

# **UCD-SeRG Lab Manual**

Kristen Aiemjoy      Ezra Morrison

Last updated: 2026-01-10

# Contents

<b>1</b>	<b>Welcome to UCD-SeRG!</b>	<b>1</b>
1.1	About the lab . . . . .	1
1.2	About this lab manual . . . . .	1
<b>2</b>	<b>Culture and conduct</b>	<b>2</b>
2.1	Lab culture . . . . .	2
2.2	Diversity, equity, and inclusion . . . . .	2
2.3	Protecting human subjects . . . . .	2
2.4	Authorship . . . . .	3
<b>3</b>	<b>Communication and coordination</b>	<b>4</b>
3.1	Microsoft Teams . . . . .	4
3.2	Email . . . . .	4
3.3	Trello . . . . .	4
3.4	Google Drive . . . . .	5
3.5	UC Davis Box and SharePoint . . . . .	5
3.6	Meetings . . . . .	5
<b>4</b>	<b>Reproducibility</b>	<b>6</b>
4.1	What is the reproducibility crisis? . . . . .	6
4.2	Study design . . . . .	7
4.3	Register study protocols . . . . .	7
4.4	Write and register pre-analysis plans . . . . .	7
4.5	Create reproducible workflows . . . . .	7
4.6	Process and analyze data with internal replication and masking . . . . .	8
4.7	Use reporting checklists with manuscripts . . . . .	8
4.8	Publish preprints . . . . .	8
4.9	Publish data (when possible) and replication scripts . . . . .	8
<b>5</b>	<b>Code repositories</b>	<b>9</b>
5.1	Project Structure . . . . .	9
5.2	.Rproj files . . . . .	10
5.3	Configuration ('config') File . . . . .	10
5.4	Order Files and Directories . . . . .	10
5.5	Using Bash scripts to ensure reproducibility . . . . .	11
<b>6</b>	<b>R Coding</b>	<b>12</b>
6.1	Lab Protocols for Code and Data . . . . .	12
6.2	R Package Structure for All Projects . . . . .	12
6.3	Essential R Package Development Tools . . . . .	15
6.4	Complete Package Development Workflow . . . . .	17
6.5	Code Style Guidelines . . . . .	19
6.6	Testing Requirements . . . . .	24
6.7	Iterative Operations . . . . .	25

6.8	Reading and Saving Data . . . . .	26
6.9	Version Control and Collaboration . . . . .	27
6.10	Quality Assurance Checklist . . . . .	29
6.11	Continuous Integration . . . . .	30
6.12	Automated Code Styling . . . . .	30
6.13	Additional Resources . . . . .	31
<b>7</b>	<b>Big data</b>	<b>32</b>
7.1	The data.table package . . . . .	32
7.2	Using downsampled data . . . . .	32
7.3	Optimal RStudio set up . . . . .	32
<b>8</b>	<b>Data masking</b>	<b>34</b>
8.1	General Overview . . . . .	34
8.2	Technical Overview . . . . .	35
<b>9</b>	<b>Github</b>	<b>38</b>
9.1	Basics . . . . .	38
9.2	Github Desktop . . . . .	38
9.3	Git Branching . . . . .	38
9.4	Example Workflow . . . . .	39
9.5	Commonly Used Git Commands . . . . .	39
9.6	How often should I commit? . . . . .	40
9.7	What should be pushed to Github? . . . . .	40
<b>10</b>	<b>Unix</b>	<b>42</b>
10.1	Basics . . . . .	42
10.2	Syntax for both Mac/Windows . . . . .	43
10.3	Running Bash Scripts . . . . .	45
10.4	Running Rscripts in Windows . . . . .	45
10.5	Checking tasks and killing jobs . . . . .	45
10.6	Running big jobs . . . . .	46
<b>11</b>	<b>Reproducible Environments</b>	<b>50</b>
11.1	Package Version Control with renv . . . . .	50
<b>12</b>	<b>Code Publication</b>	<b>52</b>
12.1	Checklist overview . . . . .	52
12.2	Fill out file headers . . . . .	52
12.3	Clean up comments . . . . .	52
12.4	Document functions . . . . .	52
12.5	Remove deprecated filepaths . . . . .	53
12.6	Ensure project runs via bash . . . . .	53
12.7	Complete the README . . . . .	53
12.8	Clean up feature branches . . . . .	55
12.9	Create Github release . . . . .	55
<b>13</b>	<b>Data Publication</b>	<b>56</b>
13.1	Overview . . . . .	56
13.2	Removing PHI . . . . .	57
13.3	Create public IDs . . . . .	58
13.4	Create a data repository . . . . .	59

13.5 Edit and test analysis scripts . . . . .	60
13.6 Create a public GitHub page for public scripts . . . . .	60
13.7 Go live . . . . .	61
<b>14 High-performance computing (HPC)</b>	<b>62</b>
14.1 UC Davis Computing Resources . . . . .	62
14.2 Getting started with SLURM clusters . . . . .	62
14.3 Moving files to the cluster . . . . .	64
14.4 Installing packages on the cluster . . . . .	64
14.5 Testing your code . . . . .	65
14.6 Storage & group storage access . . . . .	67
14.7 Running big jobs . . . . .	68
<b>15 Checklists</b>	<b>70</b>
15.1 Pre-analysis plan checklist . . . . .	70
15.2 Code checklist . . . . .	70
15.3 Manuscript checklist . . . . .	70
15.4 Figure checklist . . . . .	71
<b>16 Resources</b>	<b>73</b>
16.1 Resources for R . . . . .	73
16.2 Resources for Git & Github . . . . .	73
16.3 Scientific figures . . . . .	73
16.4 Writing . . . . .	74
16.5 Presentations . . . . .	74
16.6 Professional advice . . . . .	74
16.7 Funding . . . . .	74
16.8 Ethics and global health research . . . . .	74

# 1 Welcome to UCD-SeRG!

## 1.1 About the lab

Welcome to the Seroepidemiology Research Group (SeRG) at the University of California, Davis, led by Drs. Kristen Aiemjoy and Ezra Morrison. Accurate methods to measure infectious disease burden are essential for guiding public health decisions, yet many infectious diseases remain under-recognized due to limited diagnostics and costly, resource-intensive surveillance systems. Our work addresses this gap by developing seroepidemiologic methods to characterize infection burden in populations. Currently, we focus on enteric fever (*Salmonella Typhi* and *Paratyphi*), Scrub Typhus (*Orientia tsutsugamushi*), Melioidosis (*Burkholderia pseudomallei*), Shigella (*Shigella spp.*), and Cholera (*Vibrio cholerae*). We are supported by the US National Institutes of Health, the Bill and Melinda Gates Foundation, and the Department of Defense, and collaborate with partners around the world. To learn more about the lab, visit [ucdserg.ucdavis.edu](https://ucdserg.ucdavis.edu)<sup>1</sup>.

## 1.2 About this lab manual

This lab manual covers our communication strategy, code of conduct, and best practices for reproducibility of computational workflows. It is a living document that is updated regularly.

This manual is a fork of the Benjamin-Chung Lab's manual<sup>2</sup>, adapted for UCD-SeRG. We are grateful to Dr. Jade Benjamin-Chung and her team for developing and openly sharing their excellent lab manual. You can view the original manual at [jadebc.github.io/lab-manual](https://jadebc.github.io/lab-manual)<sup>3</sup>. Original contributors include Jade Benjamin-Chung, Kunal Mishra, Stephanie Djajadi, Nolan Pokpongkiat, Anna Nguyen, Iris Tong, and Gabby Barratt Heitmann.

Feel free to draw from this manual (and please cite it if you do!).

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

---

<sup>1</sup><https://ucdserg.ucdavis.edu>

<sup>2</sup><https://github.com/jadebc/lab-manual>

<sup>3</sup><https://jadebc.github.io/lab-manual/index.html>

## 2 Culture and conduct

Adapted by UCD-SeRG team from original by Jade Benjamin-Chung<sup>1</sup>

### 2.1 Lab culture

We are committed to a lab culture that is collaborative, supportive, inclusive, open, and free from discrimination and harassment.

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.

### 2.2 Diversity, equity, and inclusion

UCD-SeRG recognizes the importance of and is committed to cultivating a culture of diversity, equity, and inclusion. This means being a safe, supportive, and anti-racist environment in which students from diverse backgrounds are equally and inclusively supported in their education and training. Diversity takes many forms, and includes, but is not limited to, differences in race, ethnicity, gender, sexuality, socioeconomic status, religion, disability, and political affiliation.

### 2.3 Protecting human subjects

All lab members must complete CITI Human Subjects Biomedical Group 1<sup>2</sup> training and share their certificate with the lab leadership. Team members will be added to relevant Institutional Review Board protocols prior to their start date to ensure they have permission to work with identifiable datasets.

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

- Be sure to understand and comply with project-specific policies about where data can be saved, particularly if the data include personal identifiers.
- Do not share data with anyone without permission, including to other members of the group, who might not be on the same IRB protocol as you (check with lab leadership first).

---

<sup>1</sup><https://jadebc.github.io/lab-manual/culture-and-conduct.html>

<sup>2</sup><https://research.ucdavis.edu/policiescompliance/irb-admin/education/>

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.

## **2.4 Authorship**

We adhere to the ICMJE Definition of authorship<sup>3</sup> and are happy for team members who meet the definition of authorship to be included as co-authors on scientific manuscripts.

---

<sup>3</sup><http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html>

# 3 Communication and coordination

Adapted by UCD-SeRG team from original by Jade Benjamin-Chung<sup>1</sup>

One benefit of the academic environment is its schedule flexibility. This means that lab members may choose to work in the early morning, evening, or weekends. That said, we do not expect lab members to respond outside of business hours (unless there are special circumstances).

## 3.1 Microsoft Teams

- Use Microsoft Teams for scheduling, coding related questions, quick check ins, etc. If your Teams message exceeds 200 words, it might be time to use email.
- Use channels instead of direct messages unless you need to discuss something private.
- Please make an effort to respond to messages that mention you (e.g., @username) as quickly as possible and always within 24 hours.
- 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 Teams (e.g., it could say “On vacation”) so we know not to expect to see you online.
- Please thread messages in Teams as much as possible.

## 3.2 Email

- Use email for longer messages (>200 words) or messages that merit preservation.
- Generally, strive to respond within 24 hours hours. As noted above, 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.

## 3.3 Trello

- Lab leadership will add new cards within our shared Trello board that outline your tasks.
- The higher a card is within your list, the higher priority it is.
- Generally, strive to complete the tasks in your card by the date listed.
- Use checklists to break down a task into smaller chunks. Sometimes leadership will write this for you, but you can also add this yourself.
- Lab leadership will move your card to the “Completed” list when it is done.

---

<sup>1</sup><https://jadebc.github.io/lab-manual/communication-and-coordination.html>

## 3.4 Google Drive

- We mostly use Google Drive to create shared documents with longer descriptions of tasks. These documents are linked to in Trello. Lab leadership often shares these with the whole team since tasks are overlapping, and even if a task is assigned to one person, others may have valuable insights.

## 3.5 UC Davis Box and SharePoint

- Human subjects data for research studies are generally stored in UC Davis Box or SharePoint. Please check with lab leadership about whether there are special storage and transfer requirements for the datasets you are working with for each study.
- You can access Box via your UC Davis credentials. For more information, visit UC Davis Box Support<sup>2</sup>.
- SharePoint is also used for collaborative document storage and team file sharing. Access SharePoint through your UC Davis Microsoft 365 account.

## 3.6 Meetings

- Our meetings start on the hour.
- If you are going to be late, please send a message in our Teams channel.
- If you are regularly not able to come on the hour, notify the team and we might choose to modify the agenda order or the start time.

---

<sup>2</sup><https://kb.ucdavis.edu/?id=6485>

# 4 Reproducibility

Adapted by UCD-SeRG team from original by Jade Benjamin-Chung<sup>1</sup>

Our lab adopts the following practices to maximize the reproducibility of our work.

1. Design studies with appropriate methodology and adherence to best practices in epidemiology and biostatistics
2. Register study protocols
3. Write and register pre-analysis plans
4. Create reproducible workflows
5. Process and analyze data with internal replication and masking
6. Use reporting checklists with manuscripts
7. Publish preprints
8. Publish data (when possible) and replication scripts

## 4.1 What is the reproducibility crisis?

In the past decade, an increasing number of studies have found that published study findings could not be reproduced. Researchers found that it was not possible to reproduce estimates from published studies: 1) with the same data and same or similar code and 2) with newly collected data using the same (or similar) study design. These “failures” of reproducibility were frequent enough and broad enough in scope, occurring across a range of disciplines (epidemiology, psychology, economics, and others) to be deeply troubling. Program and policy decisions based on erroneous research findings could lead to wasted resources, and at worst, could harm intended beneficiaries. This crisis has motivated new practices in reproducibility, transparency, and openness. Our lab is committed to adopting these best practices, and much of the remainder of the lab manual focuses on how to do so.

Recommended readings on the “reproducibility crisis”:

- Nuzzo R. How scientists fool themselves – and how they can stop. 2015. <https://www.nature.com/articles/526182a><sup>2</sup>
- Stoddart C. Is there a reproducibility crisis in science? 2016. <https://www.nature.com/articles/d41586-019-00067-3>
- Munafo MR, et al. A manifesto for reproducible science. *Nature Human Behavior* 2017 <http://dx.doi.org/10.1038/s41562-016-0021>

---

<sup>1</sup><https://jadebc.github.io/lab-manual/reproducibility.html>

<sup>2</sup>

## 4.2 Study design

Appropriate study design is beyond the scope of this lab manual and is something trainees develop through their coursework and mentoring.

## 4.3 Register study protocols

We register all randomized trials on [clinicaltrials.gov](https://clinicaltrials.gov), and in some cases register observational studies as well.

## 4.4 Write and register pre-analysis plans

We write pre-analysis plans for most original research projects that are not exploratory in nature, although in some cases, we write pre-analysis plans for exploratory studies as well. The format and content of pre-analysis plans can vary from project to project. Here is an example of one: <https://osf.io/tgbxr/>. Generally, these include:

1. Brief background on the study (a condensed version of the introduction section of the paper)
2. Hypotheses / objectives
3. Study design
4. Description of data
5. Definition of outcomes
6. Definition of interventions / exposures
7. Definition of covariates
8. Statistical power calculation
9. Statistical analysis:
  - Type of model
  - Covariate selection / screening
  - Standard error estimation method
  - Missing data analysis
  - Assessment of effect modification / subgroup analyses
  - Sensitivity analyses
  - Negative control analyses

## 4.5 Create reproducible workflows

Reproducible workflows allow a user to reproduce study estimates and ideally figures and tables with a “single click”. In practice, this typically means running a single bash script that sources all replication scripts in a repository. These replication scripts complete data processing, data analysis, and figure/table generation. The following chapters provide detailed guidance on this topic:

- Chapter 5: Code repositories
- Chapter 6: Coding practices
- Chapter 7: Coding style

- Chapter 8: Code publication
- Chapter 9: Working with big data
- Chapter 10: Github
- Chapter 11: Unix

## 4.6 Process and analyze data with internal replication and masking

See my video on this topic: <https://www.youtube.com/watch?v=WoYkY9MkbRE>

## 4.7 Use reporting checklists with manuscripts

Using reporting checklists helps ensure that peer-reviewed articles contain the information needed for readers to assess the validity of your work and/or attempt to reproduce it. A collection of reporting checklists is available here: <https://www.equator-network.org/about-us/what-is-a-reporting-guideline/><sup>3</sup>

## 4.8 Publish preprints

A preprint is a scientific manuscript that has not been peer reviewed. Preprint servers create digital object identifiers (DOIs) and can be cited in other articles and in grant applications. Because the peer review process can take many months, publishing preprints prior to or during peer review enables other scientists to immediately learn from and build on your work. Importantly, NIH allows applicants to include preprint citations in their biosketches. In most cases, we publish preprints on medRxiv<sup>4</sup>.

## 4.9 Publish data (when possible) and replication scripts

Publishing data and replication scripts allows other scientists to reproduce your work and to build upon it. We typically publish data on Open Science Framework<sup>5</sup>, share links to Github<sup>6</sup> repositories, and archive code on Zenodo<sup>7</sup>.

---

<sup>3</sup><https://www.equator-network.org/about-us/what-is-a-reporting-guideline/>

<sup>4</sup><https://www.medrxiv.org/>

<sup>5</sup>[osf.io](https://osf.io)

<sup>6</sup>[github.com](https://github.com)

<sup>7</sup>[zenodo.org](https://zenodo.org)

# 5 Code repositories

Adapted by UCD-SeRG team from original by Kunal Mishra, Jade Benjamin-Chung, and Stephanie Djajadi<sup>1</sup>

Each study has at least one code repository that typically holds R code, shell scripts with Unix code, and research outputs (results .RDS files, tables, figures). Repositories may also include datasets. This chapter outlines how to organize these files. Adhering to a standard format makes it easier for us to efficiently collaborate across projects.

## 5.1 Project Structure

We recommend the following directory structure:

```
0-run-project.sh
0-config.R
1 - Data-Management/
    0-prep-data.sh
    1-prep-cdph-fluseas.R
    2a-prep-absentee.R
    2b-prep-absentee-weighted.R
    3a-prep-absentee-adj.R
    3b-prep-absentee-adj-weighted.R
2 - Analysis/
    0-run-analysis.sh
    1 - Absentee-Mean/
        1-absentee-mean-primary.R
        2-absentee-mean-negative-control.R
        3-absentee-mean-CDC.R
        4-absentee-mean-peakwk.R
        5-absentee-mean-cdph2.R
        6-absentee-mean-cdph3.R
    2 - Absentee-Positivity-Check/
    3 - Absentee-P1/
    4 - Absentee-P2/
3 - Figures/
    0-run-figures.sh
    ...
4 - Tables/
    0-run-tables.sh
    ...
5 - Results/
```

---

<sup>1</sup><https://jadebc.github.io/lab-manual/code-repositories.html>

```

1 - Absentee-Mean/
  1-absentee-mean-primary.RDS
  2-absentee-mean-negative-control.RDS
  3-absentee-mean-CDC.RDS
  4-absentee-mean-peawkw.RDS
  5-absentee-mean-cdph2.RDS
  6-absentee-mean-cdph3.RDS

  ...
.gitignore
.Rproj

```

For brevity, not every directory is “expanded”, but we can glean some important takeaways from what we *do* see.

## 5.2 .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 I’d 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.

## 5.3 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 which will be reference in multiple scripts (i.e. a `merged_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 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.

## 5.4 Order Files and Directories

This makes the jumble of alphabetized filenames much more coherent and places similar code and files next to one another. This also helps us understand how data flows from start to finish and allows us to easily map a script to its output (i.e. `2 - Analysis/1 - Absentee-Mean/1-absentee-mean-primary.R => 5 - Results/1 - Absentee-Mean/1-absentee-mean-primary.RDS`). 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.

## 5.5 Using Bash scripts to ensure reproducibility

Bash scripts are useful components of a reproducible workflow. At many of the directory levels (i.e. in 3 - `Analysis`), there is a bash script that runs each of the analysis scripts. This is exceptionally useful when data “upstream” changes – you simply run the bash script. See Chapter 10 for further details.

After running bash scripts, `.Rout` log files will be generated for each script that has been executed. It is important to check these files. Scripts may appear to have run correctly in the terminal, but checking the log files is the only way to ensure that everything has run completely.

# 6 R Coding

## 6.1 Lab Protocols for Code and Data

Just as wet labs have strict safety protocols to ensure reproducible results and prevent contamination, our computational lab has protocols for coding and data management. These protocols are not suggestions—they are essential practices that:

- **Ensure reproducibility:** Others (including your future self) can recreate your analysis
- **Prevent errors:** Systematic approaches reduce the risk of mistakes
- **Enable collaboration:** Consistent practices allow team members to work together efficiently
- **Maintain data integrity:** Proper handling prevents data corruption and loss
- **Support publication:** Well-documented, reproducible code is increasingly required for publication

**Violating these protocols can have serious consequences**, including invalid results, wasted time, inability to publish, and damage to scientific credibility. Treat coding and data management protocols with the same seriousness as you would safety protocols in a wet lab.

## 6.2 R Package Structure for All Projects

All R projects in our lab should be structured as R packages, even if they are primarily analysis projects and not intended for distribution on CRAN or Bioconductor. This standardized structure provides numerous benefits:

### 6.2.1 Why Use R Package Structure?

1. **Organized code:** Clear separation of functions (`R/`), documentation (`man/`), tests (`tests/`), data (`data/`), and vignettes/analyses
2. **Dependency management:** `DESCRIPTION` file explicitly declares all package dependencies and versions
3. **Automatic documentation:** `roxygen2` generates help files from inline comments
4. **Built-in testing:** `testthat` framework integrates seamlessly with package structure
5. **Code quality:** Tools like `devtools::check()` and `lintr` enforce best practices
6. **Reproducibility:** Package structure makes it easy to share and reproduce analyses
7. **Reusable functions:** Decompose complex analyses into well-documented, testable functions
8. **Version control:** Track changes to code, documentation, and data together

## 6.2.2 Basic Package Structure

```
myproject/
  DESCRIPTION           # Package metadata and dependencies
  NAMESPACE            # Auto-generated, don't edit manually
  R/
    analysis_functions.R
    data_prep.R
    plotting.R
  man/                 # Auto-generated documentation
  tests/
    testthat/          # Unit tests
  data/                # Processed data objects (.rda files)
  data-raw/            # Raw data and data processing scripts
  vignettes/
    intro.Rmd         # Main vignettes (shipped with package)
    tutorial.Rmd
    articles/          # Website-only articles (not shipped)
      advanced-topics.Rmd
      case-studies.Rmd
  inst/                # Additional files to include in package
  extdata/             # External data files
  analyses/            # Analyses using restricted data (see below)
.Rproj                # RStudio project file
```

## 6.2.3 Where to Place Analysis Files

### 6.2.3.1 Vignettes vs Articles

**Vignettes** (`vignettes/*.Rmd`): - Shipped with the package when installed - Accessible via `vignette()` and `browseVignettes()` in R - Displayed on CRAN - Built at package build time - Use for core package documentation and tutorials - Created with `usethis::use_vignette("name")`

**Articles** (`vignettes/articles/*.Rmd`): - Website-only (not shipped with the package) - Only appear on the pkgdown website - Not accessible via `vignette()` in R - Not displayed on CRAN - Use for supplementary content, blog posts, extended tutorials, or frequently updated material - Created with `usethis::use_article("name")` - Automatically added to `.Rbuildignore`

**When to use each:** - **Vignette:** Essential tutorials users need offline, core package workflows - **Article:** Supplementary material, case studies, advanced topics, blog-style content

### 6.2.3.2 Public Analyses (vignettes/)

Use `vignettes/` for analysis workbooks that: - Use publicly available data - Should be accessible to all package users - Are core to understanding the package

Use `vignettes/articles/` for: - Extended case studies - Blog-style posts - Supplementary analyses - Material that updates frequently

All vignettes and articles will be rendered by `pkgdown::build_site()` on your package website.

### 6.2.3.3 Analyses with Restricted Data (`inst/analyses/`)

For analyses that rely on **private, sensitive, or restricted data**, place `.qmd` or `.Rmd` files in `inst/analyses/`:

```
myproject/
  inst/
    analyses/
      01-confidential-data-analysis.qmd
      02-unpublished-results.qmd
      README.md # Document data access requirements
    extdata/
  vignettes/
    01-public-analysis.Rmd
    02-demo-with-simulated-data.Rmd
```

**Benefits of this approach:**

- Analyses with restricted data are included in version control alongside your code
- They're clearly separated from public documentation
- `inst/analyses/` is **excluded from `pkgdown` builds** and package documentation
- Collaborators with data access can still run these analyses
- You maintain a complete record of all project work

**Note on privacy:** Files in `inst/analyses/` are not inherently private—they will be visible if your repository is public. Use this folder for analyses that rely on restricted data that is stored separately, not for storing the restricted data itself. If you need to keep the analysis code private, use a private repository.

**Best practices for analyses with restricted data:**

1. **Document data requirements:** Include a `README.md` in `inst/analyses/` explaining:
  - What data is required
  - Where to obtain it (if permissible)
  - Data access restrictions
  - How to set up data paths

2. **Use relative paths carefully:** Structure your code so data paths can be configured:

```
# In inst/analyses/01-analysis.qmd
# Users should set this based on their local setup
data_dir <- Sys.getenv("MYPROJECT_DATA",
                      default = "~/restricted_data/myproject")
raw_data <- readr::read_csv(file.path(data_dir, "sensitive.csv"))
```

3. **Create public alternatives:** When possible, create companion vignettes in `vignettes/` using:

- Simulated data that mimics the structure
  - Publicly available datasets
  - Aggregated/de-identified summaries
4. **Add to .Rbuildignore:** Ensure `inst/analyses/` doesn't cause package checks to fail:

```
# In .Rbuildignore
^inst/analyses$
```

### 6.2.4 Keep Analysis Workbooks Tidy

**Decompose reusable functions** from your analysis notebooks into the `R/` directory. Your vignettes should:

- Be clean, readable narratives of your analysis
- Call well-documented functions from your package
- Focus on the “what” and “why” rather than implementation details
- Be reproducible by others with a single click (or with documented data access for private analyses)

**Example of what NOT to do** (all code in vignette):

```
# Bad: 100 lines of data manipulation in vignette
raw_data <- read_csv("data.csv")
# ... 100 lines of cleaning, transforming, reshaping ...
cleaned_data <- final_result
```

**Example of what TO do** (functions in `R/`, simple calls in vignette):

```
# Good: Clean vignette calling documented functions
raw_data <- read_csv("data.csv")
cleaned_data <- prep_study_data(raw_data) # Function in R/data_prep.R
```

## 6.3 Essential R Package Development Tools

The following tools are essential for R package development in our lab:

### 6.3.1 usethis: Package Setup and Management

`usethis` automates common package development tasks:

```
# Install usethis
install.packages("usethis")

# Create a new package
usethis::create_package("~/myproject")

# Add common components
```

```

usethis::use_mit_license()          # Add a license
usethis::use_git()                 # Initialize git
usethis::use_github()              # Connect to GitHub
usethis::use_testthat()            # Set up testing infrastructure
usethis::use_vignette("intro")     # Create a vignette (shipped with package)
usethis::use_article("case-study") # Create an article (website-only)
usethis::use_data_raw("dataset")   # Create data processing script
usethis::use_package("dplyr")      # Add a dependency
usethis::use_pipe()                # Import pipe operator

# Increment version
usethis::use_version()           # Increment package version

```

### 6.3.2 devtools: Development Workflow

`devtools` provides the core development workflow:

```

# Install devtools
install.packages("devtools")

# Load your package for interactive development
devtools::load_all()             # Like library(), but for development

# Documentation
devtools::document()             # Generate documentation from roxygen2

# Testing
devtools::test()                 # Run all tests
devtools::test_active_file()      # Run tests in current file

# Checking
devtools::check()                # R CMD check (comprehensive validation)
devtools::check_man()             # Check documentation only

# Dependencies
devtools::install_dev_deps()      # Install all development dependencies

# Building
devtools::build()                # Build package bundle
devtools::install()               # Install package locally

```

### 6.3.3 pkgdown: Package Websites

`pkgdown` builds beautiful documentation websites from your package:

```

# Install pkgdown
install.packages("pkgdown")

```

```
# Set up pkgdown
usethis::use_pkgdown()

# Build website locally
pkgdown::build_site()

# Preview in browser
pkgdown::build_site(preview = TRUE)

# Build components separately
pkgdown::build_reference()                      # Function reference
pkgdown::build_articles()                       # Vignettes
pkgdown::build_home()                           # Home page from README
```

Configure your pkgdown site with `_pkgdown.yml`:

```
url: https://ucd-serg.github.io/myproject

template:
  bootstrap: 5

reference:
  - title: "Data Preparation"
    desc: "Functions for preparing and cleaning data"
    contents:
      - prep_study_data
      - validate_data

    - title: "Analysis"
      desc: "Core analysis functions"
      contents:
        - run_primary_analysis
        - sensitivity_analysis

articles:
  - title: "Analysis Workflow"
    navbar: Analysis
    contents:
      - 01-data-preparation
      - 02-primary-analysis
      - 03-sensitivity-analysis
```

## 6.4 Complete Package Development Workflow

Here's the typical workflow for developing an R package in our lab:

### 6.4.1 1. Initial Setup

```
# Create package structure
usethis::create_package("~/myproject")

# Set up infrastructure
usethis::use_git()
usethis::use_github()
usethis::use_testthat()
usethis::use_pkgdown()
usethis::use_mit_license()
usethis::use_readme_rmd()
```

### 6.4.2 2. Add Dependencies

```
# Add packages your project depends on
usethis::use_package("dplyr")
usethis::use_package("ggplot2")
usethis::use_package("readr")

# Add packages only needed for development/testing
usethis::use_package("testthat", type = "Suggests")
```

### 6.4.3 3. Write Functions

Create functions in R/ directory with roxygen2 documentation:

```
'#' Prepare Study Data
'#
' #' Clean and prepare raw study data for analysis.
'#
' #' @param raw_data A data frame containing raw study data
' #' @param validate Logical; whether to run validation checks
'#
' #' @returns A cleaned data frame ready for analysis
'#
' #' @examples
' #' raw_data <- read_csv("data.csv")
' #' clean_data <- prep_study_data(raw_data)
'#
' #' @export
prep_study_data <- function(raw_data, validate = TRUE) {
  # Function implementation
}
```

#### 6.4.4 4. Document

```
# Generate documentation from roxygen2 comments
devtools::document()
```

#### 6.4.5 5. Test

Create tests in `tests/testthat/`:

```
# tests/testthat/test-data_prep.R
test_that("prep_study_data handles missing values", {
  raw_data <- data.frame(x = c(1, NA, 3))
  result <- prep_study_data(raw_data)
  expect_false(anyNA(result$x))
})
```

Run tests:

```
devtools::test()
```

#### 6.4.6 6. Check

```
# Comprehensive package check
devtools::check()
```

Fix any warnings or errors before proceeding.

#### 6.4.7 7. Build Documentation Site

```
pkgdown::build_site()
```

#### 6.4.8 8. Share and Publish

```
# Push to GitHub
# The pkgdown site can be automatically deployed to GitHub Pages
# using GitHub Actions
```

### 6.5 Code Style Guidelines

Follow these code style guidelines for all R code:

### 6.5.1 General Principles

- Follow **tidyverse style guide**: <https://style.tidyverse.org>
- Use **native pipe**: `|>` not `%>%` (available in R  $\geq 4.1.0$ )
- **Naming**: Use `snake_case` for functions and variables; acronyms may be uppercase (e.g., `prep_IDs_data`)
- **Write tidy code**: Keep code clean, readable, and well-organized

### 6.5.2 File Organization

Just as your data “flows” through your project, code should flow naturally through files in the package structure:

1. **In R/ files**: Put reusable functions with roxygen2 documentation
2. **In vignettes**: Put narrative analysis with calls to those functions
3. **In data-raw/**: Put scripts that create data objects saved to `data/`
4. **In inst/analyses/**: Put analyses using restricted data

### 6.5.3 Function Structure and Documentation

Every function should follow this pattern:

```
#' Short Title (One Line)
#'
#' Longer description providing details about what the function does,
#' when to use it, and important considerations.
#'
#' @param param1 Description of first parameter, including type and constraints
#' @param param2 Description of second parameter
#'
#' @returns Description of return value, including type and structure
#'
#' @examples
#' # Example usage
#' result <- my_function(param1 = "value", param2 = 10)
#'
#' @export
my_function <- function(param1, param2) {
  # Implementation
}
```

### 6.5.4 Comments

Use comments to explain *why*, not *what*:

```
# Good: Explains reasoning
# Use log scale because distribution is highly skewed
ggplot(data, aes(x = log10(income))) + geom_histogram()

# Bad: States the obvious
# Create a histogram
ggplot(data, aes(x = income)) + geom_histogram()
```

**File headers** (for scripts in `data-raw/` or `inst/analyses/`):

```
#####
# @Project - Myproject
# @Description - Process raw survey data and create analysis dataset
#####
```

**Sections and subsections** in longer files:

```
# Data Loading -----
## Read survey data -----
## Read lab results -----
# Data Cleaning -----
```

### 6.5.5 Messaging and User Communication

Use `cli` package functions for all user-facing messages in package functions:

```
# Good
cli::cli_inform("Analysis complete")
cli::cli_warn("Missing data detected")
cli::cli_abort("Invalid input: {x}")

# Bad - don't use these in package code
message("Analysis complete")
warning("Missing data detected")
stop("Invalid input")
```

### 6.5.6 Package Code Practices

- **No `library()` in package code:** Use `::` notation or declare in `DESCRIPTION` Imports
- **Document all exports:** Use roxygen2 (`@title`, `@description`, `@param`, `@returns`, `@examples`)
- **Avoid code duplication:** Extract repeated logic into helper functions

### 6.5.7 Line Breaks and Formatting

For readability, use line breaks in function calls and pipelines:

```
# Good: Named arguments on separate lines
mean_Y <- calc_fluseas_mean(
  data = flu_data,
  yname = "maari_yn",
  silent = FALSE
)

# Good: Pipeline with clear flow
cleaned_data <- raw_data |>
  filter(!is.na(outcome)) |>
  mutate(
    log_value = log10(value),
    category = case_when(
      value < 10 ~ "low",
      value < 100 ~ "medium",
      TRUE ~ "high"
    )
  ) |>
  group_by(category) |>
  summarize(
    n = n(),
    mean = mean(log_value)
  )
```

For complex `ggplot` calls:

```
ggplot(data, aes(x = year, y = rate, group = group)) +
  geom_point(
    aes(col = group, shape = group),
    position = position_dodge(width = 0.2),
    size = 2.5
  ) +
  geom_errorbar(
    aes(ymin = lb, ymax = ub, col = group),
    position = position_dodge(width = 0.2),
    width = 0.2
  ) +
  scale_y_continuous(
    limits = c(0, 100),
    breaks = seq(0, 100, 25)
  ) +
  scale_color_manual(values = colors) +
  theme_minimal() +
  labs(
    title = "Title",
    x = "Year",
```

```
y = "Rate (%)"
)
```

### 6.5.8 Tidyverse Replacements

Use modern tidyverse/alternatives for base R functions:

```
# Data structures
tibble::tibble()           # instead of data.frame()
tibble::tribble()          # instead of manual data.frame creation

# I/O
readr::read_csv()          # instead of read.csv()
readr::write_csv()          # instead of write.csv()
readr::read_rds()           # instead of readRDS()
readr::write_rds()          # instead of saveRDS()

# Data manipulation
dplyr::bind_rows()         # instead of rbind()
dplyr::bind_cols()          # instead of cbind()

# String operations
stringr::str_which()        # instead of grep()
stringr::str_replace()       # instead of gsub()

# Session info
sessioninfo::session_info() # instead of sessionInfo()
```

### 6.5.9 The here Package

The `here` package helps manage file paths in projects:

```
library(here)

# Automatically finds project root and builds paths
data <- readr::read_csv(here("data-raw", "survey.csv"))
saveRDS(results, here("inst", "analyses", "results.rds"))
```

This works regardless of where collaborators clone the repository.

### 6.5.10 Object Naming

Use descriptive names that are both expressive and explicit:

```
# Good
vaccination_coverage_2017_18
absentee_flu_residuals

# Less good
vaxcov_1718
flu_res
```

Prefer nouns for objects and verbs for functions:

```
# Good
clean_data <- prep_study_data(raw_data) # verb for function, noun for object

# Less clear
data <- process(input)
```

## 6.6 Testing Requirements

**ALWAYS establish tests BEFORE modifying functions.** This ensures changes preserve existing behavior and new behavior is correctly validated.

### 6.6.1 When to Use Snapshot Tests

Use snapshot tests (`expect_snapshot()`, `expect_snapshot_value()`) when:

- Testing complex data structures (data frames, lists, model outputs)
- Validating statistical results where exact values may vary slightly
- Output format stability is important

```
test_that("prep_study_data produces expected structure", {
  result <- prep_study_data(raw_data)
  expect_snapshot_value(result, style = "serialize")
})
```

### 6.6.2 When to Use Explicit Value Tests

Use explicit tests (`expect_equal()`, `expect_identical()`) when:

- Testing simple scalar outputs
- Validating specific numeric thresholds
- Testing Boolean returns or categorical outputs

```
test_that("calculate_mean returns correct value", {
  expect_equal(calculate_mean(c(1, 2, 3)), 2)
  expect_equal(calculate_ratio(3, 7), 0.4285714, tolerance = 1e-6)
})
```

### 6.6.3 Testing Best Practices

- **Seed randomness:** Use `withr::local_seed()` for reproducible tests
- **Use small test cases:** Keep tests fast
- **Test edge cases:** Missing values, empty inputs, boundary conditions
- **Test errors:** Verify functions fail appropriately with invalid input

```
test_that("prep_study_data handles edge cases", {
  # Empty input
  expect_error(prep_study_data(data.frame()))

  # Missing required columns
  expect_error(prep_study_data(data.frame(x = 1)))

  # Valid input with missing values
  result <- prep_study_data(data.frame(id = 1:3, value = c(1, NA, 3)))
  expect_true(all(!is.na(result$value)))
})
```

## 6.7 Iterative Operations

When applying analyses with different variations (outcomes, exposures, subgroups), use functional programming approaches:

### 6.7.1 `lapply()` and `sapply()`

```
# Apply function to each element
results <- lapply(outcomes, function(y) {
  run_analysis(data, outcome = y)
})

# Simplify to vector if possible
summary_stats <- sapply(data_list, mean)
```

### 6.7.2 `purrr::map()` Family

The `purrr` package provides type-stable alternatives:

```
library(purrr)

# Always returns a list
results <- map(outcomes, ~ run_analysis(data, outcome = .x))

# Type-specific variants
means <- map_dbl(data_list, mean)          # Returns numeric vector
models <- map(splits, ~ lm(y ~ x, data = .x)) # Returns list of models
```

### 6.7.3 purrr::pmap() for Multiple Arguments

When iterating over multiple parameter lists:

```
params <- tibble(
  outcome = c("outcome1", "outcome2", "outcome3"),
  exposure = c("exp1", "exp2", "exp3"),
  covariate_set = list(c("age", "sex"), c("age"), c("age", "sex", "bmi")))
)

results <- pmap(params, function(outcome, exposure, covariate_set) {
  run_analysis(
    data = study_data,
    outcome = outcome,
    exposure = exposure,
    covariates = covariate_set
  )
})
```

### 6.7.4 Parallel Processing

For computationally intensive work, use `future` and `furrr`:

```
library(future)
library(furrr)

# Set up parallel processing
plan(multisession, workers = availableCores() - 1)

# Parallel version of map()
results <- future_map(large_list, time-consuming_function, .progress = TRUE)
```

## 6.8 Reading and Saving Data

### 6.8.1 RDS Files (Preferred)

Use RDS format for R objects:

```
# Save single object
readr::write_rds(analysis_results, here("results", "analysis.rds"))

# Read back
results <- readr::read_rds(here("results", "analysis.rds"))
```

**Avoid .RData files** because: - You can't control object names when loading - Can't load individual objects - Creates confusion in older code

## 6.8.2 CSV Files

For tabular data that may be shared with non-R users:

```
# Write
readr::write_csv(data, here("data-raw", "clean_data.csv"))

# Read
data <- readr::read_csv(here("data-raw", "clean_data.csv"))

# For very large files, use data.table
data.table::fwrite(large_data, "big_file.csv")
data <- data.table::fread("big_file.csv")
```

## 6.9 Version Control and Collaboration

### 6.9.1 Version Numbers

Follow semantic versioning (MAJOR.MINOR.PATCH):

- Development versions: 0.0.0.9000, 0.0.0.9001, etc.
- First release: 0.1.0
- Bug fixes: increment PATCH (e.g., 0.1.0 → 0.1.1)
- New features: increment MINOR (e.g., 0.1.1 → 0.2.0)
- Breaking changes: increment MAJOR (e.g., 0.2.0 → 1.0.0)

```
# Increment version
usethis::use_version()
```

### 6.9.2 NEWS File

Document all user-facing changes in NEWS.md:

```
# myproject 0.2.0

## New Features

* Added `prep_study_data()` function for data preparation
* Implemented sensitivity analysis in `sensitivity_analysis()`

## Bug Fixes

* Fixed handling of missing values in `calculate_mean()`

## Breaking Changes

* Renamed `old_function()` to `new_function()`
```

### 6.9.3 Git Workflow

```
# Set up Git and GitHub
usethis::use_git()
usethis::use_github()

# Before committing changes
devtools::document()          # Update documentation
devtools::test()               # Run tests
devtools::check()              # Full package check
```

### 6.9.4 Managing Dependencies with renv

**Use renv to tightly control dependency versions during development.** The `renv` package creates a project-specific library that isolates your package dependencies, ensuring reproducibility across different machines and over time.

#### 6.9.4.1 When to Use renv

- **During active development:** Use `renv` to lock down specific versions of dependencies
- **For collaborative projects:** Ensure all team members use the same package versions
- **For long-term reproducibility:** Preserve the exact dependency environment

#### 6.9.4.2 Setting Up renv

```
# Install renv
install.packages("renv")

# Initialize renv in your package
renv::init()

# This creates:
# - renv/ directory with package library
# - renv.lock file with exact package versions
# - .Rprofile to activate renv on startup
```

#### 6.9.4.3 Using renv in Package Development

```
# Install/update packages (renv tracks changes)
install.packages("dplyr")
renv::install("tidyverse@1.3.0") # Install specific version

# Save the current state
```

```
renv::snapshot()

# Restore packages from lockfile
renv::restore()

# Update packages
renv::update()
```

#### 6.9.4.4 Preparing for CRAN Release

For packages ready for CRAN submission, disable `renv` and test with latest dependency versions:

1. Remove `renv` before CRAN submission:

```
renv::deactivate()
# Delete renv/ directory and renv.lock
# Remove .Rprofile modifications
```

2. Test with latest versions:

```
# Install latest versions of all dependencies
devtools::install_dev_deps()

# Run comprehensive checks
devtools::check()
```

3. Why disable `renv` for CRAN: CRAN tests packages against the latest versions of dependencies. Your package should work with current CRAN versions, not pinned older versions.

Workflow summary:

- Development phase: Use `renv` for stability and reproducibility
- Pre-CRAN release: Disable `renv`, test with latest dependency versions
- After CRAN release: Can re-enable `renv` for continued development if desired

## 6.10 Quality Assurance Checklist

Before submitting a pull request or finalizing analysis, verify:

- All functions have complete roxygen2 documentation
- All functions have corresponding tests
- `devtools::document()` has been run
- `devtools::test()` passes with no failures
- `devtools::check()` passes with no errors, warnings, or notes
- `lintr::lint_package()` shows no issues (or only acceptable ones)
- `spelling::spell_check_package()` passes
- Version number has been incremented
- `NEWS.md` has been updated with changes
- `README.Rmd` has been updated (if needed) and `README.md` regenerated

- `pkgdown::build_site()` builds successfully
- All changes committed and pushed to GitHub

## 6.11 Continuous Integration

Set up automated checks using GitHub Actions:

```
# Add standard R package CI workflows
usethis::use_github_action("check-standard")          # R CMD check
usethis::use_github_action("test-coverage")            # Code coverage
usethis::use_github_action("pkgdown")                  # Deploy website
```

These workflows automatically:

- Run R CMD check on multiple platforms (Linux, macOS, Windows)
- Calculate test coverage
- Build and deploy your pkgdown website
- Ensure code quality before merging

## 6.12 Automated Code Styling

### 6.12.1 RStudio Built-in Formatting

Use RStudio's built-in autoformatter (keyboard shortcut: **CMD-Shift-A** or **Ctrl-Shift-A**) to quickly format highlighted code.

### 6.12.2 styler Package

For automated styling of entire projects:

```
# Install styler
install.packages("styler")

# Style all files in R/ directory
styler::style_dir("R/")

# Style entire package
styler::style_pkg()

# Note: styler modifies files in-place
# Always use with version control so you can review changes
```

### 6.12.3 lintr Package

For checking code style without modifying files:

```
# Install lintr
install.packages("lintr")

# Lint the entire package
lintr::lint_package()

# Lint a specific file
lintr::lint("R/my_function.R")
```

The linter checks for: - Unused variables - Improper whitespace - Line length issues - Style guide violations

You can customize linting rules by creating a `.lintr` file in your project root.

## 6.13 Additional Resources

- R Packages book<sup>1</sup> by Hadley Wickham and Jenny Bryan
- Tidyverse style guide<sup>2</sup>
- usethis documentation<sup>3</sup>
- devtools documentation<sup>4</sup>
- pkgdown documentation<sup>5</sup>
- testthat documentation<sup>6</sup>

---

<sup>1</sup><https://r-pkgs.org/>

<sup>2</sup><https://style.tidyverse.org/>

<sup>3</sup><https://usethis.r-lib.org/>

<sup>4</sup><https://devtools.r-lib.org/>

<sup>5</sup><https://pkgdown.r-lib.org/>

<sup>6</sup><https://testthat.r-lib.org/>

# 7 Big data

Adapted by UCD-SeRG team from original by Kunal Mishra and Jade Benjamin-Chung<sup>1</sup>

## 7.1 The `data.table` package

It may also be the case that you’re working with very large datasets. Generally I would define this as 10+ million rows. As is outlined in this document, the 3 main players in the data analysis space are Base R, `Tidyverse` (more specifically, `dplyr`), and `data.table`. For a majority of things, Base R is inferior to both `dplyr` and `data.table`, with concise but less clear syntax and less speed. `Dplyr` is architected for medium and smaller data, and while it’s very fast for everyday usage, it trades off maximum performance for ease of use and syntax compared to `data.table`. An overview of the `dplyr` vs `data.table` debate can be found in this stackoverflow post<sup>2</sup> and all 3 answers are worth a read.

You can also achieve a performance boost by running `dplyr` commands on `data.tables`, which I find to be the best of both worlds, given that a `data.table` is a special type of `data.frame` and fairly easy to convert with the `as.data.table()` function. The speedup is due to `dplyr`’s use of the `data.table` backend and in the future this coupling should become even more natural.

If you want to test whether using a certain coding approach increases speed, consider the `tic toc` package. Run `tic()` before a code chunk and `toc()` after to measure the amount of system time it takes to run the chunk. For example, you might use this to decide if you *really* need to switch a code chunk from `dplyr` to `data.table`.

## 7.2 Using downsampled data

In our 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.

## 7.3 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

<sup>1</sup><https://jadebc.github.io/lab-manual/big-data.html>

<sup>2</sup><https://stackoverflow.com/questions/21435339/data-table-vs-dplyr-can-one-do-something-well-the-other-cant-or-does-poorly/27840349#27840349>

- **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.

# 8 Data masking

Adapted by UCD-SeRG team from original by Anna Nguyen, Jade Benjamin-Chung, and Gabby Barratt Heitmann<sup>1</sup>

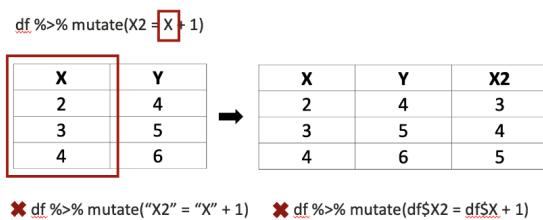
For information about UC Davis computing resources for data-intensive work, see Chapter 14.

## 8.1 General Overview

This chapter covers data masking, a unique process in R in which columns are treated as distinct objects within their dataframe's environment. In our lab, data masking most frequently comes up when writing wrapper functions where arguments to indicate column names are supplied as strings. We often do this when we repeat the same code on multiple columns, and want to apply a function to a vector of strings that correspond to column names in a dataframe. For example, we might want to clean multiple columns using the same function or estimate the same model under different feature sets. Here, we try to break down what data masking is, why this error comes up, and common approaches to solve this problem.

### 8.1.1 What is Data Masking?

Within certain tidyverse operations, columns are called as if they were variables. For example, while running `df %>% mutate(X = ...)` R recognizes that `X` specifically references a column in `df` without explicitly stating its membership `df %>% mutate(df$X = ...)` or calling the column name as a string `df %>% mutate("X" = ...)`.



However, this behavior may introduce errors when we attempt to incorporate variables from the global environment within these tidyverse pipelines. In the example above, `column_name = "X"` followed by `df %>% mutate(X2 = column_name + 1)` would yield an error, since `column_name` is not a column in `df` and the variable `column_name` is not defined within the environment of `df`.

<sup>1</sup><https://jadebc.github.io/lab-manual/data-masking.html>

### 8.1.2 Using tidy evaluation for data masking

In dplyr-based R programming, we make use of tidy evaluation. This allows us to avoid using base R syntax to reference specific columns in a data frame. By leveraging Tidy evaluation-based data masking, we can employ long pipes with several dplyr verbs to manipulate our data using stand-alone variables that store column names as strings.

For example, consider a data frame “df” that contains a column called “heavyrain” that we want to manipulate. Suppose we wanted to convert the values of “heavyrain” into a factor.

Using base R, which does not mask data, heavyrain must have quotes to be treated as a data-variable:

```
df[["outcome"]] = as.factor(df[["heavyrain"]])
```

In a dplyr pipe, heavyrain is being masked using tidy evaluation and will be correctly interpreted as a column because it is recognized as a data-variable: `df %>% mutate(outcome = as.factor(heavyrain))`

With modified data masking, heavyrain is a string that is coerced into being recognized as a data-variable:

```
var_name = "heavyrain"
df %>% mutate(outcome = as.factor (!!sym(var_name)))
```

While cleaner and often more convenient, the data frame that var\_name is in is now “masked” and we refer to the vectors in the dataframe (data-variables) as though it is an object of its own (an environmental-variable). This is why we can just say the variable’s name in the context of a pipe – we treat it as though it’s an object defined in our environment. Within normal scripts, this is usually fine, because the data frame is “held on to” in the pipe. However, it can cause some programming hurdles when writing functions that take strings of variable/column names as arguments. In the next section, we briefly describe how to troubleshoot common errors in data masking, as relevant to our lab’s work.

## 8.2 Technical Overview

This section covers the R functions and tools that we often use in the context of data masking, focusing on the bang bang operator (!! with symbol coercion (`sym()`) and the Walrus operator (`:=`).

The combined use of !! and `sym()` allows us to use strings, rather than data-variables, to reference column names within dplyr. Together, `!!sym("column_name")` forces dplyr to recognize “column\_name” as a data-variable prior to evaluating the rest of the expression, enabling the ability to perform calculations on the column while referring to it as a string. `sym()` is a function that turns strings into symbols. In the context of a dplyr pipe, these symbols are interpreted as data-variables. The !! (bang bang) operator tells dplyr to evaluate the `sym()` expression first, e.g. to unquote its expression (e.g. “column\_name”) and evaluate it as a pre-existing object, first. This is helpful because often we use `sym("column_name")` within a larger expression, and dplyr might evaluate other elements of the expression first without !!, causing errors.

When we want to create a new column (via `mutate` or `summarize`), the Walrus operator (`:=`) allows us to specify the new column's name using a string. For example, while `df %>% mutate("new_column" = values)` would yield an error, `df %>% mutate("new_column" := values)` will correctly create a new column called "new\_column". If we want to use a variable representing a string, we can use `!!` to force the variable to be evaluated before using `:=` to assign the value of the new column.

```
col_name = "new_column"
df %>% mutate (!!col_name := values)
```

### 8.2.1 Example

Suppose we want to write a function "generate\_descriptive\_table" to summarize how the prevalence of "outcome" varies under different levels of a "risk\_factor" in a data frame "df"

We can start by writing the function shell:

```
generate_descriptive_table <- function (df, outcome, rf) {
  outcome_dist_by_rf <- ....
  return(outcome_dist_by_rf)
}
```

Next, we can filter the data frame for only rows in which "rf" and "outcome" are not missing. We can use `!!` and `sym()` within `filter` to evaluate the strings stored in "rf" and "outcome". Note that defining `!!sym(outcome)` or `!!sym(outcome)` in variables *outside of the dplyr pipeline* will *not* work.

```
generate_descriptive_table <- function (df, outcome, rf,) {
  outcome_dist_by_rf <- df %>%
    filter(!is.na(!!sym(outcome)), !is.na(!!sym(rf))) %>%
    ....
  return(outcome_dist_by_rf)
}
```

Similarly, we use `!!` and `sym()` in `group_by` to evaluate column name, stored as a string in the argument "rf"

```
generate_descriptive_table <- function (df, outcome, rf,) {
  outcome_dist_by_rf <- df %>%
    filter(!is.na(!!sym(outcome)), !is.na(!!sym(rf))) %>%
    ....
  return(outcome_dist_by_rf)
}
```

Finally, we can use the walrus operator, `!!` and `sym()` with "summarize" to create a new column that takes the mean of the column referenced in "rf". We also use "glue" or "paste" to give the new column an informative name that includes the "outcome" it describes.

```
generate_descriptive_table <- function (df, outcome, rf,) {
  outcome_dist_by_rf <- df %>%
    filter(!is.na (!!sym(outcome)), !is.na (!!sym(rf))) %>%
    group_by (!!sym(rf)) %>%
    summarize (!!(glue:::glue("{outcome}_prev")) := mean (!!sym(outcome)))
  return(outcome_dist_by_rf)
}
```

OR

```
generate_descriptive_table <- function (df, outcome, rf,) {
  outcome_dist_by_rf <- df %>%
    filter(!is.na (!!sym(outcome)), !is.na (!!sym(rf))) %>%
    group_by (!!sym(rf)) %>%
    summarize (!!(paste0(outcome, "_prev")) := mean (!!sym(outcome)))
  return(outcome_dist_by_rf)
}
```

OR

```
generate_descriptive_table <- function (df, outcome, rf,) {
  new_column_name = paste0(outcome, "_prev")
  outcome_dist_by_rf <- df %>%
    filter(!is.na (!!sym(outcome)), !is.na (!!sym(rf))) %>%
    group_by (!!sym(rf)) %>%
    summarize (!!new_column_name) := mean (!!sym(outcome)))
  return(outcome_dist_by_rf)
}
```

# 9 Github

Adapted by UCD-SeRG team from original by Stephanie Djajadi and Nolan Pokpongkiat<sup>1</sup>

## 9.1 Basics

- A detailed tutorial of Git can be found here on the CS61B website<sup>2</sup>.
- If you are already familiar with Git, you can reference the summary at the end of Section B<sup>3</sup>.
- If you have made a mistake in Git, you can refer to this article<sup>4</sup> to undo, fix, or remove commits in git.

## 9.2 Github Desktop

While knowing how to use Git on the command line will always be useful since the full power of Git and its customizations and flexibility is designed for use with the command line, Github also provides Github Desktop<sup>5</sup> as a graphical interface to do basic git commands; you can do all of the basic functions of Git using this desktop app. Feel free to use this as an alternative to Git on the command line if you prefer.

## 9.3 Git Branching

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<sup>6</sup>. You can also find instructions on

---

<sup>1</sup><https://jadebc.github.io/lab-manual/github.html>

<sup>2</sup><https://sp19.datastructur.es/materials/guides/using-git#/b-local-repositories-narrative-introduction>

<sup>3</sup><https://sp19.datastructur.es/materials/guides/using-git#/b-local-repositories-narrative-introduction>

<sup>4</sup><https://sethrobertson.github.io/GitFixUm/fixup.html>

<sup>5</sup><https://desktop.github.com/>

<sup>6</sup><https://sp19.datastructur.es/materials/guides/using-git#/e-git-branching-advanced-git-optional>

how to handle merge conflicts when joining branches together.

## 9.4 Example Workflow

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

Command	Description
SETUP: FIRST TIME ONLY: <code>git clone &lt;url&gt;</code> <code>&lt;directory_name&gt;</code>	Clone the repo. This copies of all the project files in its current state on Github to your local computer.
1. <code>git pull origin master</code>	update the state of your files to match the most current version on GitHub
2. <code>git checkout -b</code> <code>&lt;new_branch_name&gt;</code>	create new branch that you'll be working on and go to it
3. Make some file changes	work on your feature/implementation
4. <code>git add -p</code>	add changes to stage for commit, going through changes line by line
5. <code>git commit -m &lt;commit message&gt;</code>	commit files with a message
6. <code>git push -u origin</code> <code>&lt;branch_name&gt;</code>	push branch to remote and set to track (-u only needed if this is first push)
7. Repeat step 4-5.	work and commit often
8. <code>git push</code>	push work to remote branch for others to view
9. Follow the link given from the <code>git push</code> command to submit a pull request (PR) on GitHub online <sup>7</sup>	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. <code>git fetch --all</code> <code>--prune</code>	clean up your local git by untracking deleted remote branches

Other helpful commands are listed below.

## 9.5 Commonly Used Git Commands

Command	Description
<code>git clone &lt;url&gt;</code> <code>&lt;directory_name&gt;</code>	clone a repository, only needs to be done the first time
<code>git pull origin master</code>	pull from <code>master</code> before making any changes
<code>git branch</code>	check what branch you are on

<sup>7</sup><https://help.github.com/en/github/collaborating-with-issues-and-pull-requests/creating-a-pull-request#creating-the-pull-request>

Command	Description
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 add -p	adds changes to commit, showing you changes one by one
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/<branch_name> (use if you forgot -u on first push)
git push origin <branch_name>	push work to branch
git checkout <branch_name>	switch to branch and merge changes from master into <branch_name> (two commands)
git merge master	switch to branch and merge changes from master into <branch_name> (one command)
git merge <branch_name> master	
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)
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

## 9.6 How often should I commit?

It is good practice to commit every 15 minutes, or every time you make a significant change. It is better to commit more rather than less.

## 9.7 What should be pushed to Github?

Never push .Rout 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 named and date timestamped.

There is a standardized `.gitignore` for R which you can download<sup>8</sup> and add to your project. This ensures you're not committing log files or things that would otherwise best be left

---

<sup>8</sup><https://github.com/github/gitignore/blob/master/R.gitignore>

## *9 Github*

ignored to GitHub. This is a great discussion of project-oriented workflows<sup>9</sup>, extolling the virtues of a self-contained, portable projects, for your reference.

---

<sup>9</sup><https://www.tidyverse.org/articles/2017/12/workflow-vs-script/>

# 10 Unix

Adapted by UCD-SeRG team from original by Stephanie Djajadi, Kunal Mishra, Anna Nguyen, and Jade Benjamin-Chung<sup>1</sup>

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 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; it is recommended that Windows users use Git Bash, not Windows PowerShell.

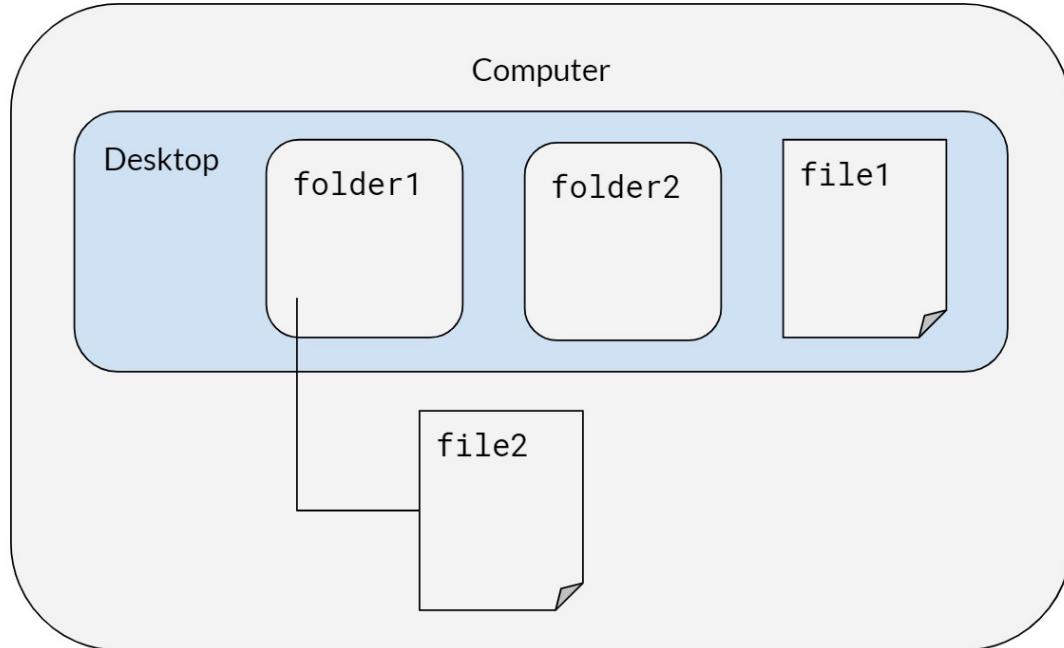


Figure 10.1: Here is our example desktop.

<sup>1</sup><https://jadebc.github.io/lab-manual/unix.html>

## 10.2 Syntax for both Mac/Windows

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

Command	Description
<code>cd desktop/folder1</code>	Change directory to <code>folder1</code>
<code>pwd</code>	Print working directory
<code>ls</code>	List files in the directory
<code>cp "file2" "newfile2"</code>	Copy file (remember to include file extensions when typing in file names like <code>.pdf</code> or <code>.R</code> )
<code>mv "newfile2" "file3"</code>	Rename <code>newfile2</code> to <code>file3</code>
<code>cd ..</code>	Go to parent of the working directory (in this case, <code>desktop</code> )
<code>mv "file1" folder2</code>	Move <code>file1</code> to <code>folder2</code>
<code>mkdir folder3</code>	Make a new folder in <code>folder2</code>
<code>rm &lt;filename&gt;</code>	Remove files
<code>rm -rf folder3</code>	Remove directories ( <code>-r</code> will attempt to remove the directory recursively, <code>-rf</code> will force removal of the directory)
<code>clear</code>	Clear terminal screen of all previous commands

```

MINGW64:/c/Users/Stephanie Djajadi/desktop/folder2
Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~
$ cd ~/desktop

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop
$ pwd
/c/Users/Stephanie Djajadi/desktop

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop
$ ls
desktop.ini  file1.txt  folder1/  folder2/

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop
$ cd folder1

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder1
$ ls
file2.txt

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder1
$ cp file2.txt newfile2.txt

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder1
$ ls
file2.txt  newfile2.txt

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder1
$ mv newfile2.txt file3.txt

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder1
$ ls
file2.txt  file3.txt

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder1
$ cd ..

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop
$ ls
desktop.ini  file1.txt  folder1/  folder2/

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop
$ mv file1.txt folder2

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop
$ cd folder2

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder2
$ ls
file1.txt

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder2
$ mkdir folder3

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder2
$ ls
file1.txt  folder3/

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder2
$ rm file1.txt

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder2
$ rm -rf folder3

Stephanie Djajadi@DESKTOP-L0H5VOO MINGW64 ~/desktop/folder2
$ ls

```

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

## 10.3 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

## 10.4 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<sup>2</sup>. I strongly recommend completing this option.

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')"

### 10.4.1 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 /.

## 10.5 Checking tasks and killing jobs

---

<sup>2</sup><https://www.howtogeek.com/118594/how-to-edit-your-system-path-for-easy-command-line-access/>

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

## 10.6 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<sup>3</sup> has useful tips on using `tmux`.

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

# run your job of interest
```

<sup>3</sup><https://medium.com/@jeongwhanchoi/install-tmux-on-osx-and-basics-commands-for-beginners-be22520fd95e>

```
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 -t session_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. A utility file is included in this repository, called `runFileSaveLogs`, and is used by the example bash scripts to... run files and save the generated logs. It is an awesome utility and one I definitely recommend using. Before using `runFileSaveLogs`, it is necessary to put the file in the home working directory. For help and documentation, you can use the command `./runFileSaveLogs -h`. See example code and example usage for `runFileSaveLogs` below.

### 10.6.1 Example code for `runfileSaveLogs`

```
#!/usr/bin/env python3
# Type "./runFileSaveLogs -h" for help

import os
import sys
import argparse
import getpass
import datetime
import shutil
import glob
import pathlib

# Setting working directory to this script's current directory
os.chdir(os.path.dirname(os.path.abspath(__file__)))

# Setting up argument parser
parser = argparse.ArgumentParser(description='Runs the argument R script(s) - in parallel if multiple are specified')  
if __name__ == '__main__':
    args = parser.parse_args()
    # Check if user has provided a valid R script name
    if len(args.r_script) > 0 and os.path.isfile(args.r_script):
        # Run the R script
        run_file(args.r_script, args.parallel)
    else:
        print("Error: No R script provided or file does not exist")
```

```

# Function ensuring that the file is valid
def is_valid_file(parser, arg):
    if not os.path.exists(arg):
        parser.error("The file %s does not exist!" % arg)
    else:
        return arg

# Function ensuring that the directory is valid
def is_valid_directory(parser, arg):
    if not os.path.isdir(arg):
        parser.error("The specified path (%s) is not a directory!" % arg)
    else:
        return arg

# Additional arguments that can be added when running runFileSaveLogs
parser.add_argument('-p', '--parallel', action='store_true', help="Runs the argument R scri
parser.add_argument("-i", "--identifier", help="Adds an identifier to the directory name wh
parser.add_argument('filenames', nargs='+', type=lambda x: is_valid_file(parser, x))

args = parser.parse_args()
args_dict = vars(args)

print(args_dict)

# Run given R Scripts
for filename in args_dict["filenames"]:
    system_call = "R CMD BATCH" + " " + filename
    if args_dict["parallel"]:
        system_call = "nohup" + " " + system_call + " &"

    os.system(system_call)

# Create the directory (and any parents) of the log files
currentUser = getpass.getuser()
currentTime = datetime.datetime.now().strftime("%Y-%m-%d %H:%M:%S")
logDirPrefix = "/home/kaiserData/logs/" # Change to the directory where the logs should be
logDir = logDirPrefix + currentTime + "-" + currentUser

# If specified, adds the identifier to the filename of the log
if args.identifier is not None:
    logDir += "-" + args.identifier

logDir += "/"

pathlib.Path(logDir).mkdir(parents=True, exist_ok=True)

# Find and move all logs to this new directory
currentLogPaths = glob.glob('./.*.Rout')

for currentLogPath in currentLogPaths:

```

```
filename = currentLogPath.split("/")[-1]
shutil.move(currentLogPath, logDir + filename)
```

### 10.6.2 Example usage for runfileSaveLogs

This example bash script runs files and generates logs for five scripts in the `kaiserflu/3-figures` folder. Note that the `-i` flag is used as an identifier to add `figures` to the filename of each log.

```
#!/bin/bash

# Copy utility run script into this folder for concision in call
cp ~/kaiserflu/runFileSaveLogs ~/kaiserflu/3-figures/

# Run folder scripts and produce output
cd ~/kaiserflu/3-figures/
./runFileSaveLogs -i "figures" \
fig-mean-season-age.R \
fig-monthly-rate.R \
fig-point-estimates-combined.R \
fig-point-estimates.R \
fig-weekly-rate.R

# Remove copied utility run script
rm runFileSaveLogs
```

# 11 Reproducible Environments

Adapted by UCD-SeRG team from original by Anna Nguyen<sup>1</sup>

## 11.1 Package Version Control with `renv`

### 11.1.1 Introduction

Replicable code should produce the same results, regardless of when or where it's run. However, our analyses often leverage open-source R packages that are developed by other teams. These packages continue to be developed after research projects are completed, which may include changes to analysis functions that could impact how code runs for both other team members and external replicators.

For example, suppose we had used a function that took in one argument, such that our code contained `example_function(arg_a = "a")`. A few months after we publish our code, the package developers update the function to take in another mandatory argument `arg_b`. If someone runs our code, but has the most recent version of the package, they'll receive an error message that the argument `arg_b` is missing and will not be able to full reproduce our results.

To ensure that the right functions are used in replication efforts, it is important for us to keep track of package versions used in each project.

`renv` can be used to promote reproducible environments within R projects. `renv` creates individual package libraries for each project instead of having all projects, which may use different versions of the same package, share the same package library. However, for projects that use many packages, this process can be memory intensive and increase the time needed for a new users to start running code.

In this lab manual chapter, we provide a quick tutorial for integrating `renv` into research workflows. For more detailed instructions, please refer to the `renv` package vignette.

### 11.1.2 Implementing `renv` in projects

Ideally, `renv` should be initiated at the start of projects and updated continuously when new packages are introduced in the codebase. However, this process can be initiated at any point in a project

To add `renv` to your workflow, follow these steps:

1. Install the `renv` package by running `install.packages("renv")`
2. Create an RProject file and ensure that your working directory is set to the correct folder

---

<sup>1</sup><https://jadebc.github.io/lab-manual/reproducible-environments.html>

3. In the R console, run `renv::init()` to initialize renv in your R Project
4. This will create the following files: `renv.lock`, `.Rprofile`, `renv/settings.json` and `renv/activate.R`. Commit and push these files to GitHub so that they're accessible to other users.
5. As you write code, update the project's R library by running `renv::snapshot()` in the R console
6. Add `renv::restore()` to the head of your config file, to make sure that all users that run your code are on the same package versions.

### 11.1.3 Using projects with renv

If you're starting to work on an ongoing project that already has `renv` set up, follow these steps to ensure that you're using the same project versions.

1. Install the `renv` package by running `install.packages("renv")`
2. Pull the most updated version of the project from GitHub
3. Open the project's RProject file
4. Run `renv::restore()`. In our lab's projects, this is often already found at the top of the config file, so you can just run scripts as is.
5. This will pull up a list of the project's packages that need to be updated for you to be consistent with the project. The console will ask if you want to proceed with updating these packages - type "Y" to continue.
6. Wait for the correct versions of each package to install/update. This may take some time, depending on how many packages the project uses.
7. Your R environment should now be using the same package versions as specified in the `renv` lock file. You should now be able to replicate the code.
8. If you make edits to the code and introduce new/updated packages, see the section above for instructions on how to make updates.

# 12 Code Publication

Adapted by UCD-SeRG team from original by Nolan Pokpongkiat<sup>1</sup>

## 12.1 Checklist overview

1. Fill out file headers
2. Clean up comments
3. Document functions
4. Remove deprecated filepaths
5. Ensure project runs via bash
6. Complete the README
7. Clean up feature branches
8. Create Github release

## 12.2 Fill out 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. See template here.<sup>2</sup>

## 12.3 Clean up comments

Make sure comments in the code are for code documentation purposes only. Do not leave comments to self in the final script files.

## 12.4 Document functions

Every function you write must include a header to document its purpose, inputs, and outputs. See template for the function documentation here.<sup>3</sup>

---

<sup>1</sup><https://jadebc.github.io/lab-manual/code-publication.html>

<sup>2</sup><https://jadebc.github.io/lab-manual/coding-practices.html#file-headers>

<sup>3</sup><https://jadebc.github.io/lab-manual/coding-practices.html#function-documentation>

## 12.5 Remove deprecated filepaths

All file paths should be defined in 0-config.R, and should be set relative to the project working directory. All absolute file paths from your local computer should be removed, and replaced with a relative path. If a third party were to re-run this analysis, if they need to download data from a separate source and change a filepath in the 0-config.R to match, make sure to specify in the README which line of 0-config.R needs to be substituted.

## 12.6 Ensure project runs via bash

The project should be configured to be entirely reproducible by running a master bash script, run-project.sh, which should live at the top directory. This bash script can call other bash scripts in subfolders, if necessary. Bash scripts should use the runFileSaveLogs utility script, which is a wrapper around the Rscript command, allowing you to specify where .Rout log files are moved after the R scripts are run.

See usage and documentation here.<sup>4</sup>

## 12.7 Complete the README

A README.md should live at the top directory of the project. This usually includes a Project Overview and a Directory Structure, along with the names of the contributors and the Creative Commons License. See below for a template:

### Overview

To date, coronavirus testing in the US has been extremely limited. Confirmed COVID-19 case counts underestimate the total number of infections in the population. We estimated the total COVID-19 infections – both symptomatic and asymptomatic – in the US in March 2020. We used a semi-Bayesian approach to correct for bias due to incomplete testing and imperfect test performance.

### Directory structure

- 0-config.R: configuration file that sets data directories, sources base functions, and loads required libraries
- 0-base-functions: folder containing scripts with functions used in the analysis
  - 0-base-functions.R: R script containing general functions used across the analysis
  - 0-bias-corr-functions.R: R script containing functions used in bias correction
  - 0-bias-corr-functions-undertesting.R: R script containing functions used in bias correction to estimate the percentage of underestimation due to incomplete testing vs. imperfect test accuracy
  - 0-prior-functions.R: R script containing functions to generate priors

---

<sup>4</sup><https://jadebc.github.io/lab-manual/unix.html#example-code-for-runfilesavelogs>

- 1-data: folder containing data processing scripts NOTE: some scripts are deprecated
- 2-analysis: folder containing analysis scripts. To rerun all scripts in this subdirectory, run the bash script 0-run-analysis.sh.
  - 1-obtain-priors-state.R: obtain priors for each state
  - 2-est-expected-cases-state.R: estimate expected cases in each state
  - 3-est-expected-cases-state-perf-testing.R: estimate expected cases in each state, estimate the percentage of underestimation due to incomplete testing vs. imperfect test accuracy
  - 4-obtain-testing-protocols.R: find testing protocols for each state.
  - 5-summarize-results.R: summarize results; obtain results for in text numerical results.
- 3-figure-table-scripts: folder containing figure scripts. To rerun all scripts in this subdirectory, run the bash script 0-run-figs.sh.
  - 1-fig-testing.R: creates plot of testing patterns by state over time
  - 2-fig-cases-usa-state-bar.R: creates bar plot of confirmed vs. estimated infections by state
  - 3a-fig-map-usa-state.R: creates map of confirmed vs. estimated infections by state
  - 3b-fig-map-usa-state-shiny.R: creates map of confirmed vs. estimated infections by state with search functionality by state
  - 4-fig-priors.R: creates figure with priors for US as a whole
  - 5-fig-density-usa.R: creates figure of distribution of estimated cases in the US
  - 6-table-data-quality.R: creates table of data quality grading from COVID Tracking Project
  - 7-fig-testpos.R: creates figure of the probability of testing positive among those tested by state
  - 8-fig-percent-undertesting-state.R: creates figure of the percentage of under estimation due to incomplete testing
- 4-figures: folder containing figure files.
- 5-results: folder containing analysis results objects.
- 6-sensitivity: folder containing scripts to run the sensitivity analyses

**Contributors:** UCD-SeRG team (adapted from original contributors: Jade Benjamin-Chung, Sean L. Wu, Anna Nguyen, Stephanie Djajadi, Nolan N. Pokpongkiat, Anmol Seth, Andrew Mertens)

Wu SL, Mertens A, Crider YS, Nguyen A, Pokpongkiat NN, Djajadi S, et al. Substantial underestimation of SARS-CoV-2 infection in the United States due to incomplete testing and imperfect test accuracy. medRxiv. 2020; 2020.05.12.20091744. doi:10.1101/2020.05.12.20091744

When possible, also include a description of the RDS results that are generated, detailing what data sources were used, where the script lives that creates it, and what information the RDS results hold.

## 12.8 Clean up feature branches

In the remote repository on Github, all feature branches aside from master should be merged in and deleted. All outstanding PRs should be closed.

## 12.9 Create Github release

Once all of these items are verified, create a tag to make a Github release, which will tag the repository, creating a marker at this specific point in time.

Detailed instructions here.<sup>5</sup>

---

<sup>5</sup><https://docs.github.com/en/enterprise/2.13/user/articles/creating-releases>

# 13 Data Publication

Adapted from Fanice Nyatigo and Ben Arnold's chapter in the Proctor-UCSF Lab Manual<sup>1</sup>

## 13.1 Overview

---

**Warning!** **NEVER** push a dataset into the public domain (e.g., GitHub, OSF) without first checking with lab leadership 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 unwieldy. Instead, we recommend using another stable repository that has version control, such as the Open Science Framework ([osf.io](https://osf.io)<sup>2</sup>). 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](https://datadryad.org)<sup>3</sup>) or institutional digital repositories.

---

<sup>1</sup>[https://urlisolation.com/browser?clickId=524DE241-3F8F-4C98-B619-3C278374BF64&traceToken=1728923499%3Bucsfmed\\_hosted%3Bhttps%3A%2F%2Fproctor-ucsf.github.io%2Fd&url=https%3A%2F%2Fproctor-ucsf.github.io%2Fdcc-handbook%2Fpublicdata.html](https://urlisolation.com/browser?clickId=524DE241-3F8F-4C98-B619-3C278374BF64&traceToken=1728923499%3Bucsfmed_hosted%3Bhttps%3A%2F%2Fproctor-ucsf.github.io%2Fd&url=https%3A%2F%2Fproctor-ucsf.github.io%2Fdcc-handbook%2Fpublicdata.html)

<sup>2</sup><https://osf.io>

<sup>3</sup><https://datadryad.org>

We recommend cross-linking public files in GitHub (scripts/notebooks only) and OSF/Dryad/institutional digital repositories.

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/institutional digital 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

## 13.2 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:

### 13.2.1 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

### 13.2.2 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)!*

---

### 13.2.3 Geographic information

Do not include GPS coordinates (longitude, latitude) except in special circumstances where they have been obfuscated/shifted. Reach out to lab leadership 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>

## 13.3 Create public IDs

### 13.3.1 Rationale

The UC Davis 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.

### 13.3.2 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 to 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.

---

### 13.3.3 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>

## 13.4 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<sup>4</sup>. 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.

### 13.4.1 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>

---

<sup>4</sup><https://proctor-ucsf.github.io/dcc-handbook/datawrangling.html#documenting-datasets>

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<sup>5</sup>
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>

## 13.5 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.

## 13.6 Create a public GitHub page for public scripts

At minimum, we should include all of the scripts required to run the analyses. **IMPORTANT:** ensure you have taken a snapshot and saved your computing environment using the `renv` package (`renv`).

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.

---

<sup>5</sup><https://researchdata.princeton.edu/research-lifecycle-guide/whats-doi-and-what-should-i-know-about-citing-datasets>

## 13.7 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 studies from UCSF’s Proctor Foundation:

- 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<sup>6</sup>

---

<sup>6</sup><https://plos.org/resource/how-to-store-and-manage-your-data/#data-management-plan>

# 14 High-performance computing (HPC)

Adapted by UCD-SeRG team from original by Anna Nguyen, Jade Benjamin-Chung, and Gabby Barratt Heitmann<sup>1</sup>

When you need to run a script that requires a large amount of RAM, large files, or that uses parallelization, UC Davis provides several high-performance computing (HPC) resources.

## 14.1 UC Davis Computing Resources

### 14.1.1 Available Resources

**UC Davis HPC Clusters:** - **Farm Cluster** ([hpc.ucdavis.edu<sup>2</sup>](https://hpc.ucdavis.edu)): UC Davis's primary HPC cluster providing shared computing resources for research

**PHS Shared Compute Environments:** For lab members affiliated with the School of Public Health Sciences (PHS), additional shared computing environments are available. These environments provide secure, HIPAA-compliant computing resources suitable for working with sensitive health data.

- **Shiva** ([shiva.ucdavis.edu](https://shiva.ucdavis.edu)): SLURM-based cluster for computational work
- **Mercury** ([mercury.ucdavis.edu](https://mercury.ucdavis.edu)): RStudio GUI computing environment

For detailed information about PHS shared compute environments, including access procedures, security guidelines, and usage policies, please refer to the PHS Shared Compute Environments Guide<sup>3</sup>.

Contact lab leadership for assistance with:

- Requesting access to computing resources
- Choosing the appropriate computing environment for your project
- Setting up your computing environment

## 14.2 Getting started with SLURM clusters

To access a UC Davis HPC cluster, in terminal, log in using SSH. For example, to access shiva:

```
ssh USERNAME@shiva.ucdavis.edu
```

---

<sup>1</sup><https://jadebc.github.io/lab-manual/slurm.html>

<sup>2</sup><https://hpc.ucdavis.edu/>

<sup>3</sup>[assets/files/PHS\\_Shared\\_Compute\\_Environments.pdf](#)

You will be prompted to enter your UC Davis credentials and may need to complete two-factor authentication.

Once you log in, you can view the contents of your home directory in command line by entering `cd $HOME`. You can create subfolders within this directory using the `mkdir` command. For example, you could make a “code” subdirectory and clone a Github repository there using the following code:

```
cd $HOME
mkdir code
git clone https://github.com/jadebc/covid19-infections.git
```

### 14.2.1 One-Time System Set-Up

To keep the install packages consistent across different nodes, you will need to explicitly set the pathway to your R library directory.

Open your `~/.Renviron` file (`vi ~/.Renviron`) and append the following line:

*Note: Once you open the file using `vi [file_name]`, you must press `i` (on Mac OS) or `Insert` (on Windows) to make edits. After you finish, hit `Esc` to exit editing mode and type `:wq` to save and close the file.*

```
R_LIBS=~/R/x86_64-pc-linux-gnu-library/4.0.2
```

Alternatively, run an R script with the following code on the cluster:

```
r_environ_file_path = file.path(Sys.getenv("HOME"), ".Renviron")
if (!file.exists(r_environ_file_path)) file.create(r_environ_file_path)

cat("\nR_LIBS=~/R/x86_64-pc-linux-gnu-library/4.0.2",
    file = r_environ_file_path, sep = "\n", append = TRUE)
```

To load packages that run off of C++, you’ll need to set the correct compiler options in your R environment.

Open the `Makevars` file (`vi ~/.R/Makevars`) and append the following lines

```
CXX14FLAGS=-O3 -march=native -mtune=native -fPIC
CXX14=g++
```

Alternatively, create an R script with the following code, and run it on the cluster:

```
dotR = file.path(Sys.getenv("HOME"), ".R")
if (!file.exists(dotR)) dir.create(dotR)

M = file.path(dotR, "Makevars")
if (!file.exists(M)) file.create(M)

cat("\nCXX14FLAGS=-O3 -march=native -mtune=native -fPIC",
    "CXX14=g++",
    file = M, sep = "\n", append = TRUE)
```

## 14.3 Moving files to the cluster

The \$HOME directory is a good place to store code and small test files. Save large files to the \$SCRATCH directory or other designated storage areas. Check with the UC Davis HPC documentation<sup>4</sup> for specific quotas and retention policies. It's best to create a bash script that records the file transfer process for a given project. See example code below:

```
# note: the following steps should be done from your local
# (not after ssh-ing into the cluster)

# securely transfer folders from Box to cluster home directory
# note: the -r option is for folders and is not needed for files
scp -r "Box/project-folder/folder-1/" USERNAME@shiva.ucdavis.edu:/home/users/USERNAME/

# securely transfer folders from Box to your cluster scratch directory
scp -r "Box/project-folder/folder-2/" USERNAME@shiva.ucdavis.edu:/scratch/users/USERNAME/

# securely transfer folders from Box to shared scratch directory
scp -r "Box/project-folder/folder-3/" USERNAME@shiva.ucdavis.edu:/scratch/group/GROUPNAME/
```

## 14.4 Installing packages on the cluster

When you begin working on a cluster, you will most likely encounter problems with installing packages. To install packages, login to the cluster on the command line and open a development node. Do not attempt to do this in RStudio Server, as you will have to re-do it for every new session you open.

```
ssh USERNAME@shiva.ucdavis.edu

sdev
```

You should only have to install packages once. The cluster may require that you specify the repository where the package is downloaded from. You may also need to add an additional argument to `install.packages` to prevent the packages from locking after installation:

```
install.packages(<PACKAGE NAME>, repos="http://cran.us.r-project.org",
                 INSTALL_opts = "--no-lock")
```

In order for some R packages to work on clusters, it is necessary to load specific software modules before running R. These must be loaded each time you want to use the package in R. For example, for spatial and random effects analyses, you may need the modules/packages below. These modules must also be loaded on the command line prior to opening R in order for package installation to work.

---

<sup>4</sup><https://hpc.ucdavis.edu/>

```

module --force purge # remove any previously loaded modules, including math and devel
module load math
module load math gmp/6.1.2
module load devel
module load gcc/10
module load system
module load json-glib/1.4.4
module load curl/7.81.0
module load physics
module load physics udunits geos
module load physics gdal/2.2.1 # for R/4.0.2
module load physics proj/4.9.3 # for R/4.0.2
module load pandoc/2.7.3

module load R/4.0.2

R # Open R in the Shell window to install individual packages or test code
Rscript install-packages.R # Alternatively, run a package installation script in the Shell

```

Figuring out the issues with some packages will require some trial and error. If you are still encountering problems installing a package, you may have to install other dependencies manually by reading through the error messages. If you try to install a dependency from CRAN and it isn't working, it may be a module. You can search for it using the `module spider` command:

```
module spider DEPENDENCY NAME
```

You can also reach out to UC Davis HPC support for help. Visit [hpc.ucdavis.edu<sup>5</sup>](https://hpc.ucdavis.edu) for support information.

## 14.5 Testing your code

Both of the following ways to test code on a cluster are recommended for making small changes, such as editing file paths and making sure the packages and source files load. You should write and test the functionality of your script locally, only testing on the cluster once major bugs are out.

### 14.5.1 The command line

There are two main ways to explore and test code on computing clusters. The first way is best for users who are comfortable working on the command line and editing code in base R. Even if you are not comfortable yet, this is probably the better way because these commands will transfer between different cluster computers using Slurm.

Typically, you will want to initially test your scripts by initiating a development node using the command `sdev`. This will allocate a small amount of computing resources for 1 hour. You can access R via command line using the following code.

---

<sup>5</sup><https://hpc.ucdavis.edu>

```
# open development node
sdev

# Load all the modules required by the packages you are using
module load MODULE NAME

# Load R (default version)*
module load R

# initiate R in command line
R
```

\*Note: for collaboration purposes, it's best for everyone to work with one version of R. Check what version is being used for the project you are working on. Some packages only work with some versions of R, so it's best to keep it consistent.

### 14.5.2 RStudio Server

For RStudio GUI computing, UC Davis provides mercury.ucdavis.edu. This is accessed through a web browser and provides an RStudio interface. You will be prompted to authenticate with your UC Davis credentials. This is the best way to work with R for people who are not comfortable accessing & editing in base R in a Shell application.

Note that mercury does not have SLURM, so it's best suited for interactive work and smaller computations. For large-scale computations requiring SLURM job scheduling, use shiva.ucdavis.edu instead.

When using RStudio Server, you can test your code interactively. However, do NOT use the RStudio Server's Terminal to install packages and configure your environment for SLURM-based clusters, as you will likely need to re-do it for every session/project. For SLURM clusters, use the command line approach described earlier.

### 14.5.3 Filepaths & configuration on the cluster

In most cases, you will want to test that the file paths work correctly on the cluster. You will likely need to add code to the configuration file in the project repository that specifies cluster-specific file paths. Here is an example:

```
# set cluster-specific file paths
if(Sys.getenv("LMOD_SYSHOST")!=""){

  cluster_path = paste0(Sys.getenv("HOME"), "/project-name/")

  data_path = paste0(cluster_path, "data/")
  results_path = paste0(cluster_path, "results/")
}
```

## 14.6 Storage & group storage access

### 14.6.1 Individual storage

There are multiple places to store your files on computing clusters. Each user has their own \$HOME directory as well as a \$SCRATCH directory. These are directories that can be accessed via the command line once you've logged in to the cluster:

```
cd $HOME
cd /home/users/USERNAME # Alternatively, use the full path

cd $SCRATCH
cd /scratch/users/USERNAME # Full path
```

You can also navigate to these using the File Explorer if available through a web interface.

\$HOME typically has a volume quota (e.g., 15 GB). \$SCRATCH typically has a larger volume quota (e.g., 100 TB), but files here may get deleted after a certain period of inactivity. Thus, use \$SCRATCH for test files, exploratory analyses, and temporary storage. Use \$HOME for long-term storage of important files and more finalized analyses.

Check with the UC Davis HPC documentation<sup>6</sup> for specific storage options and quotas.

### 14.6.2 Group storage

The lab may have shared \$GROUP\_HOME and \$GROUP\_SCRATCH directories to store files for collaborative use. These typically have larger quotas and may have different retention policies. You can access these via the command line or navigate to them using the File Explorer:

```
cd $GROUP_HOME
cd /home/groups/GROUPNAME

cd $GROUP_SCRATCH
cd /scratch/groups/GROUPNAME
```

However, saving files to group storage can be tricky. You can try using the scp command in the section “Moving files to the cluster” to see if you have permission to add files to group directories. Read the next section to ensure any directories you create have the right permissions.

### 14.6.3 Folder permissions

Generally, when we put folders in \$GROUP\_HOME or \$GROUP\_SCRATCH, it is so that we can collaborate on an analysis within the research group, so multiple people need to be able to access the folders. If you create a new folder in \$GROUP\_HOME or \$GROUP\_SCRATCH, please check the folder's permissions to ensure that other group members are able to access its contents. To check the permissions of a folder, navigate to the level above it, and enter ls -l. You will see output like this:

---

<sup>6</sup><https://hpc.ucdavis.edu/>

```
drwxrwxrwx 2 jadabc jadabc 2204 Jun 17 13:12 myfolder
```

Please review this website<sup>7</sup> to learn how to interpret the code on the left side of this output. The website also tells you how to change folder permissions. In order to ensure that all users and group members are able to access a folder's contents, you can use the following command:

```
chmod ugo+rwx FOLDER_NAME
```

## 14.7 Running big jobs

Once your test scripts run successfully, you can submit an sbatch script for larger jobs. These are text files with a `.sh` suffix. Use a text editor like Sublime to create such a script. Documentation on sbatch options is available here<sup>8</sup>. Here is an example of an sbatch script with the following options:

- `job-name=run_inc`: Job name that will show up in the SLURM system
- `begin=now`: Requests to start the job as soon as the requested resources are available
- `dependency=singleton`: Jobs can begin after all previously launched jobs with the same name and user have ended.
- `mail-type=ALL`: Receive all types of email notification (e.g., when job starts, fails, ends)
- `cpus-per-task=16`: Request 16 processors per task. The default is one processor per task.
- `mem=64G`: Request 64 GB memory per node.
- `output=00-run_inc_log.out`: Create a log file called `00-run_inc_log.out` that contains information about the Slurm session
- `time=47:59:00`: Set maximum run time to 47 hours and 59 minutes. If you don't include this option, the cluster will automatically exit scripts after 2 hours of run time (default may vary by cluster).

The file `analysis.out` will contain the log file for the R script `analysis.R`.

```
#!/bin/bash

#SBATCH --job-name=run_inc
#SBATCH --begin=now
#SBATCH --dependency=singleton
#SBATCH --mail-type=ALL
#SBATCH --cpus-per-task=16
#SBATCH --mem=64G
#SBATCH --mem=64G
#SBATCH --output=00-run_inc_log.out
#SBATCH --time=47:59:00

cd $HOME/project-code-repo/2-analysis/
```

---

<sup>7</sup><https://www.chriswrits.com/how-to-change-file-permissions-using-the-terminal/>

<sup>8</sup><https://slurm.schedmd.com/sbatch.html>

```
module purge

# load R version 4.0.2 (required for certain packages)
module load R/4.0.2

# load gcc, a C++ compiler (required for certain packages)
module load gcc/10

# load software required for spatial analyses in R
module load physics gdal
module load physics proj

R CMD BATCH --no-save analysis.R analysis.out
```

To submit this job, save the code in the chunk above in a script called `myjob.sh` and then enter the following command into terminal:

```
sbatch myjob.sh
```

To check on the status of your job, enter the following code into terminal:

```
squeue -u $USERNAME
```

# 15 Checklists

Adapted by UCD-SeRG team from original by Jade Benjamin-Chung<sup>1</sup>

## 15.1 Pre-analysis plan checklist

- Brief background on the study (a condensed version of the introduction section of the paper)
- Hypotheses / objectives
- Study design
- Description of data
- Definition of outcomes
- Definition of interventions / exposures
- Definition of covariates
- Statistical power calculation
- Statistical model description
- Covariate selection / screening
- Standard error estimation method
- Missing data analysis
- Assessment of effect modification / subgroup analyses
- Sensitivity analyses
- Negative control analyses

## 15.2 Code checklist

- Does the script run without errors?
- Is code self-contained within repo and/or associated Box folder?
- Is all commented out code / remarks removed?
- Does the header accurately describe the process completed in the script?
- Is the script pushed to its github repository?
- Does the code adhere to the coding style guide<sup>2</sup>?
- Are all warnings ignorable? Should any warnings be intentionally suppressed or addressed?

## 15.3 Manuscript checklist

*This is adapted in part from this article<sup>3</sup>.*

---

<sup>1</sup><https://jadebc.github.io/lab-manual/checklists.html>

<sup>2</sup><https://jadebc.github.io/lab-manual/coding-style.html>

<sup>3</sup><https://www.nature.com/articles/d41586-019-01431-z>

- Have you completed the relevant reporting checklist, if applicable? (Collection of checklists<sup>4</sup>)
- Are the study results within the manuscript replicable (i.e., if you rerun the code in the study's repository, the tables and figures will be exactly replicated?)
- Is a target journal selected?
- Is the title declarative, in other words, does it state the object/findings rather than suggest them?
- Is the word count of the manuscript close to the target journal's allowance?
- Does the manuscript adhere to the formatting guide of the target journal?
- Does the manuscript use a consistent voice (passive or active – usually active is preferred ... pun intended)?
- Is each figure and table (including supplementary material) referenced in the main text?
- Is there a caption for each figure and table (including supplementary material)?
- Are tables/figures and supplementary material numbered in accordance with their appearance in the main text?
- Does the text use past tense if it is reporting research findings or future tense if it is a study protocol?
- Does the text avoid subjective wording (e.g., “interesting”, “dramatic”)?
- Does the text use minimal abbreviations, and are all abbreviations defined at first use?
- Does the text avoid directionless words? (e.g., instead of writing, ‘Precipitation influences disease risk’, write, ‘Precipitation was associated with increased disease risk’).
- Does the text avoid making causal claims that are not supported by the study design? Be careful about the words “effect”, “increase”, and “decrease”, which are often interpreted as causal.
- Does the text avoid describing results with the word “significant”, which can easily be confused with statistical significance? (see references on this topic here<sup>5</sup>)
- Have you drafted author contributions? Do they follow the CRediT Taxonomy<sup>6</sup> for author contributions?

## 15.4 Figure checklist

- Are the x-axis and y-axis labeled?
- If the figure includes panels, is each panel labeled?
- Are there sufficient numerical / text labels and breaks on the x-axis and y-axis?
- Is the font size appropriate (i.e., large enough to read, not so large that it distracts from the data presented in the figure?)
- Are the colors used colorblind friendly? See a colorblind-friendly palette here<sup>7</sup>, a neat palette generator with colorblind options here<sup>8</sup>, and an article on why this matters here<sup>9</sup>

<sup>4</sup><https://www.equator-network.org/about-us/what-is-a-reporting-guideline/>

<sup>5</sup>[https://journals.lww.com/epidem/Fulltext/2001/05000/The\\_Value\\_of\\_P.2.aspx](https://journals.lww.com/epidem/Fulltext/2001/05000/The_Value_of_P.2.aspx)

<sup>6</sup>[https://journals.plos.org/plosone/s/authorship/?utm\\_source=plos&utm\\_medium=blog&utm\\_campaign=plos-1607-credit#loc-author-contributions](https://journals.plos.org/plosone/s/authorship/?utm_source=plos&utm_medium=blog&utm_campaign=plos-1607-credit#loc-author-contributions)

<sup>7</sup>[http://www.cookbook-r.com/Graphs/Colors\\_\(ggplot2\)/#a-colorblind-friendly-palette](http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/#a-colorblind-friendly-palette)

<sup>8</sup>[https://medialab.github.io/iwanthue/?utm\\_source=Nature+Briefing&utm\\_campaign=2c68711076-briefing-dy-20211006&utm\\_medium=email&utm\\_term=0\\_c9dfd39373-2c68711076-44335685](https://medialab.github.io/iwanthue/?utm_source=Nature+Briefing&utm_campaign=2c68711076-briefing-dy-20211006&utm_medium=email&utm_term=0_c9dfd39373-2c68711076-44335685)

<sup>9</sup><https://www.nature.com/articles/d41586-021-02696-z>

- Are colors/shapes/line types defined in a legend?
- Are the legends and other labels easy to understand with minimal abbreviations?
- If there is overplotting, is transparency used to show overlapping data?
- Are 95% confidence intervals or other measures of precision shown, if applicable?

# 16 Resources

Adapted by UCD-SeRG team from original by Jade Benjamin-Chung and Kunal Mishra<sup>1</sup>

## 16.1 Resources for R

- dplyr and tidyr cheat sheet<sup>2</sup>
- ggplot cheat sheet<sup>3</sup>
- data table cheat sheet<sup>4</sup>
- RMarkdown cheat sheet<sup>5</sup>
- Hadley Wickham's R Style Guide<sup>6</sup>
- Jade's R-for-epi course<sup>7</sup>
- Tidy Eval in 5 Minutes<sup>8</sup> (video)
- Tidy Evaluation<sup>9</sup> (e-book)
- Data Frame Columns as Arguments to Dplyr Functions<sup>10</sup> (blog)
- Standard Evaluation for \*\_join<sup>11</sup> (stackoverflow)
- Programming with dplyr<sup>12</sup> (package vignette)

## 16.2 Resources for Git & Github

- Data Camp introduction to Git<sup>13</sup>
- Introduction to Github<sup>14</sup>

## 16.3 Scientific figures

- Ten Simple Rules for Better Figures<sup>15</sup>

---

<sup>1</sup><https://jadebc.github.io/lab-manual/resources.html>

<sup>2</sup><https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf>

<sup>3</sup><https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>

<sup>4</sup>[https://s3.amazonaws.com/assets.datacamp.com/blog\\_assets/datatable\\_Cheat\\_Sheet\\_R.pdf](https://s3.amazonaws.com/assets.datacamp.com/blog_assets/datatable_Cheat_Sheet_R.pdf)

<sup>5</sup><https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>

<sup>6</sup><http://adv-r.had.co.nz/Style.html>

<sup>7</sup><https://ucb-epi-r.github.io>

<sup>8</sup><https://www.youtube.com/watch?v=nERXS3ssntw>

<sup>9</sup><https://tidyeval.tidyverse.org/index.html>

<sup>10</sup><https://www.brodrigues.co/blog/2016-07-18-data-frame-columns-as-arguments-to-dplyr-functions/>

<sup>11</sup><https://stackoverflow.com/questions/28125816/r-standard-evaluation-for-join-dplyr>

<sup>12</sup><https://dplyr.tidyverse.org/articles/programming.html>

<sup>13</sup><https://www.datacamp.com/courses/introduction-to-git-for-data-science>

<sup>14</sup><https://lab.github.com/githubtraining/introduction-to-github>

<sup>15</sup><https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833>

## 16.4 Writing

- Tips on how to write a great science paper<sup>16</sup>
- ICMJE Definition of authorship<sup>17</sup>
- Nature article on elements of style for scientific writing<sup>18</sup>
- The Pathway to Publishing: A Guide to Quantitative Writing in the Health Sciences<sup>19</sup>
- Secret, actionable writing tips<sup>20</sup>

## 16.5 Presentations

- How to tell a compelling story in scientific presentations<sup>21</sup>
- How to give a killer narratively-driven scientific talk<sup>22</sup>
- How to make a better poster<sup>23</sup>
- How to make an even better poster<sup>24</sup>

## 16.6 Professional advice

- Professional advice, especially for your first job<sup>25</sup>

## 16.7 Funding

- Building Your Funding Train<sup>26</sup>
- NIH Grant Writing Resources<sup>27</sup>

## 16.8 Ethics and global health research

- Global Code of Conduct For Research in Resource-Poor Settings<sup>28</sup>
- Who is a global health expert? Advice for aspiring global health experts<sup>29</sup>
- Transforming Global Health Partnerships<sup>30</sup>

<sup>16</sup><https://www.nature.com/articles/d41586-019-02918-5>

<sup>17</sup><http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html>

<sup>18</sup><https://www.nature.com/articles/nphys724>

<sup>19</sup><https://link.springer.com/book/10.1007/978-3-030-98175-4>

<sup>20</sup><https://x.com/acagamic/status/1680381737816424450>

<sup>21</sup><https://www.nature.com/articles/d41586-021-03603-2>

<sup>22</sup><https://www.sciencedirect.com/science/article/pii/S1471491423000928>

<sup>23</sup><https://www.youtube.com/watch?v=1RwJbhkCA5&t=1171s>

<sup>24</sup><https://mitcommllab.mit.edu/be/2023/09/27/toward-an-evenbetterposter-improving-the-betterposter-template/>

<sup>25</sup>[https://docs.google.com/document/d/1ckgRCcr7FFPyymyMA6Y\\_-cB\\_vftOyrrxT28c6gRg0PI/edit](https://docs.google.com/document/d/1ckgRCcr7FFPyymyMA6Y_-cB_vftOyrrxT28c6gRg0PI/edit)

<sup>26</sup><https://grantwriting.stanford.edu/funding-train/#ep>

<sup>27</sup><https://www.niaid.nih.gov/grants-contracts/write-grant-application>

<sup>28</sup><https://www.globalcodeofconduct.org/wp-content/uploads/2018/05/Global-Code-of-Conduct-Brochure.pdf>

<sup>29</sup><https://journals.plos.org/globalpublichealth/article?id=10.1371/journal.pgph.0002269>

<sup>30</sup><https://link.springer.com/book/9783031537929>