


# Notes on Workflow for Mental Health and the Justice System in Durham County

Maria Tackett

2023-06-05

## Table of contents

<b>Reproducibility</b>	<b>1</b>
<b>Data</b>	<b>2</b>
Why use it in PACE? . . . . .	3
How to use it in PACE . . . . .	3
<b>File organization</b>	<b>3</b>
<b>Workflow</b>	<b>3</b>
<b>Resources</b>	<b>4</b>
	<b>4</b>

 Note

The source code for this document is available at [github.com/matackett/mhjs-workflow](https://github.com/matackett/mhjs-workflow). Please [open an issue](#) or [submit a pull request](#) if you have any edits or additions.

## Reproducibility

In *Telling Stories with Data* (Alexander 2023), Rohan Alexander describes reproducibility as the following (emphasis mine):

*“Alexander (2019) defines reproducible research as that which can be exactly redone, given all the materials used. This underscores the importance of providing the **code, data, and environment**. The minimum expectation is that another person is independently able to use your code, data, and environment to get your results, including figures and tables.”*

Using this definition as a guide, we can think about how to make our research reproducible by establishing good practices in each of the following components:

1. Data
2. Code (and analysis)
3. Organization
4. Environment

We’ll address each of these components separately to help organize our thinking, but in reality all these components intertwine, so we’ll often need to discuss one while discussing the other. For example, we can’t establish best practices for coding without thinking about the data.

- Why is it important to do the work on this project in a reproducible way?
- What are practical considerations to keep in mind as we develop reproducible practices for this project?

## Data

In the 2021 paper [Datasets for Datasheets](#) (Gebru et al. 2021), Gebru et al. name two stakeholders for a data set: *data creators* and *data consumers*. For many data sets the creator and consumer are one in the same. In either case, clear and detailed data documentation is critical for both parties to understand what’s in a data set, how the data can be analyzed, and the context and limitations of any results gleaned from the data.

At a minimum, every data set should have a *codebook* (aka data dictionary) that documents what each row represents, and every every column in the data set

- Column name
- Data type
- Precise definition
- Possible values / categories
- Units (if applicable)

Another important piece of documentation is the datasheet that provides more information regarding the motivation for the data set, data collection, and ethical considerations. Gebru et al. (2021) gives a list of questions we can use to create “a datasheet that documents its

motivation, composition, collection process, recommended uses, and so on.” The datasheet is intended for *data creators* and *data consumers*. Datasheets serve a specific purpose for each group:

- **Data creators:** “to encourage careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use”
- **Data consumers:** “to ensure they have the information they need to make informed decisions about using a dataset”

! Something about looking at the paper and identifying what questions are relevant for a datasheet for our group.

Version control

### Why use it in PACE?

### How to use it in PACE

- Download git

### File organization

### Workflow

- Choose file type based on the coding purpose
- Use R **script** to create data sets or other one time tasks
  - use comments throughout code
- Use Quarto for analysis
  - include narrative and some comments
  - can use bullet points and notes, does not need to be fancy narrative until we’re ready to present results

## Resources

- [Chapter 3: Reproducible Workflows](#) in *Telling Stories with Data*
- [Chapter 29: Quarto](#) in *R For Data Science* (2nd edition)
- [Datasheets for Datasets](#) by Timnit Gebru et al.

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. <https://tellingstorieswithdata.com>.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2021. “Datasheets for Datasets.” <https://arxiv.org/abs/1803.09010>.