# RDC reproducibility

Grant Gibson, AD Research & Evaluation - CRDCN

CRDCN and McMaster University

Date: 2024-05-15

CRDCN Canadian Research Data Centre Network
RCCDR Réseau canadien des Centres de données de recherche

# Intro & Motivation

- Growing scope of reproducibility interest generally
- Pressures can come from funders, journals, universities, supervisors
- Not a lot of comprehensive, accessible resources either within or across disciplines

# Reproducibility

- Using definition B1 from Barba(2018)[1]
  *"Reproducibility" refers to instances in which the original researcher's data and computer codes are used to regenerate the results, while "replicability" refers to instances in which a researcher collects new data to arrive at the same scientific findings as a previous study.*
- If a study is reproducible it makes it *dramatically* easier to replicate so these two concepts are obviously intertwined.

[1] Barba, Lorena A. "Terminologies for reproducible research." arXiv preprint arXiv:1802.03311 (2018).

# 2 key elements

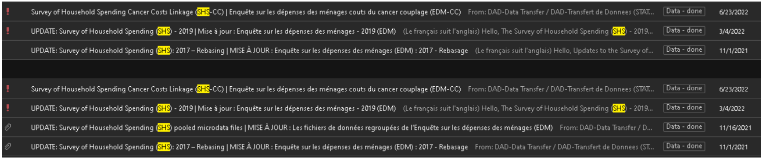
1. Finding and using the data from the original study
2. Very thorough documentation of procedures & decisions made during the analysis (also findable)

# 1. Finding and using the data from the original study (normally)

- Deposit and/or cite data with a PID (doi or similar)

# 1. Finding and using the data from the original study (RDC)

- This is the hardest part because you don't have control.
- Updates not documented using numbers or persistent identifiers, not always recoverable



| | | | | |
|---|---|---|---|---|
| ! | Survey of Household Spending Cancer Costs Linkage (SHS-CC) \| Enquête sur les dépenses des ménages couts du cancer couplage (EDM-CC) | From: DAD-Data Transfer / DAD-Transfert de Donnees (STAT... | Data - done | 6/23/2022 |
| ! | UPDATE: Survey of Household Spending (SHS) - 2019 \| Mise à jour : Enquête sur les dépenses des ménages - 2019 (EDM) | (Le français suit l'anglais) Hello, The Survey of Household Spending (SHS) - 2019... | Data - done | 3/4/2022 |
| | UPDATE: Survey of Household Spending (SHS): 2017 – Rebasing \| MISE À JOUR : Enquête sur les dépenses des ménages (EDM) : 2017 - Rebasage | (Le français suit l'anglais) Hello, Updates to the Survey of... | Data - done | 11/1/2021 |
| | | | | |
| ! | Survey of Household Spending Cancer Costs Linkage (SHS-CC) \| Enquête sur les dépenses des ménages couts du cancer couplage (EDM-CC) | From: DAD-Data Transfer / DAD-Transfert de Donnees (STAT... | Data - done | 6/23/2022 |
| ! | UPDATE: Survey of Household Spending (SHS) - 2019 \| Mise à jour : Enquête sur les dépenses des ménages - 2019 (EDM) | (Le français suit l'anglais) Hello, The Survey of Household Spending (SHS) - 2019... | Data - done | 3/4/2022 |
| 📎 | UPDATE: Survey of Household Spending (SHS) pooled microdata files \| MISE À JOUR : Les fichiers de données regroupées de l'Enquête sur les dépenses des ménages (EDM) | From: DAD-Data Transfer / D... | Data - done | 11/16/2021 |
| 📎 | UPDATE: Survey of Household Spending (SHS): 2017 – Rebasing \| MISE À JOUR : Enquête sur les dépenses des ménages (EDM) : 2017 - Rebasage | From: DAD-Data Transfer / DAD-Transfert de Donnees (ST... | Data - done | 11/1/2021 |

- Make sure you cite your data with the Date (see **How to Cite Statistics Canada Products**). Or, more usefully **Citation of Datasets**
- **Prepare a data accessibility statement**

# 2. Very thorough documentation of procedures & decisions

- This one is on you.
- Concept of computational reproducibility, push-button reproducibility
- Structuring your files

# 2a. RDC-specific issues with procedures documentation

- You likely have a trail of files related to disclosure requests, intermediate disclosures frustrate the process.
- Your code can be released as part of the disclosure request *IF* you don't have anything confidential in it.
- This will be intuitive for people who use R or Python, simply use intermediate objects:

```
inc_sd<-sd(income)

inc_mean<-mean(income)

inc_CV<- (inc_sd/inc_mean)
```

then put in `inc_CV` rather than ever using a numeric expression. Makes it easier if your data gets updated, and disclosure-friendly.

## 2a. RDC-specific issues with procedures documentation

- In Stata, you can assign things from the rreturn or ereturn following a command to a local or global macro. (see my **Stata Training** ->Varia/Using Stored Values for more info)

# 3. What happens next

- Gather your data accessibility statement and statistical code in one place once you have your final result. Publish them to a dataverse. The journal in which you publish may have one, but if not, you can use **the one from McMaster**.
- **See an example** from an RDC project using the LFS.
- Note that you will need to create a *Readme* that explains the contents of your repository and how the files interact.
- Now you can reference this item in your article - work with the journal to do this or include it in your references. Reference your article in your repository.

# RDM within the RDCs

- It's not like it used to be: Info Management policies have changed
- Storage limitations in the RDCs and eventually in vRDC = no viable long-term storage/archival option
- Generally think that you have 5-6 years from project inception for data storage.
  - So even if you want to re-use your code for a follow-up project, make sure you have a copy of your code.

# Software within the RDCs

- Also not up to you, but versioning info is visible to you and should be recorded.
  - Most tools require you to have control over your file system, or make use of a cloud environment. Can't use checkpoint, can't bundle using renv, other plugins are generally not useful (**repado** for example).
  - Many RDC projects are multi-software.

# Example for these slides

E.g. in R-studio I'm using RStudio 2024.04.0+735 "Chocolate Cosmos" Release (a00d0e775dbc93e0d79a1bf474e3e8e8de677383, 2024-04-24) for windows Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) RStudio/2024.04.0+735 Chrome/120.0.6099.291 Electron/28.2.6 Safari/537.36, Quarto 1.4.553 (Copied form Help/About RStudio)

and knitr 1.46; rmarkdown 2.26

```r
cat("R version:", R.version$version.string, "\n")
```

```
## R version: R version 4.4.0 (2024-04-24 ucrt)
```

```r
cat("Platform:", R.version$platform, "\n")
```

```
## Platform: x86_64-w64-mingw32
```

# Final note for ‘push button’ reproducibility

- Take care to make sure that the code produces the results exactly.
- For survey data, this means you ought to rerun all the code one time at the end in sequence.
- If this isn’t feasible with a big admin database you need to be better disciplined.You can use logs, a date last run line in your code, or any other mechanism to make sure things are happening in the right sequence.
- Make use of the random number generator and set the seed appropriately (no more than once per problem).

# More info (links)

- Content from my Reproducibility & RDM module
- Reproducibility checklist
- Renv on GitHub - can be used with Python if using reticulate.