Reproducible Research

Joshua Entrop

NLG Epi Methods Retreat 10th of May 2023

Agenda

I. Statistical Analysis Plans
II. Good Coding Practices
III. Code Validation
IV. Automation
V. Research Documentation

0. Reproducibility

- Research Documentation
- Automatization
- Code Validation
- Good Coding Practices
- Statistical Analysis Plans

I. Statistical Analysis Plans

I. Statistical Analysis Plans

Aims

1. Lower the likelihood of type 1 error due to ad hoc analyses
2. Increase transparency of decisions made during the research process

Example Table of Content

I. Statistical Analysis Plans (cont'd)

Hiemstra *et al. BMC Medical Research Methodology* (2019) 19:233
https://doi.org/10.1186/s12874-019-0879-5

BMC Medical Research Methodology

DEBATE **Open Access**

DEBATE-statistical analysis plans for observational studies

Bart Hiemstra[1*], Frederik Keus[2], Jørn Wetterslev[3], Christian Gluud[3] and Iwan C. C. van der Horst[4]

American Journal of Epidemiology

Vol. 183, No. 8
DOI: 10.1093/aje/kwv254
Advance Access publication: March 18, 2016

Practice of Epidemiology

Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available

Miguel A. Hernán* and James M. Robins

I. Statistical Analysis Plans (cont'd)

IV. INCLUSION CRITERIA

HL patients aged between 18-40 years at diagnosis were identified through the Swedish Lymphoma Register, the Danish Lymphoma Register (LYFO) and the Clinical Lymphoma Database at Oslo University Hospital in Norway using the codes listed in Table 1.

Table 1 List of codes used to identify HL patients by registers.

Register	Identification of HL cases	Years included
Clinical Lymphoma Database at Oslo University Hospital	ICD-O 3ed: 9560-9667 If no ICD-O code was available ICD-10 C81	1995 - 2016
Swedish Lymphoma Register	ICD-O 3ed: 9560-9667	2000 - 2018
Danish Lymphoma Register	ICD-O 3ed: 9560-9667	2000 - 2019

A. OUTCOME - EFFICACY VARIABLES

`entry_dt` Start of follow-up (date of diagnosis + 9 months)
Variable class: date ("YYYY-MM-DD").

`exit_dt` Minimum of date of childbirth, date of relapse, date of stem cell transplantation, date of death, date a comparator became a case, or date of administrative censoring (Table 2).
Variable class: date ("YYYY-MM-DD").

`st_yrs` Follow-up time in years (i.e. difference between `exit_dt` and `entry_dt` divided by 365.24).
Variable class: double.

`exit_event` Event that lead to right censoring (`exit_dt`)
Variable class: factor

1 = "CB" (if `exit_dt` is equal to `cb_1_dt`)
2 = "SCT" (if `exit_dt` is equal to `sct_dt`)
3 = "death" (if `exit_dt` is equal to `death_dt`)
4 = "END" (if `exit_dt` is equal to `dx_dt_comp`, `fup_10_dt`, or `ac_dt`)

I. Statistical Analysis Plans (cont'd)

Statistical analysis plans are the prefect place for defining your DAGs

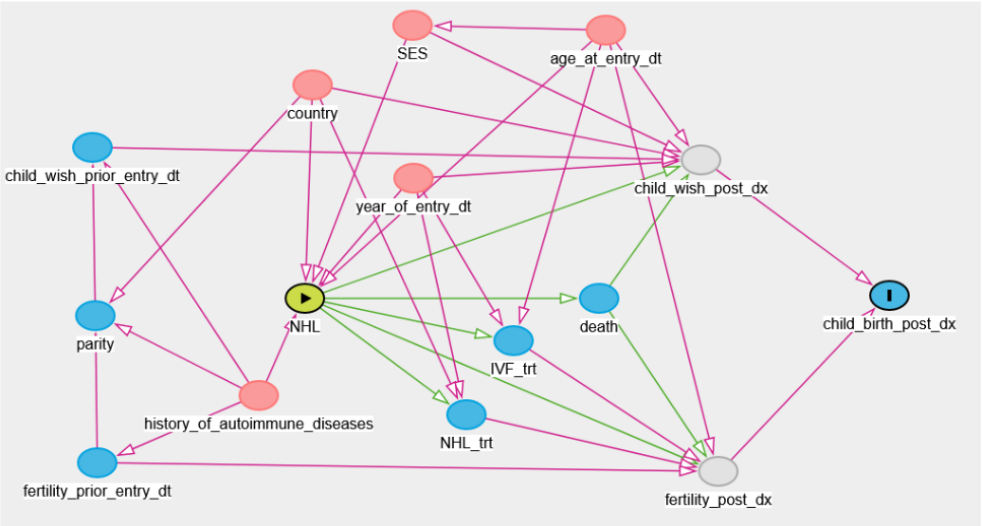


Figure 1: Hypothesised unadjusted DAG of the association between NHL subtypes and childbirth in NHL survivors including a differentiation between childbearing behaviour and fertility. Legend: Pink lines: Biased paths; Green lines: Causal paths; Grey nodes: Unmeasured variables; Red nodes: Potential confounders; Blue nodes: Ancestors of the outcome.

I. Statistical Analysis Plans – to pre-registration

OSFHOME ▾ Search Support Donate Sign Up Sign In

Reproductive Patterns Among Non-Ho... Metadata Files Wiki Analytics Registrations

1.2MB Public 0 •••

Reproductive Patterns Among Non-Hodgkin Lymphoma Survivors by Subtypes in Sweden, Denmark and Norway

Contributors: Joshua Philipp Entrop, Karin Ekström Smedby, Caroline Weibull, Tarec El-Galaly, Sandra Eloranta

Date created: 2022-01-28 11:10 AM | Last Updated: 2022-12-08 11:21 AM

Identifier: DOI 10.17605/OSF.IO/A4U3D

Category: Project

Description: *This study aims to investigate the total effect of NHL subtypes on live childbirths in NHL survivors in Sweden, Denmark, and the South-Eastern Health region in Norway.Specifically, we aim to estimate:*

Aim 1: The absolute and relative rates of childbirth over time since index date (date of diagnosis + 9 months) among NHL survivors, by lymphoma subtype, compared to matched general population comparators (free from NHL).

Aim 2: The cumulative incidence of first (and possibly recurrent) childbirths over time since index date among NHL survivors, by lymphoma subtype, compared to matched general population comparators (free from NHL).

License: *CC-By Attribution 4.0 International*

Wiki

Changes to the pre-registration on June 27th, 2022

- Replaced *childbearing* with *reproductive patterns* in the title: This change was made in order to find a term that includes both children born to female lymphoma survivors and children born to partners of male lymphoma survivors.
- The ICD-10 code *Z71.1* was replaced by *Z71.7* in the definition of HIV history: The code *Z71.1* has been falsely used in th...

Read More

Citation

Recent Activity

Joshua Philipp Entrop linked GitHub repo entjos/Reproduction_Among_NHL_Patients to Reproductive Patterns Among Non-Hodgkin Lymphoma Survivors by Subtypes in Sweden, Denmark and Norway

2022-12-08 11:21 AM

Joshua Philipp Entrop authorized the GitHub addon for Reproductive Patterns Among Non-Hodgkin Lymphoma Survivors by Subtypes in Sweden, Denmark and Norway

Available at https://osf.io/a4u3d/

I. Statistical Analysis Plans – to pre-registration

What to do if you would like to change you pre-registration?

Wiki

Changes to the pre-registration on June 27th, 2022

- Replaced *childbearing* with *reproductive patterns* in the title: This change was made in order to find a term that includes both children born to female lymphoma survivors and children born to partners of male lymphoma survivors.
- The ICD-10 code *Z71.1* was replaced by *Z71.7* in the definition of HIV history: The code *Z71.1* has been falsely used in th...

Read More

I. Statistical Analysis Plans – discussion

Are you using statistical analysis plans in your research project?

What is the purpose of using statistical analysis plans in your projects?

What are you recording in your statistical analysis plans?

II. Good Coding Practices

II. Good Coding Practices

Aim: Increase readability and usability of programme code

Coding styles

1

2

{Styler}[3] & {Lintr}[4]

Project structure

```
./example_project
   +--- data
   +--- scripts
   +--- outputs
   |    +--- graphs
   |    +--- tables
   +--- library
```


[1]https://style.tidyverse.org/; [2]https://google.github.io/styleguide/; [3]https://cran.r-project.org/web/packages/styler/index.html; [4]https://cran.r-project.org/web/packages/lintr/index.html

II. Good Coding Practices – coding styles

Some basic rules

- Use verbs for function names
- Use nouns for object names
- Keep to maximum 80 characters in one line

```
# Good
add_row()
permute()

# Bad
row_adder()
permutation()
```

Some more advanced rules

- Use indentation to highlight code structure
- Do not use spaces before closing and after opening brackets
- Use spaces after commas

```
# Good
cox.ph(Surv(st, event) ~ x1 + x2,
       data = brcancer)

# Bad
cox.ph(Surv(st, event) ~ x1 + x2,
```

II. Good Coding Practices – project structure

Portability

```
./example_project
   +--- data
   +--- scripts
   +--- outputs
   |    +--- graphs
   |    +--- tables
   +--- library
```

R-projects[1]

[2]

[1]https://support.posit.co/hc/en-us/articles/200526207-Using-RStudio-Projects; [2]https://rstudio.github.io/renv/articles/renv.html

II. Good Coding Practices – project structure

Portability

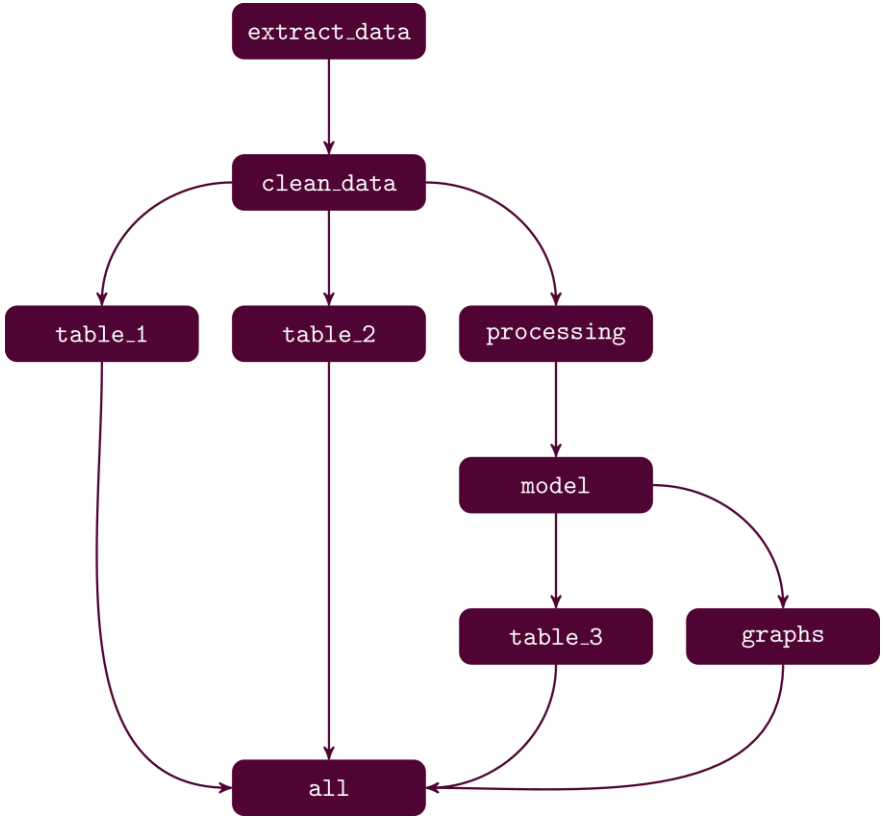
```
# Good
dta <- read("./data/index_pop")

# Bad
dta <- read("C:/data/index_pop")
```

R-projects[1]

[2]

[1]https://support.posit.co/hc/en-us/articles/200526207-Using-RStudio-Projects; [2]https://rstudio.github.io/renv/articles/renv.html

II. Good Coding Practices – modularise your code




```
# Example set up of scripts
./scripts
   +--- /data_cleaning
   +--- /modelling
   +--- /tables
   +--- /graphs
   +--- /user_functions
           +--- __init__.R
           +--- gen_risksets.R
           +--- gen_nhl_subtypes.R
           +--- summary_table.R
```

[1]https://cran.r-project.org/web/packages/box/index.html

II. Good Coding Practices – discussion

Have you worked with a coding style before?

How do you usually set up your research project folders?

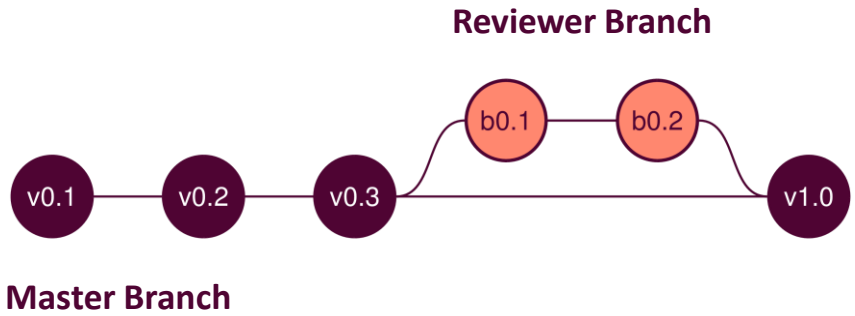
III. Code Validation

III. Code Validation

Aims

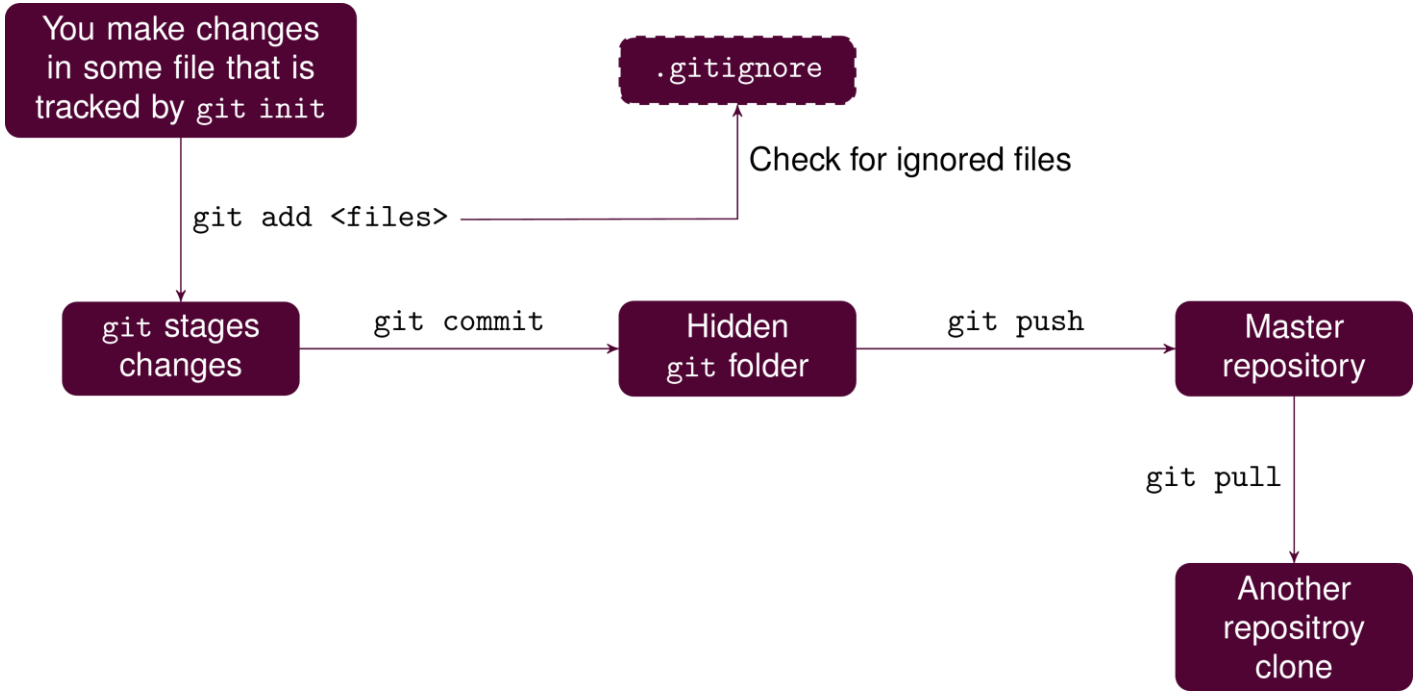
1. Assure logical correctness of programming code
2. Assure technical correctness of programming code

[1]



[1]https://git-scm.com/

III. Code Validation – Git workflow



Format cli #1

Merged entjos merged 5 commits into `entjos:main` from `simonsteiger:format-cli` on Mar 27

Conversation 1 | Commits 5 | Checks 0 | Files changed 3

simonsteiger commented on Mar 23 — Contributor

- Formatted error messages and print method with cli
- Added scalar operators to if-statements where necessary
- Fixed typos

simonsteiger and others added 4 commits 2 months ago

Add cli formatting		6798013
Remove rlang dependency	Verified	14ac2f4
Fix if statement with scalar operator		81b8ea9
Merge branch 'format-cli' of https://github.com/simonsteiger/Exclusio…		449ab4b

entjos added the enhancement label on Mar 27

entjos requested changes on Mar 27 — View reviewed changes

entjos left a comment — Owner

Enhancing the warning messages with `{cli}` is very nice and useful. I, unfortunately, however, prefer the previous printing method for `ex1_tbl` due to the printing of helping lines in the table. It would be great if you could remove the changes to the printing method while keeping the updated warning messages.

III. Code Validation – example

Fix if statement with scalar operator

main (#1)

simonsteiger committed on Mar 23

2 R/exclusion_table.R

```diff
@@ -96,7 +96,7 @@ exclusion_table <- function(
96  96      cli::cli_abort("{.var data} is not a {.cls data.frame} object.")
97  97    }
98  98
99      - if(is.null(inclusion_criteria) & is.null(exclusion_criteria)){
    99  + if(is.null(inclusion_criteria) && is.null(exclusion_criteria)){
100 100     cli::cli_abort(c(
101 101       "Require at least one criterion",
102 102       "x" = "Both {.var inclusion_criteria} and {.var exclusion_criteria} are unspecified.",
```

III. Code Validation – discussion

Have you done code validations before?

Do use guidelines for your code reviews?

IV. Automation

IV. Automation

Aim: Enable others to reproduce the exact outputs of your research project.

Automated processes ⟷ Manual processes

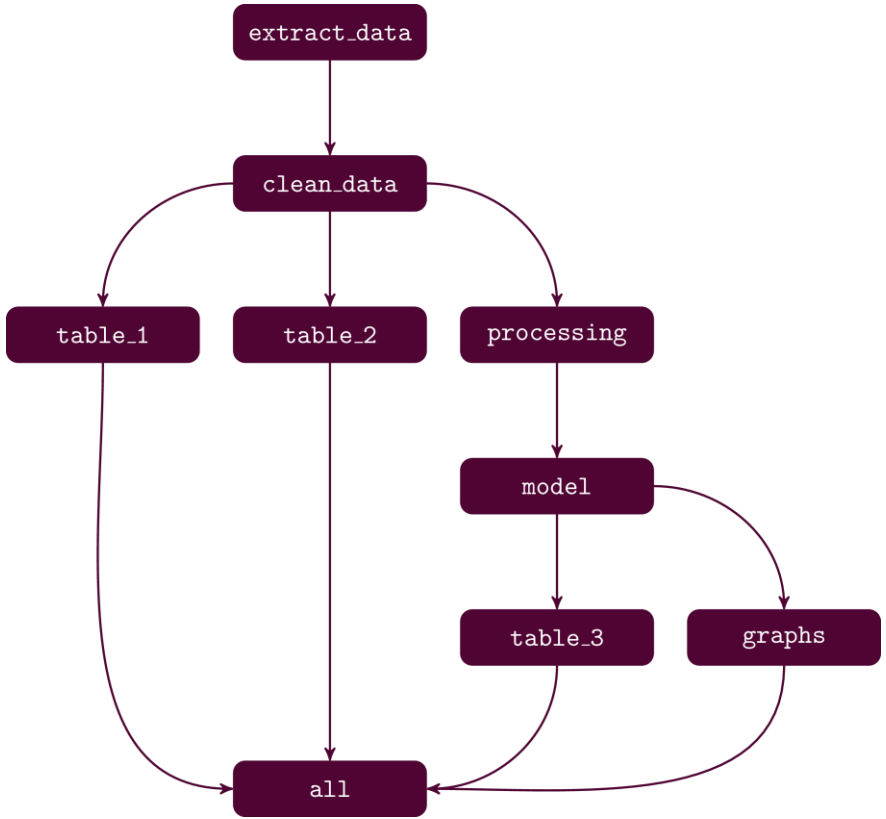
- Automated table creation
- Using automated workflows
- Using local package libraries

- Copy pasting of numbers
- Writing down in and output files
- Manually running scripts
- Installing required packages

[1] [2]

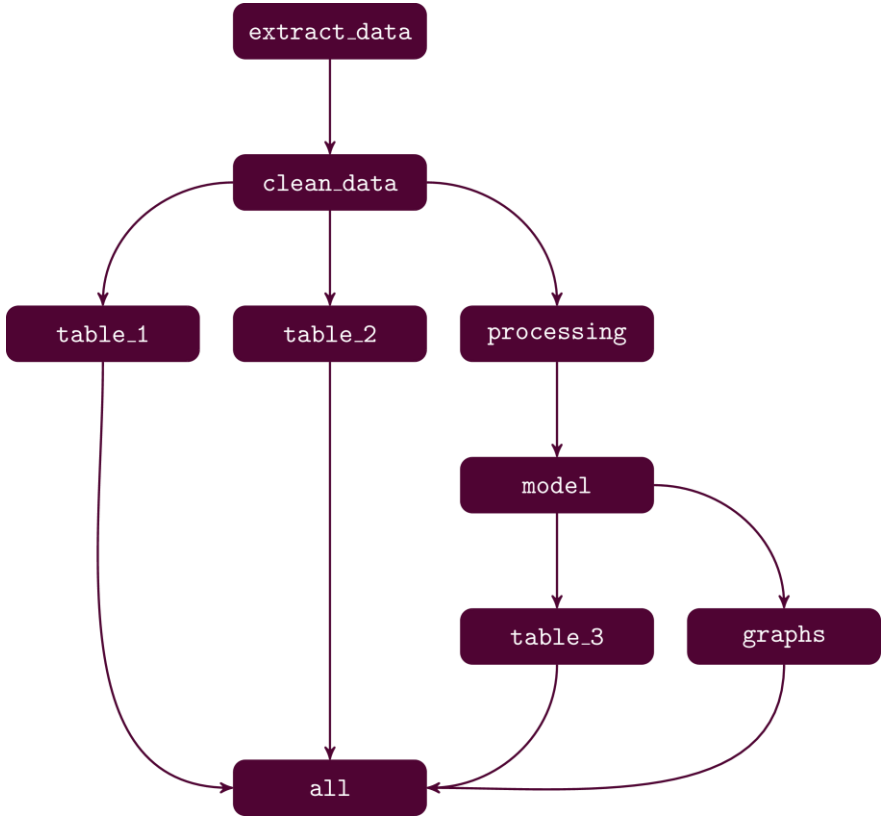
[1]https://www.gnu.org/software/make/; [2]https://quarto.org/

IV. Automation – automated workflows




```
# Stata master file

# Data preparation
do extract_data.do
do clean_data.do
do def_st.do

# Modelling
do model.do

# Tables
do table_1.do
do table_2.do
do table_3.do

# Graphs
do graphs.do
```

IV. Automation – automated workflows (cont'd)




```
# Make example
<target>: <dependencies>
   <rule>

table_2.out: table_2.R \
             clean_data.out
   R CMD BATCH table_2.R
```


[1]https://snakemake.readthedocs.io/en/stable/

IV. Automation – automated workflows (cont'd)

```
:: Example Run.bat file

Welcome to my project please use one of the
following options:

        1. Compile the project
        2. Show the log files form the last run
        3. Delete files from a previous run
```

IV. Automation – discussion

Have you ever tried to rerun some of your previous analysis?

Which parts of your research are you automating?

V. Documentation

V. Documentation

Aim: Enable others to recreate your research project and understand your reasoning.

Internal documentation

- National regulations
- Funder regulations
- University regulation
- University archiving systems

External documentation

OSF[1]

[1]https://osf.io/

V. Documentation - external

Files

Click on a storage provider or drag and drop to upload

Filter

Name	Modified
Childbearing rates in Hodgkin lymphoma survivors treated with BEA...	
GitHub: entjos/Reproduction_After_ABVD_BEACOPP (main)	
analysing	
data_processing	
README.md	
user_defined_functions	
OSF Storage (United States)	
pre_registration20210630.pdf	2021-07-06 03:47 PM
Updated pre-registration (2022-01-27)	
OSF Storage (United States)	
updated_pre_registration20220126.pdf	2022-01-27 09:26 AM

Available at https://osf.io/eumy5/

V. Documentation – discussion

What are your archiving/documentation regulations?

How do we document Nordic collaboration projects?

Karolinska Institutet