStudio
- **Note**: Depending on how important your function is, the complexity of your function code, and the complexity of different types of data in your project, you can also add "type-checking" to your function with the `assertthat::assert_that()` function. You can, for example, `assert_that(is.data.frame(statistical_input))`, which will ensure that collaborators or reviewers of your project attempting to use your function are using it in the way that it is intended by calling it with (at the minimum) the correct type of arguments. You can extend this to ensure that certain assumptions regarding the inputs are fulfilled as well (i.e. that `time_column`, `location_column`, `value_column`, and `population_column` all exist within the `statistical_input` tibble).

4.3 Object naming

Generally we recommend using nouns for objects and verbs for functions. This is because functions are performing actions, while objects are not.

Try to make your variable names both more expressive and more explicit.

Being a bit more verbose is useful and easy in the age of autocompletion! For example, instead of naming a variable `vaxcov_1718`, try naming it `vaccination_coverage_2017_18`. Similarly, `flu_res` could be named `absentee_flu_residuals`, making your code more readable and explicit.

- For more help, check out Be Expressive: How to Give Your Variables Better Names

We recommend you use **Snake_Case**.

- Base R allows `.` in variable names and functions (such as `read.csv()`), but this goes against best practices for variable naming in many other coding languages. For consistency's sake, `snake_case` has been adopted across languages, and modern packages and functions typically use it (i.e. `readr::read_csv()`). As a very general rule of thumb, if a package you're using doesn't use `snake_case`, there may be an updated version or more modern package that *does*, bringing with it the variety of performance improvements and bug fixes inherent in more mature and modern software.

- **Note**: you may also see `camelCase` throughout the R code you come across. This is *okay* but not ideal – try to stay consistent across all your code with `snake_case`.

- **Note**: there is nothing inherently wrong with using `.` in variable names, just that it goes against style best practices that are cropping up in data science, so its worth getting rid of these bad habits now.

4.4 Function calls

In a function call, use "named arguments" and separate arguments by to make your code more readable.

Here's an example of what not to do when calling the function a function `calc_fluseas_mean` (defined above):

```
mean_Y = calc_fluseas_mean(flu_data, "maari_yn", FALSE)
```

And here it is again using the best practices we've outlined:

```
mean_Y = calc_fluseas_mean(
  data = flu_data,
  yname = "maari_yn",
  silent = FALSE
)
```

4.5 The here package

The `here` package is one great R package that helps multiple collaborators deal with the mess that is working directories within an R project structure. Let's say we have an R project at the path `/home/oski/Some-R-Project`. My collaborator might clone the repository and work with it at some other path, such as `/home/bear/R-Code/Some-R-Project`. Dealing with working directories and paths explicitly can be a very large pain, and as you might imagine, setting up a Config with paths requires those paths to flexibly work for all contributors to a project. This is where the `here` package comes in and this a great vignette describing it.

For more motivation on why you should use the `here` and R projects (`.Rproj`), read this excellent blog post from Tidyverse.

4.6 Tidyverse

Throughout this document there have been references to the Tidyverse, but this section is to explicitly show you how to transform your Base R tendencies to Tidyverse (or Data.Table, Tidyverse's performance-optimized competitor). For most of our work that does not utilize very large datasets, we recommend that you code in Tidyverse rather than Base R. Tidyverse is quickly becoming the gold standard in R data analysis and modern data science packages and code should use Tidyverse style and packages unless there's a significant reason not to (i.e. big data pipelines that would benefit from Data.Table's performance optimizations).

The package author has published a great textbook on R for Data Science, which leans heavily on many Tidyverse packages and may be worth checking out.

The following list is not exhaustive, but is a compact overview to begin to translate Base R into something better:

Base R	Better Style, Performance, and Utility
—	—
`read.csv()`	`readr::read_csv()` or `data.table::fread()`
`write.csv()`	`readr::write_csv()` or `data.table::fwrite()`
`readRDS`	`readr::read_rds()`
`saveRDS()`	`readr::write_rds()`
—	—
`data.frame()`	`tibble::tibble()` or `data.table::data.table()`
`rbind()`	`dplyr::bind_rows()`
`cbind()`	`dplyr::bind_cols()`

Base R	Better Style, Performance, and Utility
`df$some_column`	`df %>% dplyr::pull(some_column)`
`df$some_column = ...`	`df %>% dplyr::mutate(some_column = ...)`
`df[get_rows_condition,]`	`df %>% dplyr::filter(get_rows_condition)`
`df[,c(col1, col2)]`	`df %>% dplyr::select(col1, col2)`
`merge(df1, df2, by = ..., all.x = ..., all.y = ...)`	`df1 %>% dplyr::left_join(df2, by = ...)` or `dplyr::full_join` or `dplyr::inner_join` or `dplyr::right_join`
—	—
`str()`	`dplyr::glimpse()`
`grep(pattern, x)`	`stringr::str_which(string, pattern)`
`gsub(pattern, replacement, x)`	`stringr::str_replace(string, pattern, replacement)`
`ifelse(test_expression, yes, no)`	`if_else(condition, true, false)`
Nested: `ifelse(test_expression1, yes1, ifelse(test_expression2, yes2, ifelse(test_expression3, yes3, no)))`	`case_when(test_expression1 ~ yes1, test_expression2 ~ yes2, test_expression3 ~ yes3, TRUE ~ no)`
`proc.time()`	`tictoc::tic()` and `tictoc::toc()`
`stopifnot()`	`assertthat::assert_that()` or `assertthat::see_if()` or `assertthat::validate_that()`

For a more extensive set of syntactical translations to Tidyverse, you can check out this document.

Working with Tidyverse within functions can be somewhat of a pain due to non-standard evaluation (NSE) semantics. If you're an avid function writer, we'd recommend checking out the following resources:

- Tidy Eval in 5 Minutes (video)
- Tidy Evaluation (e-book)
- Data Frame Columns as Arguments to Dplyr Functions (blog)
- Standard Evaluation for *_join (stackoverflow)
- Programming with dplyr (package vignette)

4.7 Coding with R and Python

If you're using both R and Python, you may wish to check out the Feather package for exchanging data between the two languages extremely quickly.

Chapter 5

Coding Style

Contributors: Kunal Mishra, Jade Benjamin-Chung, and Ben Arnold

5.1 Line breaks

- For `ggplot` calls and `dplyr` pipelines, do not crowd single lines. Here are some nontrivial examples of "beautiful" pipelines, where beauty is defined by coherence:

```
# Example 1
school_names <- list(
  OUSD_school_names = absentee_all %>%
    filter(dist.n == 1) %>%
    pull(school) %>%
    unique() %>%
    sort(),

  WCCSD_school_names <- absentee_all %>%
    filter(dist.n == 0) %>%
    pull(school) %>%
    unique() %>%
    sort()
)
```

```
# Example 2
absentee_all <- fread(file = raw_data_path) %>%
  mutate(program = case_when(schoolyr %in% pre_program_schoolyrs ~ 0,
                             schoolyr %in% program_schoolyrs ~ 1)) %>%
  mutate(period = case_when(schoolyr %in% pre_program_schoolyrs ~ 0,
                            schoolyr %in% LAIV_schoolyrs ~ 1,
                            schoolyr %in% IIV_schoolyrs ~ 2)) %>%
```

```
  filter(schoolyr != "2017-18")
```

And of a complex `ggplot` call:

```
# Example 3
ggplot(data=data,
       mapping=aes_string(x="year", y="rd", group=group)) +

  geom_point(mapping=aes_string(col=group, shape=group),
             position=position_dodge(width=0.2),
             size=2.5) +

  geom_errorbar(mapping=aes_string(ymin="lb", ymax="ub", col=group),
                position=position_dodge(width=0.2),
                width=0.2) +

  geom_point(position=position_dodge(width=0.2),
             size=2.5) +

  geom_errorbar(mapping=aes(ymin=lb, ymax=ub),
                position=position_dodge(width=0.2),
                width=0.1) +

  scale_y_continuous(limits=limits,
                     breaks=breaks,
                     labels=breaks) +

  scale_color_manual(std_legend_title,values=cols,labels=legend_label) +
  scale_shape_manual(std_legend_title,values=shapes, labels=legend_label) +
  geom_hline(yintercept=0, linetype="dashed") +
  xlab("Program year") +
  ylab(yaxis_lab) +
  theme_complete_bw() +
  theme(strip.text.x = element_text(size = 14),
        axis.text.x = element_text(size = 12)) +
  ggtitle(title)
```

Imagine (or perhaps mournfully recall) the mess that can occur when you don't strictly style a complicated `ggplot` call. Trying to fix bugs and ensure your code is working can be a nightmare. Now imagine trying to do it with the same code 6 months after you've written it. Invest the time now and reap the rewards as the code practically explains itself, line by line.

5.2 Automated Tools for Style and Project Workflow

5.2.1 Styling

1. **Code Autoformatting** - RStudio includes a fantastic built-in utility (keyboard shortcut: `CMD-Shift-A`) for autoformatting highlighted chunks of code to fit many of the best practices listed here. It generally makes code more readable and fixes a lot of the small things you may not feel like fixing yourself. Try it out as a "first pass" on some code of yours that *doesn't* follow many of these best practices!

2. **Assignment Aligner** - A cool R package allows you to very powerfully format large chunks of assignment code to be much cleaner and much more readable. Follow the linked instructions and create a keyboard shortcut of your choosing (recommendation: `CMD-Shift-Z`). Here is an example of how assignment aligning can dramatically improve code readability:

```
# Before
OUSD_not_found_aliases <- list(
  "Brookfield Village Elementary" = str_subset(string = OUSD_school_shapes$schnam, pattern = "Bro
  "Carl Munck Elementary" = str_subset(string = OUSD_school_shapes$schnam, pattern = "Munck"),
  "Community United Elementary School" = str_subset(string = OUSD_school_shapes$schnam, pattern =
  "East Oakland PRIDE Elementary" = str_subset(string = OUSD_school_shapes$schnam, pattern = "Eas
  "EnCompass Academy" = str_subset(string = OUSD_school_shapes$schnam, pattern = "EnCompass"),
  "Global Family School" = str_subset(string = OUSD_school_shapes$schnam, pattern = "Global"),
  "International Community School" = str_subset(string = OUSD_school_shapes$schnam, pattern = "In
  "Madison Park Lower Campus" = "Madison Park Academy TK-5",
  "Manzanita Community School" = str_subset(string = OUSD_school_shapes$schnam, pattern = "Manzan
  "Martin Luther King Jr Elementary" = str_subset(string = OUSD_school_shapes$schnam, pattern = "
  "PLACE @ Prescott" = "Preparatory Literary Academy of Cultural Excellence",
  "RISE Community School" = str_subset(string = OUSD_school_shapes$schnam, pattern = "Rise Commun
)

# After
OUSD_not_found_aliases <- list(
  "Brookfield Village Elementary"      = str_subset(string = OUSD_school_shapes$schnam, pattern =
  "Carl Munck Elementary"              = str_subset(string = OUSD_school_shapes$schnam, pattern =
  "Community United Elementary School" = str_subset(string = OUSD_school_shapes$schnam, pattern =
  "East Oakland PRIDE Elementary"      = str_subset(string = OUSD_school_shapes$schnam, pattern =
  "EnCompass Academy"                  = str_subset(string = OUSD_school_shapes$schnam, pattern =
  "Global Family School"               = str_subset(string = OUSD_school_shapes$schnam, pattern =
  "International Community School"     = str_subset(string = OUSD_school_shapes$schnam, pattern =
  "Madison Park Lower Campus"          = "Madison Park Academy TK-5",
  "Manzanita Community School"         = str_subset(string = OUSD_school_shapes$schnam, pattern =
  "Martin Luther King Jr Elementary"   = str_subset(string = OUSD_school_shapes$schnam, pattern =
  "PLACE @ Prescott"                   = "Preparatory Literary Academy of Cultural Excellence",
```

```
  "RISE Community School"              = str_subset(string = OUSD_school_shapes$schnam
)
```

3. **StyleR** - Another cool R package from the Tidyverse that can be powerful and used as a first pass on entire projects that need refactoring. The most useful function of the package is the `style_dir` function, which will style all files within a given directory. See the function's documentation and the vignette linked above for more details.
    - **Note**: The default Tidyverse styler is subtly different from some of the things we've advocated for in this document. Most notably we differ with regards to the number of spaces before/after "tokens" (i.e. Assignment Aligner add spaces before `=` signs to align them properly). For this reason, we'd recommend the following: `style_dir(path = ..., scope = "line_breaks", strict = FALSE)`. You can also customize StyleR even more if you're really hardcore.
    - **Note**: As is mentioned in the package vignette linked above, StyleR modifies things *in-place*, meaning it overwrites your existing code and replaces it with the updated, properly styled code. This makes it a good fit on projects *with version control*, but if you don't have backups or a good way to revert back to the intial code, I wouldn't recommend going this route.

Chapter 6

Data Wrangling

Contributors: Kunal Mishra, Jade Benjamin-Chung, Ben Arnold

6.1 Overview

Processing data from raw inputs to final data for analysis is the foundation of the data science workflow. "Management" is a bad word in both academia and hip parts of the private sector, so what used to be called "data management" is now called "data wrangling" or "data munging".

By any name, it is by far the most tedious and time consuming step in the data analysis workflow. Data wrangling is also the most error prone. Errors are likely because data processing usually requires hundreds or thousands of lines of code. Generating final datasets often requires programmers to merge and append multiple different raw datasets; often, the process requires them to re-shape the data from wide format to long format.

Read the chapters on Data Wrangling in *R for Data Science*. That book covers the foundational skills and know-how.

Here, we provide some additional recommendations for best practices related to data wrangling and highlight a few topics not covered in the excellent R for Data Science book.

6.2 Cardinal rule

Cardinal rule of data wrangling: *NEVER alter untouched datasets*

Untouched datasets include data output from data collection software, distributed by the data team, outside collaborators, or downloaded from a data repository. They should be left untouched. All data cleaning and processing activities should read in untouched data and output either temporary files or final datasets. Never, ever overwrite an untouched dataset.

6.3 Data input/output (I/O)

Read the Data Import section of *R for Data Science*.

6.3.1 Excel files

Often collaborators share data as Excel files (`.xlsx`). In this case, we recommend making a copy of the file in the untouched repository and converting it into a `.csv` file (one per tab) to input into the data processing workflow.

This step should mark the end of your use of Excel in your data analysis, aside from inspecting `.csv` files. Excel is extremely error prone because it does not have good fidelity with variable types (e.g., dates, leading zeros), and has no real programmatic interface (VBA?). Avoid it. We can find no compelling case for its use in data science.

6.3.2 `.RDS` vs `.RData` Files

One of the most common ways to load and save data in Base R is with the `load()` and `save()` functions to serialize multiple objects in a single `.RData` file. The biggest problems with this practice include an inability to control the names of things getting loaded in, the inherent confusion this creates in understanding older code, and the inability to load individual elements of a saved file. For this, we recommend using the RDS format to save R objects using `saveRDS()` and its complement `readRDS()`.

- **Note**: if you have many related R objects you would have otherwise saved all together using the `save` function, the functional equivalent with `RDS` would be to create a (named) list containing each of these objects, and saving it.
- **Note**: there is an important caveat for `.rds` files: they are not automatically backward compatible across different versions of R! So, while they are very useful in general, beware. See, for example, this thread on StackExchange. `.csv` files embed slightly less information (typically), but are more stable across different versions of R.

6.3.3 `.CSV` Files

Once again, the `readr` package as part of the Tidvyerse is great, with a much faster `read_csv()` than Base R's `read.csv()`. For massive CSVs (> 5 GB),

you'll find `data.table::fread()` to be the fastest CSV reader in any data science language out there. For writing CSVs, `readr::write_csv()` and `data.table::fwrite()` outclass Base R's `write.csv()` by a significant margin as well.

6.4 Documenting datasets

Datasets need to have metadata (documentation) associated with them to help people understand them. Well documented datasets save an enormous amount of time because it helps avoid lots of back-and-forth with new people orienting themselves with the data. This applies to both private and public data used in your work flow.

Each raw and final dataset should include a codebook. Sometimes survey instruments or electronic data capture schematics can stand-in for raw dataset codebooks, as long as that information is stored alongside the raw data!

The file asembo_analysis_codebook.txt provides one example of what a codebook for a simple, final dataset could contain.

For complex studies with multiple, relational data files, it is exceptionally helpful to also include a README overview in plain text or markdown that explains the relationships between the datasets. Here is an example from the WASH Benefits Bangladesh trial primary outcomes analysis: README-WBB-primary-outcomes-datasets.md.

6.5 Mapping data from untouched -> final

An important step for the data processing is to come up with a plan for which final datasets you want to create.

Best pracice is to scope out the flow of data from the original forms/modules/tables at the time of data capture to the final datasets planned for the analysis, and ensure that there is a `Key` linking each table needed to create the final analysis datasets.

These are probably the two most common models for the end product:

Single Massive Dataset	A single, large, final dataset for the whole project with potentially multiple levels of data (e.g., household, individual) and hundreds or thousands of variables.
Multiple Relational datasets	Multiple relational datasets, each tailored to a specific type of data collected in the study. Smaller, relational datasets can be recombined as needed for analyses.

We strongly recommend using the second model, relational data tables, in almost

every case. The chapter on Relational Data in R for Data Science explains details.

> Rationale: If you create a single massive dataset that includes information at multiple levels, then it can create challenges. First, the datasets are unwieldy, often consisting of hundreds or thousands of variables; this makes it difficult to find variables, difficult to view the data through a software browser, and easy to forget about variables. *But there are bigger problems.* Datasets with an exceptionally large number of variables or multiple levels of data in the same file make it difficult to visually detect pathological problems in the data (e.g., unexpected missing values). For studies that include data from multiple levels (e.g., communities, households, individuals) the inclusion of multiple types of data in the same file makes analyses error prone because it requires the programmer to keep track of when variables include duplicated observations. For example, the inclusion of household characteristics in an individual-level data file means that if there are multiple individuals per household, that simple means of household characteristics cannot be calculated without first restricting the file to unique observations at the household level.

6.6 Relational data

Refer to Relational Data in *R for Data Science* for a good example of what a good relational database can look like.

Important: For a set of relational tables to work, you must ensure that each table relates to another with a unique `Key`. For example, a child-level dataset must also include IDs for household- or cluster to be merged to household or cluster level data.

Practical advice on `Key` variables / IDs used to link datasets. Numeric IDs can occasionally cause problems with merges due to machine rounding errors so character values can sometimes be a good alternative. Whatever format you use, ensure that they do not in some way encode identifying information (e.g., using a social security number or medical record number would be a poor choice of ID for a final dataset).

6.7 Be careful with joins / merges

Experience shows that merging datasets is one of the most error prone steps in data wrangling. Merges can be particularly tricky for less experienced programmers. Errors are likely early in the data processing work flow, and less likely when merging final datasets with clean key variables (IDs) and clear relationships between datasets. Carefully check each merge to ensure it is working as you expect.

ESSENTIAL READING: Join Problems in *R for Data Science*

In the Stata software there is a really nice feature with `merge` that allows you to specify the type of merge/join and then examine the diagnostics of that merge.

Although R does not have a similarly nice feature at this time, the different `join` functions in the `dplyr` package provide all of the tools you need to correctly join your data.

In our experience, the two most common types of joins in our work are mutating joins and filtering joints. (from the *R for Data Science* book):

> **Mutating joins**: add new variables to one data frame from matching observations in another.
>
> **Filtering joins**: filter observations from one data frame based on whether or not they match an observation in the other table.

Familiarize yourself with these types of operations. The Relational data chapter of *R for Data Science* includes details.

6.8 Reshaping data

The `dplyr` package introduced the `pivot_wider` and `pivot_longer` functions in 2019. They make reshaping data dramatically easier than any previous functions, including `reshape()` (base R) or `gather`/`spread` (Tidyverse). This excellent article includes details:

https://tidyr.tidyverse.org/dev/articles/pivot.html

6.9 Data cleaning

Data cleaning is typically an iterative process that you should conduct on a variable-by-variable basis. Occasionally it is useful to clean an entire questionnaire module at the same time since variables in the same module can be related to one another, and since it can help to proceed through the cleaning process in well-defined chunks. I find it helpful to have a copy of the survey questionnaire (if relevant) handy while cleaning specific variables. This enables you to check skip patterns and codes in the survey against the dataset.

We don't have a lot of general advice for data cleaning except the following. First, there is something to be said for "trolling" (in the fishing sense) or visually scrolling through parts of the dataset to see what the data look like and to make sure you have a feel for how information is stored in different variables (e.g., are missing values coded or simply recored as "`.`" or `NA`). Use the `View()` utility in `RStudio` to look at your data often during the cleaning process. Second, it can be helpful to consolidate all data cleaning into a single program for each dataset. The data cleaning program corrects entry errors, labels variables,

formats variables, and outputs either a temporary or final dataset (but does not, ever, overwrite the untouched data). As noted in the Workflows chapter, *key data processing should be done once, at the earliest possible place in the workflow.*

Chapter 7

Making Data Public

Contributors: Fanice Nyatigo, Ben Arnold

7.1 Overview

Warning! ***NEVER*** *push a dataset into the public domain (e.g., GitHub, OSF) without first checking with Ben to ensure that it is appropriately de-identified and we have approval from the sponsor and/or human subjects review board to do so. For example, we will need to re-code participant IDs (even if they contain no identifying information) before making data public to completely break the link between IDs and identifiable information stored on our servers.*

If you are releasing data into the public domain, then consider making available *at minimum* a `.csv` file and a codebook of the same name (note: you should have a codebook for internal data as well). We often also make available `.rds` files as well. For example, your `mystudy/data/public` directory could include three files for a single dataset, two with the actual data in `.rds` and `.csv` formats, and a third that describes their contents:

```
analysis_data_public.csv
analysis_data_public.rds
analysis_data_public_codebook.txt
```

In general, datasets are usually too big to save on GitHub, but occasionally they are small. Here is an example of where we actually pushed the data directly to GitHub: https://github.com/ben-arnold/enterics-seroepi/tree/master/data .

If the data are bigger, then maintaining them under version control in your git repository can be unweildy. Instead, we recommend using another stable repository that has version control, such as the Open Science Framework (osf.io). For example, all of the data from the WASH Benefits trials (led by investigators at Berkeley, icddr,b, IPA-Kenya and others) are all stored through data components nested within in OSF projects: https://osf.io/tprw2/. Another good option is Dryad (datadryad.org), which normally costs a nominal fee to archive data but is free to investigators in the University of California system.

As of this time, we recommend cross-linking public files in GitHub (scripts/notebooks only), OSF, and Dryad.

Below are the main steps to making data public, after finalizing the analysis datasets and scripts:
1. Remove Protected Health Information (PHI)
2. Create public IDs or join already created public IDs to the data
3. Create an OSF repository and/or Dryad repository
4. Edit analysis scripts to run using the public datasets and test (optional)
5. Create a public github page for analysis scripts and link to OSF and/or Dryad/Zenodo
6. Go live

7.2 Removing PHI

Once the data is finalized for analysis, the first step is to strip it of Protected Health Information (PHI), or any other data that could be used to link back to specific participants, such as names, birth dates, or GPS coordinates at the village/neighborhood level or below. PHI includes, but is not limited to:

7.2.1 Personal information

These are identifiers that directly point to specific individuals, such as:
- Names, addresses, photographs, date of birth
- A combination of age, sex, and geographic location (below population 20,000) is considered identifiable

7.2.2 Dates

Any specific dates (e.g., study visit dates, birth dates, treatment dates) are usually problematic.
- If a dataset requires high resolution temporal information, coarsen visit or measurement dates to be two variables: year and week of the year (1-52).
- If a dataset requires age, provide that information without a birth date (typically month resolution is sufficient)

> ***Caution!*** *If making changes to the format of dates or ages, make sure your analysis code runs on these modified versions of the data (step 3)!*

7.2.3 Geographic information

Do not include GPS coordinates (longitude, latitude) except in special circumstances where they have been obfuscated/shifted. Reach out to your PI and the study's biostatistician before doing this because it can be complicated.

Do not include place names or codes (e.g., US Zip Codes) if the place contains <20,000 people. For villages or neighborhoods, code them with uninformative IDs. For sub-districts or districts, names are fine.

If an analysis requires GPS locations (e.g., to make a map), then typically we include a disclaimer in the article's data availability statement that explains we cannot make GPS locations public to protect participant confidentiality. As a middle ground, we typically make our *code* public that runs on the geo-located data for transparency, even if independent researchers can't actually run that code (although please be careful to ensure the code itself does not in any way include geographic identifiers).

For more examples of what constitutes PHI, please refer to this link: https://cphs.berkeley.edu/hipaa/hipaa18.html

7.3 Create public IDs

7.3.1 Rationale

The UCSF IRB requires that public datasets not include the original study IDs to identify participants or other units in the study (such as village IDs). The reason is that those IDs are linked in our private datasets to PHI. By creating a new set of public IDs, the public dataset is one step further removed from the potential to link to PHI.

7.3.2 A single set of public IDs for each study

For each study, it is ideal to create a single set of public IDs whenever possible. We could create a new set of public IDs for every public dataset, but the downside is that independent researchers could no longer link data that might be related. By creating a single set of public IDs associated with each internal study ID, public files retain the link.

Maintaining a single set of public IDs requires a shared "bridge" dataset, that includes a row for each study ID and has the associated public ID. For studies

with multiple levels of ID, we would typically have separate bridge datasets for each type of ID (e.g,. cluster ID, participant ID, etc.)

Create a public ID that can be used to uniquely identify participants and that can internally be linked to the original study IDs. We recommend creating a subdirectory in the study's shared data directory to store the public IDs. The shared location enables multiple projects to use the same IDs. Create the IDs using a script that reads in the study IDs, creates a unique (uninformative) public ID for the study IDs, and then saves the bridge dataset. The script should be saved in the same directory as the public ID files.

Caution! *Note that small differences may arise if the new public IDs do not necessarily order participants in the same way as the internal IDs. The small differences are all in estimates that rely on resampling, such as Bootstrap CIs, permutation P-values, and TMLE, as the resampling process may lead tp slightly different re-samples. The key here, to ensure the results are consistent irrespective of the dataset used, is simply to not assign public IDs randomly. Use* `rank()` *on the internal ID instead of* `row_number()` *to ensure that the order is always the same.*

7.3.3 Example scripts

We have created a self-contained and reproducible example that you can run and replicate when making data public for your projects. It contains the following files and folders:

1. `data/final/`- folder containing the projects final data in both csv and rds formats

2. `code/DEMO_generate_public_IDs.R`- creates randomly generated public IDs that can be matched to the trial's assigned patient IDs.

3. `data/make_public/DEMO_internal_to_publicID.csv`- the output from step #2, a bridge dataset with two variables- the new public ID and the patient's assigned ID.

4. `code/DEMO_create_public_datasets.R`- joins the public IDs to the trial's full dataset, and strips it of the assigned patient ID.
5. `data/public/`- folder containing the output from step #3- de-identified public dataset, in csv and rds formats, with uniquely identifying public IDs that cannot be easily linked back to the patient's ID.

The example workflow is accessible via GitHub: https://github.com/proctor-ucsf/dcc-handbook/tree/master/templates/making-data-public

7.4 Create a data repository

First, ensure that you create a codebook and metadata file for each public dataset See the DCC guide on Documenting datasets. Use the same name as the datasets, but with "-codebook.txt" / "-codebook.html" / "-codebook.csv" at the end (depending on the file format for the codebook). One nice option is the R codebook package, which also generates JSON output that is machine-readable.

7.4.1 Steps for creating an Open Science Framework (OSF) repository:

1. Create a new OSF project per these instructions: https://help.osf.io/article/252-create-a-project
2. Create a data component and upload the datasets in .csv and .rds format along with the codebooks. The primary format for public dissemination is .csv but we make the .rds files available too as auxiliary files for convenience.
3. Create a notebook component and upload the final .html files (which will not be on github… but see optional item below)
4. On the OSF landing Wiki, provide some context. Here is a recent example: https://osf.io/954bt/
5. Create a Digital Object Identifier (DOI) for the repository. A DOI is a unique identifier that provides a persistent link to content, such as a dataset in this case. Learn more about DOIs
6. Optional: Complete the software checklist and system requirement guide for the analysis to guide others. Include it on the GitHub README for the project: https://github.com/proctor-ucsf/mordor-antibody

7.4.2 Steps for creating a Dryad data repository:

TBD https://datadryad.org/stash/submission_process

7.5 Edit and test analysis scripts

Make minor changes to the analysis scripts so that they run on public data. If using version control in GitHub, the most straight-forward way is to create a branch from the main git branch that reads in the public files, and then renames the new public ID variable, e.g., "id_public" to the internally recognized ID variable name, e.g. "recordID", when reading in the public data. Re-run all the analysis scripts to ensure that they still work with the public version of the dataset.

7.6 Create a public GitHub page for public scripts

At minimum, we should include all of the scripts required to run the analyses .

See examples:
- ACTION - https://github.com/proctor-ucsf/ACTION-public
- NAITRE - https://github.com/proctor-ucsf/NAITRE-primary

> ***Caution!*** *Read through the scripts carefully to ensure there is no PHI in the code itself*

Once a public GitHub page exists, you can create a new component on an OSF project (step 3, above) and link it to the public version of the GitHub repo.

7.7 Go live

On GitHub, it is useful to create an official "release" version to freeze the repository, where you can have "associated files" with each version. Include the .html notebook output as additional files — since they aren't tracked in GitHub, it does provide a way of freezing / saving the HTML output for us and others. OSF examples of a few Proctor trials:
- ACTION - https://osf.io/ca3pe/
- NAITRE - https://osf.io/ujeyb/
- MORDOR Niger antibody study - https://osf.io/dgsq3/

Further reading on end-to-end data management: How to Store and Manage Your Data - PLOS

Chapter 8

Working with Big Data

Contributors: Eric Kim, Kunal Mishra and Jade Benjamin-Chung

8.1 Basics

A pitfall of working in R is that all objects are stored in memory - this makes it very difficult to work with datasets that are larger than 1-2 Gb for most standard computers. Here, we'll explore some alternatives to working with big data.

The Berkeley Statistical Computing Facility also has many good training resources.

8.2 Using downsampled data

In studies with very large datasets, we save "downsampled" data that usually includes a 1% random sample stratified by any important variables, such as year or household id. This allows us to efficiently write and test our code without having to load in large, slow datasets that can cause RStudio to freeze. Be very careful to be sure which dataset you are working with and to label results output accordingly.

8.3 Unix

Though bash is very commonly used for management of your file system (see Chapter 10), it is also a very capable at doing basic data manipulation with big data. At the core, since the data is stored on disk, you avoid having to overload memory when using bash commands as it will work with the files directly. By default, these commands will print the results to standard output (probably your

terminal screen), but you can then redirect the results to other files on disk to save your results. These commands can also be chained via pipes (represented as `|`, similar to `%>%` in tidyverse). All of these have a list of arguments that can be passed in via flags (check the `man` page for more details on each).

Command	Description
`head/tail`	Displays the first few or last few rows of a file
`cat`	Concatenates files and prints them
`sort`	Sorts the file
`cut`	Cuts out portions of each line and prints it
`grep`	Finds lines of a file that matches inputted patterns
`sed`	Find and replace
`awk`	Similar to `grep` and `sed` but with some extra programmatic functionality
`uniq`	Unifies repeated lines (combine with `sort` to get unique rows)
`wget / curl`	Downloads data/files from websites

8.4 SQL and `dbplyr`

SQL databases are relational databases that are a collection of *tables* that consists of *fields* or *attributes*, each containing a single *type*. If you use `dplyr` a lot, you will find that it is heavily inspired with a SQL flavor in mind. Formally, data gets loaded onto a database system and it is stored on disk. This alone makes working with data fast, but the real efficiency gain is in the concept of indexing. If you are curious, most SQL databases implement their index with B trees or B+ trees, which allow for log time complexity for search operations in average and worst case scenarios while providing constant time complexity in best case scenario.

The basic structure of a SQL query is as follows:

```
SELECT [DISTINCT] (attributes)
FROM (table)
[WHERE (conditions)]
[GROUP BY (attributes) [HAVING (conditions)]]
[ORDER BY (attributes) [DESC]]
```

The equivalent `dplyr` command would look as such:

```
table %>%
  select(attributes) %>%     # distinct(attributes) for select distinct
  group_by(attributes) %>%   #
  filter(conditions) %>%     #
```

```
  arrange(attributes)        # arrange(desc(attributes)) for descending
```

There is ample support for connection to databases in R, and, in particular, there is the `dbplyr` package, which allows you to interface with the data with `dplyr` code instead of SQL code.

8.5 data.table and dtplyr

It is often possible to load large datasets into memory in R, but computations will require more consumption of memory and will probably be very slow. One way around this is to use `data.table`. You will find that operations on the data are much faster than base R or `dplyr` even though data is loaded into memory - this is because of clever programming in C as well as internally creating a *key* (the SQL equivalent of an index) by default when loading in the data. You can improve on this even more by setting extra keys for variables you know you will be doing filter or join operations on.

More recently from the tidyverse, is the implementation of `dtplyr`, which allows for `dplyr` syntax on `data.table` objects.

An overview of the `dplyr` vs `data.table` debate can be found in this stackoverflow post and all 3 answers are worth a read.

8.6 ff, bigmemory, biglm

Sometimes, it may be impossible to load data into memory. Because of the overhead required, you can expect at least twice as much memory needed as the size of the file on disk to just load in a sufficiently large dataset and with all the other things that your computer needs to put on RAM in order to just run, you'll run out of space. One way to work around this is to keep the data on disk and instead create clever data structures that allow for natural interfacing with the data by mapping operations to the data on disk. Two R packages that implement these ideas are `ff` and `bigmemory`.

We can interface with the data while avoiding loading it into memory with these packages, but we run into issues when we try to fit models on it. For an $n \times p$ dataset, linear regression has a time complexity of $O(np^2 + p^3)$ and a space complexity of $O(np + p^2)$ (this just means it will take a while and take up a lot of space for large n and even moreso for large p). So even if we *could* load the data on disk, fitting these models would be out of question. This is where standard solutions used in machine learning (iterative algorithms like stochastic gradient descent) can help us. The idea is to take a smaller portion of our data (which will fit in memory), fit the regression (which will take a reasonable amount of time), then update the coefficient based on another run of linear regression on another small portion of the data until convergence. For GLM models, this can be done with the `biglm` package which has integration with `ff` and `bigmemory`.

8.7 Parallel computing

8.7.1 Embarrassingly Parallel Problems

Sometimes, we have to do something in a loop-like structure where each iteration may be independent of each other, such as simulations or bootstrap. These types of loops are referred to as *embarrassingly parallel* problems. Each iteration takes some time and every iteration thereafter must wait because the loop is operating as a queue, which gives a very obvious way to parallelize (hence "embarrassingly"). Every computer these days come with at least 2 cores in the CPU and each CPU core can operate independently, so after some overhead, we can speed up our loop by about a factor of the number of cores our computers have.

8.7.2 Packages

In R, the popular packages to do this are `parallel`, `foreach`, and `doParallel` (the backend that connects `foreach` and `parallel`). More modern parallel computing packages in R are `future` and `furrr` (inspired by `future purrr`, it allows for `purrr` like syntax using the `future` data structure from its namesake package). In Python, the `Dask` library has similar functionality to R's `future`. Note: the `parallel` package comes with a `detectCores` function, but I sometimes find that it is not accurate. On a Mac, you can manually check the number of cores by going into About This Mac then System Report then checking the Total Number of Cores in the Hardware tab.

8.7.3 GPU's

For most everyday tasks, CPU will be sufficient, but for large problems even an 8x speed boost from a computer with 8 cores might not be enough. This is where GPU's come into play. While CPU cores are good at complex operations, GPU cores are good at many small operations like matrix multiplication. GPU cores come in the hundreds for cheaper graphics cards and thousands for top end graphics cards, so they are ideal for training machine learning models, particularly neural networks. However, as these GPU cores were intended for the rendering of graphics on our computers, we cannot easily access their computing power out of R or Python without some translation in between. Graphics manufacturers have been catching up to this market, with one of the most popular platforms for parallel computing on GPU's being Nvidia's CUDA for use with Nvidia graphics cards.

8.7.4 The *MapReduce* paradigm

The idea of the *MapReduce* paradigm is that we can distribute the data across many nodes and try to do the computation on each piece of the data in each node. One benefit of this is that if our data is too large to fit on disk for a single

machine, we can instead spread it across many then do our operations in parallel and aggregate the results back together. We can formalize this paradigm into three steps

- Map: Split the data into sub-datasets and perform an operation on each entry in each sub-dataset thereby creating key-value pairs.
- Shuffle: Merge the key-value pairs and sort them.
- Reduce: Apply an operation on the associated values for each key.

An excellent example is included at the bottom of this link. A similar paradigm that is implemented in the tidyverse is the *split-apply-combine* strategy.

The popular infrastructures for doing parallel computing with the MapReduce paradigm are Hadoop and Spark (think of Spark as an in-memory version of Hadoop). Spark can be more easily interfaced with than Hadoop through Python via `PySpark` and through R via `SparkR` (from Apache) or `sparklyr` (from RStudio). However, note that because Spark is natively implemented in Java and Scala, the overhead of serialization between R/Python to Java/Scala may be a time expensive operation.

8.8 Optimal RStudio set up

Using the following settings will help ensure a smooth experience when working with big data. In RStudio, go to the "Tools" menu, then select "Global Options". Under "General":

Workspace

- **Uncheck** Restore RData into workspace at startup
- Save workspace to RData on exit – choose **never**

History

- **Uncheck** Always save history

Unfortunately RStudio often gets slow and/or freezes after hours working with big datasets. Sometimes it is much more efficient to just use Terminal / gitbash to run code and make updates in git.

Chapter 9

GitHub and Version Control

Contributors: Stephanie Djajadi, Nolan Pokpongkiat, and Ben Arnold

9.1 Basics

Git is a version control system. It has very good documentation online: https://git-scm.com/doc

If you get into trouble with Git, the docs will help a lot!

Git is often used in conjunction with GitHub, which is an online platform to help teams collaborate while using Git for version control: https://github.com.

- A detailed tutorial of Git can be found here on UC Berkeley's CS61B website.
- If you are already familiar with Git, you can reference the summary at the end of Section B.
- If you have made a mistake in Git, you can refer to this article to undo, fix, or remove commits in git.

There have been some great articles in PLOS Computational Biology that describe using GitHub for academic research. These papers are great introductions and summary of best practices!

Blischak JD, Davenport ER, Wilson G. A Quick Introduction to Version Control with Git and GitHub. *PLoS Comput Biol.* 2016;12: e1004668. https://www.ncbi.nlm.nih.gov/pubmed/26785377

Perez-Riverol Y, Gatto L, Wang R, Sachsenberg T, Uszkoreit J, Leprevost F da V, et al. Ten Simple Rules for Taking Advantage of Git and GitHub. *PLoS*

Comput Biol. 2016;12: e1004947. https://www.ncbi.nlm.nih.gov/pubmed/27415786

9.2 Git Branching

A terrific overview of branching workflow and its rationale is here: https://guides.github.com/introduction/flow/

Branches allow you to keep track of multiple versions of your work simultaneously, and you can easily switch between versions and merge branches together once you've finished working on a section and want it to join the rest of your code. Here are some cases when it may be a good idea to branch:

- You may want to make a dramatic change to your existing code (called refactoring) but it will break other parts of your project. But you want to be able to simultaneously work on other parts or you are collaborating with others, and you don't want to break the code for them.
- You want to start working on a new part of the project, but you aren't sure yet if your changes will work and make it to the final product.
- You are working with others and don't want to mix up your current work with theirs, even if you want to bring your work together later in the future.

A detailed tutorial on Git Branching can be found here. You can also find instructions on how to handle merge conflicts when joining branches together.

9.3 Example Workflow

A standard workflow when starting on a new project and contributing code looks like this:

Command	Description
SETUP: FIRST TIME ONLY: `git clone <url> <directory_name>`	Clone the repo. This copies of all the project files in its current state on Github to your local computer.
1. `git pull origin master`	update the state of your files to match the most current version on GitHub
2. `git checkout -b <new_branch_name>`	create new branch that you'll be working on and go to it
3. Make some file changes	work on your feature/implementation
4. `git add <filename>`	add file to stage for commit
5. `git commit -m <commit message>`	commit file with a message
6. `git push -u origin <branch_name>`	push branch to remote and set to track (-u only works if this is first push)
7. Repeat step 4-5.	work and commit often

Command	Description
8. `git push`	push work to remote branch for others to view
9. Follow the link given from the `git push` command to submit a pull request (PR) on GitHub online	PR merges in work from your branch into master
(10.) Your changes and PR get approved, your reviewer deletes your remote branch upon merging	
11. `git fetch --all --prune`	clean up your local git by untracking deleted remote branches

Other helpful commands are listed below.

9.4 Commonly Used Git Commands

Command	Description
`git clone <url> <directory_name>`	clone a repository, only needs to be done the first time
`git pull origin master`	pull before making any changes
`git branch`	check what branch you are on
`git branch -a`	check what branch you are on + all remote branches
`git checkout -b <new_branch_name>`	create new branch and go to it (only necessary when you create a new branch)
`git checkout <branch name>`	switch to branch
`git add <file name>`	add file to stage for commit
`git commit -m <commit message>`	commit file with a message
`git push -u origin <branch_name>`	push branch to remote and set to track (-u only works if this is first push)
`git branch --set-upstream-to origin <branch_name>`	set upstream to origin/ (use if you forgot -u on first push)
`git push origin <branch_name>`	push work to branch
`git checkout --track origin/<branch_name>`	pulls a remote branch and creates a local branch to track it (use when trying to pull someone else's branch onto your local computer)

Command	Description
`git push --delete <remote_name> <branch_name>`	delete remote branch
`git branch -d <branch_name>`	deletes local branch, -D to force
`git fetch --all --prune`	untrack deleted remote branches

9.5 How often should I commit?

Stephanie and Nolan (trained in CS and Data Science) suggest it is good practice to commit every 15 minutes (a time-based guideline), or every time you make a significant change (progress-based guideline). Ben's perspective aligns with this view, but is weighted toward committing around completion of discrete chunks of work; for him, a discrete chunk of work will often take quite a bit longer than 15 minutes time. Take home message: *It is better to commit more rather than less.*

9.6 What should be pushed to Github?

In general, it is better to track text-based files (`.R`, `.Rmd`, `.md`,`.txt`,etc…) compared with binary files (`.pdf`,`.png`,`.docx`, etc…) because Git will store changes to a binary file as a completely new file in your Git directory. If you store 100 versions of the same binary file, your directory will quickly become very bloated. If the binary files don't change often, then you could consider including them under version control, but it is usually cleaner to keep them under a separate version control, such as through an Open Science Framework project with a specific data component.

Be careful before you push .Rout log files! If someone else runs an R script and creates an .Rout file at the same time and both of you try to push to github, it is incredibly difficult to reconcile these two logs. If you run logs, keep them on your own system or (preferably) set up a shared directory where all logs are name and date timestamped.

There is a standardized `.gitignore` for `R` which you can download and add to your project. This ensures you're not committing log files or things that would otherwise best be left ignored to GitHub. This is a great discussion of project-oriented workflows, extolling the virtues of a self-contained, portable projects, for your reference.

9.7 How should I describe my commit?

When you commit, always include a short commit message that describes what the commit does. In the command line, you can achieve this after you have staged files to commit with the `commit -m <"your commit message here">` syntax. This helps track-back through work flow. For example, if the commit message is `new`, that doesn't provide any information about what the commit includes. A more descriptive commit message would be `Create first draft of Fig 1 distribution plot` or `Change color scheme for distribution plot`.

For more lengthy and detailed commit messages, which go beyond the simple, single line `commit -m <your commit message here>` synatax, this Medium post on the anatomy of a good commit message includes additional discussion. (Note, we don't typically use commits that are this detailed!)

Chapter 10

UNIX Commands

Contributors: Stephanie Djajadi, Kunal Mishra, Anna Nguyen, Jade Benjamin-Chung, and Ben Arnold

We typically use Unix commands in Terminal (for Mac users) or Git Bash (for Windows users) to

1. Run a series of scripts in parallel or in a specific order to reproduce our work
2. To check on the progress of a batch of jobs
3. To use git and push to github

10.1 Environment

On Mac OS, there is an application named `Terminal` that provides a bash shell interface to Unix. In Windows, one option is to install the `git for Windows` package: https://gitforwindows.org/.

The default coloring in a terminal window is pretty basic. If you want to make it more colorful in Mac OS, you can do that by saving a `.bash_profile` file in your home directory (note the "." prefix on the file name). This is one example of how you can add color to your terminal by including custom coloring in your bash profile (copied from Ben's profile):

```
# color terminal
export CLICOLOR=1
export LSCOLORS=GxFxCxDxBxegedabagaced
export PS1='\[\033[01;32m\]\u@\h\[\033[00m\]:\[\033[01;34m\]\w\[\033[00m\]\$ '
```

The encoding is extremely cryptic, but there are decodings online (e.g., link).

Another bash shell that provides a large array of colors for Mac OS is iTerm2 (https://iterm2.com/). There are over 200 color schemes to choose from:

https://github.com/mbadolato/iTerm2-Color-Schemes.

10.2 Basics

On the computer, there is a desktop with two folders, `folder1` and `folder2`, and a file called `file1.` Inside `folder1`, we have a file called `file2.` Mac users can run these commands on their terminal; we recommend that Windows users use Git Bash, not Windows PowerShell.

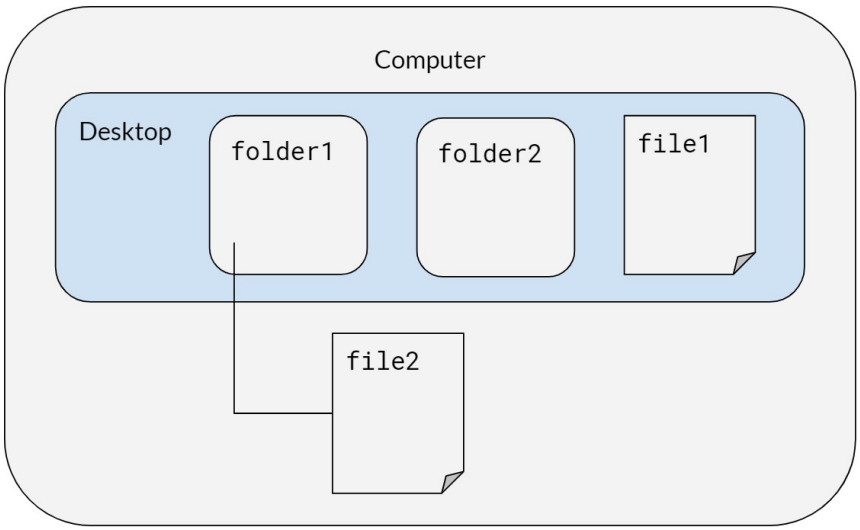


Figure 10.1: Here is our example desktop.

10.3 Syntax for both Mac/Windows

When typing in directories or file names, quotes are necessary if the name includes spaces.

Command	Description
`cd desktop/folder1`	Change directory to `folder1`
`pwd`	Print working directory
`ls`	List files in the directory
`cp "file2" "newfile2"`	Copy file (remember to include file extensions when typing in file names like `.pdf` or `.R`)
`mv "newfile2" "file3"`	Rename `newfile2` to `file3`

Command	Description
`cd ..`	Go to parent of the working directory (in this case, `desktop`)
`mv "file1" folder2`	Move `file1` to `folder2`
`mkdir folder3`	Make a new folder in `folder2`
`rm <filename>`	Remove files
`rm -rf folder3`	Remove directories (`-r` will attempt to remove the directory recursively, `-rf` will force removal of the directory)
`clear`	Clear terminal screen of all previous commands

10.4 Running Bash Scripts

Windows	Mac / Linux	Description
`chmod +750 <filename.sh>`	`chmod +x <filename.sh>`	Change access permissions for a file (only needs to be done once)
`./<filename.sh>`	`./<filename.sh>`	Run file (`./` to run any executable file)
`bash bash_script_name.sh`	`bash bash_script_name.sh &`	Run shell script in the background
`&`		

10.5 Running Rscripts in Windows

Note: This code seems to work only with Windows Command Prompt, not with Git Bash.

When R is installed, it comes with a utility called Rscript. This allows you to run R commands from the command line. If Rscript is in your `PATH,` then typing Rscript into the command line, and pressing enter, will not error. Otherwise, to use Rscript, you will either need to add it to your PATH (as an environment variable), or append the full directory of the location of Rscript on your machine. To find the full directory, search for where R is installed your computer. For instance, it may be something like below (this will vary depending on what version of R you have installed):

```
C:\Program Files\R\R-3.6.0\bin
```

For appending the `PATH` variable, please view this link. I strongly recommend completing this option.

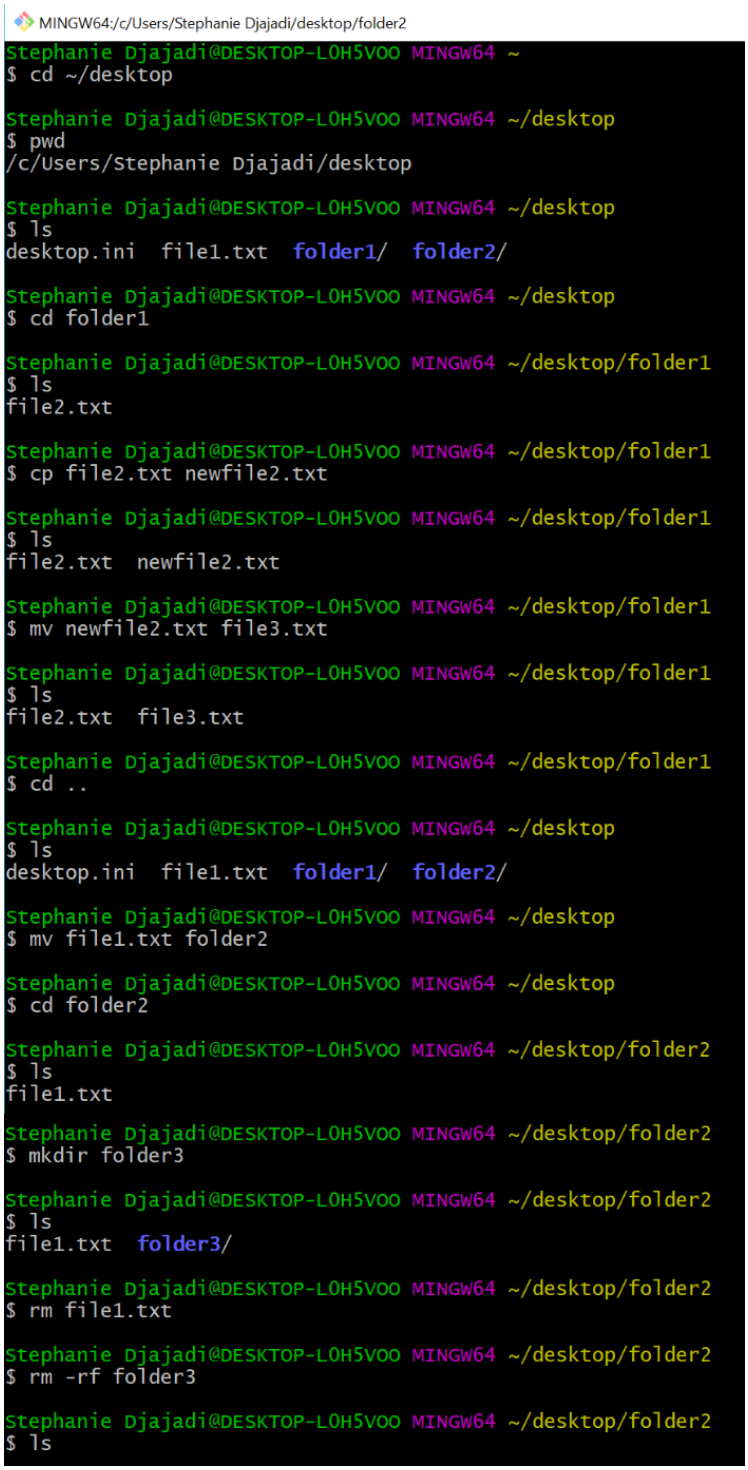


Figure 10.2: Here is an example of what your terminal might look like after executing the commands in the order listed above.

If you add the PATH as an environment variable, then you can run this line of code to test: `Rscript -e "cat('this is a test')"`, where the `-e` flag refers to the expression that will be executed.

If you do not add the PATH as an environment variable, then you can run this line of code to replicate the results from above: `"C:\Program Files\R\R-3.6.0\bin" -e "cat('this is a test')"`

To run an R script from the command line, we can say: `Rscript -e "source('C:/path/to/script/some_code.R')"`

10.5.1 Common Mistakes

- Remember to include all of the quotation marks around file paths that have a spaces.
- If you attempt to run an R script but run into `Error: '\U' used without hex digits in character string starting "'C:\U"`, try replacing all \ with \\ or /.

10.6 Checking tasks and killing jobs

Windows	Mac / Linux	Description
`tasklist`	`ps -v`	List all processes on the command line
	`top -o [cpu/rsize]`	List all running processes, sorted by CPU or memory usage
`taskkill /F /PID pid_number`	`kill <PID_number>`	Kill a process by its process ID
`taskkill /IM "process name" /F`		Kill a process by its name
`start /b program.exe`		Runs jobs in the background (exclude `/b` if you want the program to run in a new console)
	`nohup`	Prevents jobs from stopping
	`disown`	Keeps jobs running in the background even if you close R
`taskkill /?`		Help, lists out other commands

To kill a task in Windows, you can also go to Task Manager > More details > Select your desired app > Click on End Task.

10.7 Running big jobs

For big data workflows, the concept of "backgrounding" a bash script allows you to start a "job" (i.e. run the script) and leave it overnight to run. At the top level, a bash script (`0-run-project.sh`) that simply calls the directory-level bash scripts (i.e. `0-prep-data.sh`, `0-run-analysis.sh`, `0-run-figures.sh`, etc.) is a powerful tool to rerun every script in your project. See the included example bash scripts for more details.

- **Running Bash Scripts in Background**: Running a long bash script is not trivial. Normally you would run a bash script by opening a terminal and typing something like `./run-project.sh`. But what if you leave your computer, log out of your server, or close the terminal? Normally, the bash script will exit and fail to complete. To run it in background, type `./run-project.sh &; disown`. You can see the job running (and CPU utilization) with the command `top` or `ps -v` and check your memory with `free -h`.

Alternatively, to keep code running in the background even when an SSH connection is broken, you can use `tmux`. In terminal or gitbash follow the steps below. This site has useful tips on using `tmux`.

```
# create a new tmux session called session_name
tmux new -ssession_name

# run your job of interest
R CMD BATCH myjob.R &

# check that it is running
ps -v

# to exit the tmux session (Mac)
ctrl + b
d

# to reopen the tmux session to kill the job or
# start another job
tmux attach -tsession_name
```

- **Deleting Previously Computed Results**: One helpful lesson we've learned is that your bash scripts should remove previous results (computed and saved by scripts run at a previous time) so that you never mix results from one run with a previous run. This can happen when an R script errors out before saving its result, and can be difficult to catch because

your previously saved result exists (leading you to believe everything ran correctly).

- **Ensuring Things Ran Correctly**: You should check the `.Rout` files generated by the R scripts run by your bash scripts for errors once things are run.

Chapter 11

Building Automated Markdown Reports

"Never forget Murphy's Law 2.0: Anything that can't go wrong, will go wrong."
-T. Porco

Contributors: Will Godwin, Ying Lin, Ben Arnold

Many teams at Proctor conduct trials that utilize electronic data collection, with data flowing in daily. In order to identify potential errors in the field, we have found it very valuable to download, process, and visualize the data on a weekly and daily basis. On numerous occassions, Thanks to monitoring reports, we have been able to identify and rectify impossibly tall 3-month olds or a toddler with different birthdates on separate interviews. Proactively identifying preventable error saves hassle when processing the final data and ultimately makes trials more accurate and efficient. We visualize the data using reports of tables and figures generated in R markdown. This document provides a template for building your own R markdown report, whether for DSMC meeting or internal use. I've also included an example script to show how a report can look in practice.

11.1 Setting up the markdown

11.1.1 YAML

The YAML is the opening header to most R markdown documents and allows the user to specify parameters regarding the layout and output of the document. Among other things, here is where you can add a table of contents, a subtitle,

or specify the default output to be pdf file. For a monitoring report, the YAML might look as follows:

```
---
title: "EXAMPLE Weekly Report"
subtitle:
- <h1><u>CONFIDENTIAL-DO NOT DISSEMINATE</u></h1>
author: "Will Godwin"
date: "08 May 2024"
output:
  pdf_document:
    toc: true
    toc_depth: 2
---
```

11.1.2 Directory structure

Your report may begin as a couple files in a folder but as it grows in size and complexity, I highly recommend systematizing and standardizing the process for posterity and your own sanity. Refer to chapter 3 for guidance on building a standardized directory structure and chapter 4 for information on coding practices and utilizing the *here* package, which I use throughout the process.

11.2 Automating the data export process

Creating a system that enables you to run some code and download new data in a systematic manner, as opposed to manually downloading the data, is critical to saving time and headaches down the road. This is especially important if you're building a report that will be run weekly or daily. It may require a little more work on the front end but programmatically downloading data rather than manually will likely be faster, better documented, and standardized. If you work on a team that uses the World Bank's *Survey Solutions* for data collection and storage, there is a great API tool that allows you to export data directly from the project's server. Example code for that process can be found here. Salesforce, REDCap, and other electronic data capture systems likely have a similar functionality that other Proctor researchers may be more familiar with.

11.3 Loading the data

Depending on the nature of the trial, there may be multiple files to load after exporting the data. I would suggest using a table (ie simple .csv file) to track

all necessary files for the report, as shown below. Refer to the example script for how this looks in a setting that utilizes *Survey Solutions*.

	A	B	C	D	E
1	form	id	primary	type	version
2	eligibility	b8a61d18-77	1	tabular	1
3	baseline	27627906-82	1	tabular	1
4	anthro	17457a9d-44	1	tabular	1
5	treatment	0b4220c2-06	1	tabular	1
6	followup	0d968910-66	1	tabular	1
7					

11.4 Data cleaning/processing

I won't focus explicitly on data cleaning and processing since this process is usually fairly project specific. Briefly, after reading data into R, spend some time making sure all relevant variables (columns) were loaded and checking that variable classes are encoded as expected. Below is an example that checks the structure of the data and converts all variables that should be "Date" class to be so.

```
#convert all variables with "date" within the name to date class
dt.anthro %<>% mutate_at(vars(matches("date")), as.Date)

#convert all values within "childID" variable to lowercase
dt.anthro %<>% mutate_at(vars(matches("childID")), tolower)
```

11.5 Data monitoring

Now that we have data read into R and structured appropriately, we can focus on the (2) primary aims of the report: data validation to be used internally and DSMC-style presentation of the data. Of course the aims of your report may vary but the monitoring examples below represent evaluations critical to almost any RCT.

11.5.1 Data validation

11.5.1.1 Missing/inappropriately structured data

While not the most glamorous component of the report, checking for missing data is critical. It's helpful once all the data have been read in, to conduct some important checks of the data integrity. Did R (or python, etc) correctly interpret and load the data files? Are there any participant IDs missing? Are there participant's with reported birthdates in the future? This is an important

stage to think about the plethora of maladies that could befall the data in between when it was recorded and when it got to you. Below are examples of how you could check for some common issues.

```
#check for missing values across all columns
map(dt.anthro, ~sum(is.na(.)))

#check for missing values in childsex
na_childSex <- sum(is.na(dt.anthro$childSex))

There are 4 rows with missing values for child sex.
```

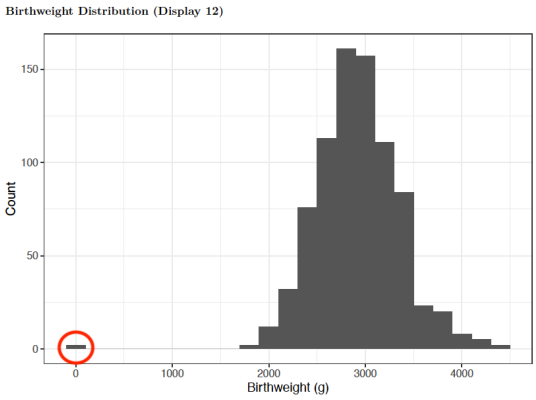
11.5.1.2 Reporting inconsistencies

Many studies may use multiple independent questionnaires to gather information at different timepoints and ask the same demographic information. For example, in an individually randomized trial, each participant receives an ID that links the participant responses across all timepoints of the study. Below is an example showing how to check that the responses are consistent across all questionnaires.

```
#join anthro and swabs data by childID to check where childSex is different
dt.anthro %>%
  select(childSex, childID) %>%
  inner_join(
    dt.swabs %>% select(childSex, childID),
    by="childID"
  ) %>%
  filter(childSex.x!=childSex.y)
```

Another component of data validation is checking for overt errors and outliers, many of which can often be fixed in the field. One example is when interviewers may accidentally record a study participant's weight in kilograms instead of grams, resulting in an error 3 magnitudes in size. This error is easily identified (red circle) from plotting the distribution of the birthweights in a histogram.

```
#age histogram
dt.anthro %>%
  mutate(weeks = difftime(startTime, date_of_birth, units = "weeks")) %>%
  ggplot(dt.anthro, aes(weeks)) +
    geom_histogram(binwidth = .2) +
    theme_bw() +
    xlab("Age (weeks)") +
    ylab("Count")
```



Distribution of participant birthweights in grams

11.5.2 Data presentation (DSMC-style)

The ultimate arbiter of whether a trial remains running is often the Data and Safety Monitoring Committee (DSMC) and these reports can be used to present preliminary findings/analyses as well as detail all relevant safety metrics.

11.5.2.1 Enrollment

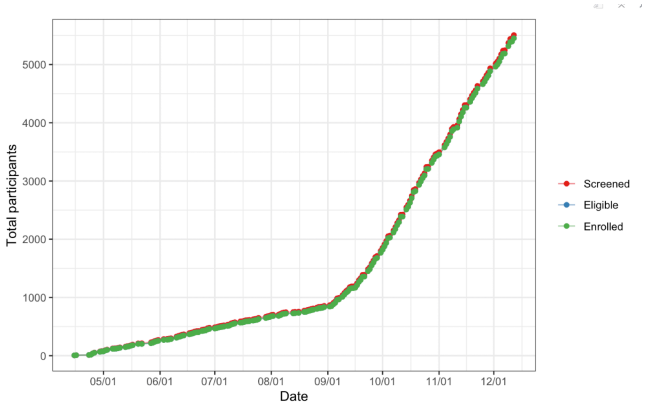
Any study investigator cares a lot about enrollment, since it determines the length and cost of the trial, among other things. Therefore, any report needs tables/figures showing enrollment trends and trajectory. Below are a couple examples of what they can look like.

Enrollment counts-total (Display 1):

Screened	Eligible	Refused	Enrolled
3864	3824	2	3822

Enrollment counts-past 7 days (Display 2):

Screened	Eligible	Refused	Enrolled
369	367	0	367



Enrollment counts across the length of the study period.

11.5.2.2 Baseline characteristics

Demographic characteristics of the enrolled participants at baseline are often necessary.

Age and sex breakdown (Display 14)

Region	Female	[8-14) Day	[14-21) Day	[21-28) Day	Total
Region 2	49.9%	79.0%	13.8%	7.19%	501
Region 1	50.2%	48.1%	27.9%	24.0%	1263
Region 6	50.4%	83.8%	11.3%	4.83%	1863
Region 4	48.0%	78.3%	15.0%	6.74%	861
Region 3	48.8%	78.5%	15.6%	5.96%	688
Region 5	48.4%	70.5%	22.9%	6.55%	275
Overall	**49.6%**	**72.9%**	**17.1%**	**10.0%**	**5451**

11.5.2.3 Treatment

There is usually a treatment component to an RCT, where participants are treated according to their assigned arm (i.e. treatment or placebo). The report should include checks that confirm particpants actually received the treatment that they were randomized to. Otherwise, the results of the study lack interpretability and are likely null.

```
#merge with treatment letters
wrong_random <- dt.treatment %>%
  left_join(dt_letters, by="childID") %>%
  #the treatment letter does not match what's in database
  tally(TL != assignment) %>% as.numeric()
```

```
Number of IDs where treatment letter given does not match database: 0
```

Treatment coverage (Display 15)

Region	Treated	Enrolled	Coverage
Region 1	1263	1262	100%
Region 2	501	501	100%
Region 3	688	688	100%
Region 4	861	861	100%
Region 5	275	275	100%
Region 6	1863	1863	100%
Overall	**5451**	**5451**	**100%**

Counts of number treated and number eligible, along with percent treated (coverage), by region.

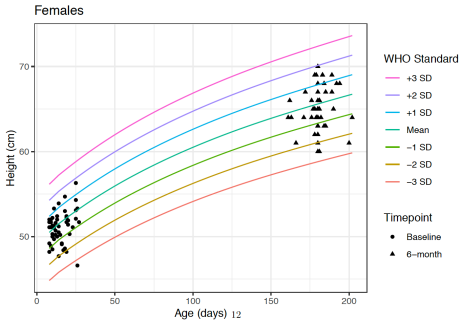
11.5.2.4 Primary/secondary outcomes (*masked!*)

This is probably the most exciting section for an investigator to see since these metrics are the motivation of the trial itself. Be especially vigilient to make sure that the figures remain masked when sending to trial investigators and coordinators. This means making sure that you send no information regarding the treatment assignments while the trial is still active.

Six-month mortality (Display 23)

Region	Deaths	Person-years	Mortality (per 1000 py)
Region 1	4	226.81	17.64
Region 2	0	119.21	0.00
Region 3	2	55.28	36.18
Region 4	0	52.99	0.00
Region 5	0	17.98	0.00
Region 6	2	132.39	15.11
Total	**8**	**604.67**	**13.23**

Preliminary mortality rate and total person-years by region.



Height growth curves of participants at Baseline and 6-month timepoints.

11.5.2.5 Follow-up/Adverse events

This report can serve as guard against loss to follow-up, an issue that plagues many trials. By actively monitoring the number of participants that have, and have not, been followed-up on, we can communicate where participants or clusters need to more attention to workers in the field. Below is an example showing follow-up percentages by region.

Follow-up verification (Display 17)

Region	Follow-up
Region 2	99.8%
Region 1	99.8%
Region 6	99.8%
Region 4	99.7%
Region 3	100%
Region 5	100%
Overall	**99.8%**

Children with at least 1 follow-up in 21 days post-treatment

As mentioned above, adverse events should be tabulated and closely watched throughout the study to monitor if the intervention may be causing increased harm.

Adverse events (Display 20)

Region	Fever	Diarrhea	Vomiting	Abdominal Pain	Rash	Constipation
Region 1	67	23	25	29	19	32
Region 2	27	7	15	39	8	33
Region 3	68	12	6	18	5	10
Region 4	8	6	6	7	0	4
Region 5	0	0	0	0	2	0
Region 6	42	12	14	76	28	23
Overall	**212**	**60**	**66**	**169**	**62**	**102**

Counts of adverse events across all follow-up visits.

11.5.3 Appendix (if necessary)

Sometimes an appendix can be helpful to display long lists or addendums to tables shown earlier in the report. Below is an example of a list enumerating all adverse events that were not included in the AE table within the main report.

List of "Other" reasons hospital/health center visited

```
##   [1] "-1ERE VISITE : DIARRHÉE+VOMISSEMENTS-2EME VISITE :  FURONCULOSE"
##   [2] "-MALNUTRITION PUIS TOUX"
##   [3] "1ere vaccination 2mois"
##   [4] "abcès au niveau du sein gauche"
##   [5] "AUBSTRUCTION NASALE ET ETERNOUILLEMENT"
##   [6] "ballonement"
##   [7] "BALLONEMENTS"
##   [8] "BALLONNEMENT"
##   [9] "Ballonnement ABDOMINALE"
##  [10] "BALLONNEMENT ABDOMINALE"
##  [11] "BALLONNEMENT ABDOMINALE ET INFECTION URINAIRE"
##  [12] "BALLONNEMENT ABDOMINALE ET ROUGEUR ANALE"
```

11.6 Summary

This chapter described and illustrated an example of a data monitoring report using R markdown. This is, by no means, an exhaustive list of what to include in a monitoring report but meant to serve as a template as you create your own report. It is specifically geared towards continuous monitoring of randomized-controlled trials but the overarching concepts apply to other types of trials or even retrospective analyses.

Chapter 12

Communication and Coordination

Contributors: Jade Benjamin-Chung, Ben Arnold

These communications guidelines are evolving as we increasingly adopt Slack, but here some general principles if you work closely with Ben.

12.1 Slack

- If you work with Ben but are not a member of Proctor's Slack workspace then ask him to invite you!
- We do not recommend using Slack for messages you'd like to save and find again in 6 months – for messages with important information and documentation that you will need in the future, use email!
- Use Slack for scheduling, coding related questions, quick check ins, etc. If your Slack message exceeds 200 words, it might be time to use email.
- Use channels instead of direct messages unless you need to discuss something private.
- Include tags on your message (e.g., `@Ben`) when you want to ensure that a person sees the message. Ben doesn't regularly read messages where he isn't tagged.
- Please make an effort to respond to messages that message you (e.g., `@Ben`) as quickly as possible and always within 24 hours, unless of course you are on vacation!

- If you are unusually busy (e.g., taking MCAT/GRE, taking many exams) or on vacation please alert the team in advance so we can expect you not to respond at all / as quickly as usual and also set your status in Slack (e.g., it could say "On vacation") so we know not to expect to see you online.
- Please thread messages in Slack as much as possible.

12.2 Email

- Use email for longer messages (>200 words) or messages that merit preservation.
- Generally, strive to respond within 24 hours. If you are unusually busy or on vacation please alert the team in advance so we can expect you not to respond at all / as quickly as usual.

12.3 Trello

- Ben manages projects and teams using a kanban board approach in Trello.
- You and/or Ben will add new cards within our team's Trello boards and assign them to team members.
- Each card represents a discrete chunk of work.
- Cards higher in a list are higher priority.
- Strive to complete the tasks in your card by the card's due date. Talk to Ben about deadlines – we can always manage the calendar!
- Use checklists to break down a task into smaller chunks. Usually, you can do this yourself (but ask Ben if you ever want input).
- Move cards to the "DONE" list on a board when they are done.

12.4 Google Drive

- We mostly use Google Drive to create shared documents with longer descriptions of tasks. These documents are linked to Trello cards. Ben often shares these docs with a whole project team since tasks are overlapping, and even if a task is assigned to one person, others may have valuable insights.
- Please invite both of Ben's email addresses to any documents you create (bfarnold@gmail.com, ben.arnold@ucsf.edu).

12.5 Calendar / Meetings

- Ben will schedule most meetings through the calendar.
- Our meetings start on the hour.
- If you are going to be late, please send a message in our Slack channel.

- If you are regularly not able to come on the hour, notify the team and we might choose the modify the agenda order or the start time.
- Add hoc meetings are welcome. If Ben's office door is open, come in!

Chapter 13

Code of conduct

Contributors: Jade Benjamin-Chung, Ben Arnold

13.1 Group culture

We strive to work in an environment that is collaborative, supportive, open, and free from discrimination and harassment, per University policies.

We encourage students / staff of all experience levels to respectfully share their honest opinions and ideas on any topic. Our group has thrived upon such respectful honest input from team members over the years, and this document is a product of years of student and staff input (and even debate) that has gradually improved our productivity and overall quality of our work.

If Ben is your PI, be forewarned that he tends to batch his email communication (~30 mins in the morning and afternoon, 15 mins mid-day), and doesn't tend to answer Slack or email during evenings or weekends. If you need to reach him urgently then give him a call or text on his mobile.

13.2 Protecting human subjects

All lab members must complete CITI Human Subjects Training and share their certificate with Ben. We will will add team members to relevant Institutional Review Board protocols to ensure they have permission to work with identifiable datasets.

One of the most relevant aspects of protecting human subjects in our work in the Data Coordinating Center is maintaining confidentiality and data privacy. For students supporting our data science efforts, in practice this means:

- If you are using a virtual computer (e.g., Google Cloud, AWS, Optum), never save the data in that system to your personal computer or any other computer without prior permission.
- Do not share data with anyone without first obtaining permission, including to other members of the Proctor Foundation, who might not be on the same IRB protocol as you (check with Ben or the relevant PI first).
- **NEVER** push a dataset into the public domain (e.g., GitHub, OSF) without first checking with Ben to ensure that it is appropriately de-identified and we have approval from the sponsor and/or human subjects review board to do so.

Remember, data that looks like it does not contain identifiers to you might still be classified as data that requires special protection by our IRB or under HIPAA, so always proceed with caution and ask for help if you have any concerns about how to maintain study participant confidentiality. For example, the combination of age, sex, and geographic location of the individual's town or neighborhood is typically considered identifiable.

13.3 Authorship

We adhere to the ICMJE Definition of authorship and are happy for team members who meet the definition of authorship to be included as co-authors on scientific manuscripts.

13.4 Work hours

Please follow the Proctor Foundation's employee guidelines for work hours, and discuss the specifics with your PI. If Ben is your PI, then work with him on your schedule to ensure we have overlap in the office and that you are around at key times for group meetings, etc.

Chapter 14

Additional Resources

TBD