Replication Package ReadMe for "Scooped! Estimating Rewards for Priority in Science"

Ryan Hill and Carolyn Stein

August 20, 2024

1 Overview

This replication package provides data and code to produce the analysis in "Scooped! Estimating Rewards for Priority in Science." The data contains a mix of publicly-available and proprietary data sources. All code is provided, but some code will not run without access to the proprietary data. Versions of the final analysis tables can be produced in part as described below. The data schematic in this pdf shows the data build and analysis process, with restricted data noted in red. All code runs through the Stata file master.do.

2 Raw Data Sources and Availability

2.1 Publicly Available Data

Protein Data Bank

Accessed from the RCSB Custom Report Web Service on May 22, 2018. The RCSB has changed the data formatting and API protocols since then, but information about accessing data can be found here: https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction-to-rcsb-pdb-apis.

- pdb_StructureSummary.csv
- pdb_Citation.csv
- pdb_RefinementDetails.csv
- pdb_refinementParameters.csv
- pdb_DataCollectionDetails.csv
- clusters'XX'.csv
- clean_citation.dta

Author-ity

A publicly available database that creates name disambiguated author ids for PubMed-linked publications. This project uses the 2018 version, available for public use here: https://experts.illinois.edu/en/datasets/author-ity-2018-pubmed-author-name-disambiguated-dataset.

- author-ity_raw.csv

Journal Citation Reports

Journal impact factors by year accessed from Thomson Reuters in 2018. Data was publicly available at the time, and is therefore included in the replication package. Since then, the product has been sold to Clarivate and integrated with the proprietary Web of Science Core Collection (see below)

- 'Year'_jifs.csv

QS Top Universities Ranking for Life Sciences and Medicines

Publicly-available data accessed in 2018. Current landing page: https://www.topuniversities.com/university-subject-rankings/life-sciences-medicine

- QS_lifesci_rankings.csv
- affiliation_crosswalk.csv (created by hand to merge QS data to Web of Science affiliations)

World Cities

Publicly-available list of cities and countries accessed in June 2019: https://simplemaps.com/data/world-cities

- world_cities.csv

Editorial Dates

Publicly-available data on publisher websites scraped in August 2019. Contains all received, accepted, and publication dates from Nature, Cell, PloS, and Science families of journals.

- doi_CELL_output.csv
- doi_NATURE_output.csv
- doi_PLOS_output.csv
- doi_SCIENCE_output.csv

PubMed

Data from PubMed listing the volume and issue of each journal linked by pubmed id. Scraped in 2019 directly from the PubMed website.

- vol_issue.dta

2.2 Proprietary Data

Below are the proprietary data sources that would need to be obtained in order to run the full code base. Some have open-access alternatives. For example, OpenAlex publications and citations database could be used instead of Web of Science. The necessary variables are described in bullets below.

Web of Science

Proprietary data owned by Stanford University (curated and provided by Heidi Williams in 2020). Data can be accessed from Web of Science through a data sharing agreement (see https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/web-of-science-core-collection/).

- wos_references_pub'year'.csv
  - citing_wosid
  - cited_wosid
  - citing_year
  - cited_year
- wos_pmid.csv
  - wos_uid
  - pmid (PubMed Identifier)
- wos_titles.csv
  - wos_uid
  - title
  - title_type
- wos_summary_names.csv
  - wos_uid
  - record_order
  - names_count
  - wos_standard
- pmid_affs.dta
  - pmid
  - record_order
  - author_id
  - affiliation

Altmetric

Proprietary data was accessed on August 2nd, 2019. At the time they offered researcher access for free, although policies may have changed. Landing page for data access is here: https://www.altmetric.com/solutions/altmetric-api/.

- altmetric_output.csv

- pmid
- score
- history
- readers_count
- cited_by_posts_count
- cited_by_tweeters_count
- cited_by_wikipedia_count
- cited_by_patents_count
- cited_by_msm_count

Survey

The raw survey data is confidential and not provided in the replication package, but a anonymized version is included for the main analysis. See below for details

- survey_results_10_17_2019.csv

3 Instructions for Running Code

3.1 Data

The Data folder houses all of the publicly available data for this project. The 01_Raw folder houses the raw data. Within this folder, the Restricted folder is empty; this is where users would input the restricted data. Running the code will populate the 02_Clean and 03_Built folders within the Data folder. However, a few files in these folders are pre-populated. These are files where we cannot share the raw data, but we can share the processed data. In some cases, variables have been removed from these processed files to allow for sharing.

3.2 Code

The Code folder houses all of the code for this project. All code provided in the replication package is run on Stata (version is pre-set to Stata 16 in the master do-file). The project runs from top to bottom using master.do. The only required step for the user is to replace INSERT_USER_ROOT with the root directory where the replication folders are stored. If the user is missing any of the installed user-written packages, they should also uncomment the lines where these are installed in the master do-file (see note in 3.1 below on user-written packages).

The do-files are arranged in four folders, with each being called by master.do. The files marked in red below use proprietary data and cannot run without the user first acquiring the data and replacing the data file. The master file currently has those cleaning files turned off so that the code can run with the available data.

The analysis will output figures into the /Output/Figures/ subfolder. The regression estimates are automatically output into sheets in /Output/Tables.xlsx. The numbers populate cells at the bottom of the sheet and are automatically called into the formatted tables at the top of each sheet. We have replaced

some of the proprietary variables (such as citations) in the clean data folder with missings or zeros. This allows the user to run the analysis with only the public-use data, therefore some of the columns will return estimates of zero. These cells are shaded in red in the excel output file.[1]

- master.do
  - Sets directories for code and output folders
  - Sets globals for the control variables used in the analysis
  - Installs user-written packages used in the analysis (if necessary)
  - Provides switches for each do-file, allowing the user to run all or some of the code in pieces.
- 00_PrepWoS
  - 01_prep_author-ity.do
  - 02_prep_stanford_wos.do
  - 03_prep_join_names.do
  - 04_prep_pubmed_panel.do
  - 05_prep_cite_counts_3_5_10.do
- 01_Clean
  - 02_clean_summary.do
  - 02_clean_citation.do
  - 03_clean_refine.do
  - 04_clean_collection.do
  - 05_clean_entity.do
  - 06_clean_pubmed.do
  - 07_clean_altmetric.do
  - 08_clean_jifs.do
  - 09_clean_affiliations.do
  - 10_clean_editor_dates.do
  - 11_clean_pdb_authority.do
  - 12_clean_survey.do
- 02_Build
  - 00_build_lab_controls.do
  - 01_build_pdb.do
  - 02_build_scooped.do

[1] In cases where we replaced outcome variables with zeroes, the number of observations may be inconsistent with these same tables in the paper. This happens when missing values (which reduce the number of observations) get overwritten by zeroes (which increases the number of observations). In general, all results in red-shaded columns should be ignored due to missing data.

  - 03_build_scooped_panel.do
  - 04_build_projects.do
  - 05_build_cite_pairs.do
  - 06_build_reputation.do
  - 07_build_survey.do

- 03_Analysis
  - 00_analysis_paper_stats.do
  - 01_analysis_summary_stats.do
  - 02_analysis_scoop.do
  - 03_analysis_long_run.do
  - 04_analysis_editor_dates.do
  - 05_analysis_pre_deposit_scoops.do
  - 06_analysis_reputation.do
  - 07_analysis_survey.do

3.3 A Note About User-Written Packages

In the master file, we have commented out all of the commands which install user-written packages. These commands should be uncommented as necessary.

In general, user-written packages installed via ssc (the Boston College Statistical Software Components archive) will have the most recent version of the packages installed. Updated packages are overwritten in the ssc archive, so it is not possible to install older versions of packages. In our testing, the results appeared robust to different versions of the user-written commands installed using ssc, but we cannot guarantee that this will be the case across all versions or going forward.

Two packages (lassopack and pdslasso) did lead to result sensitivities depending on the version. Fortunately the authors of these packages have saved older versions of their packages on github. Therefore, we use the net install command to install these packages and specified the versions. We use version 13 of lassopack and version 11 of pdslasso. Using other versions of these packages may lead to errors or slightly different results.

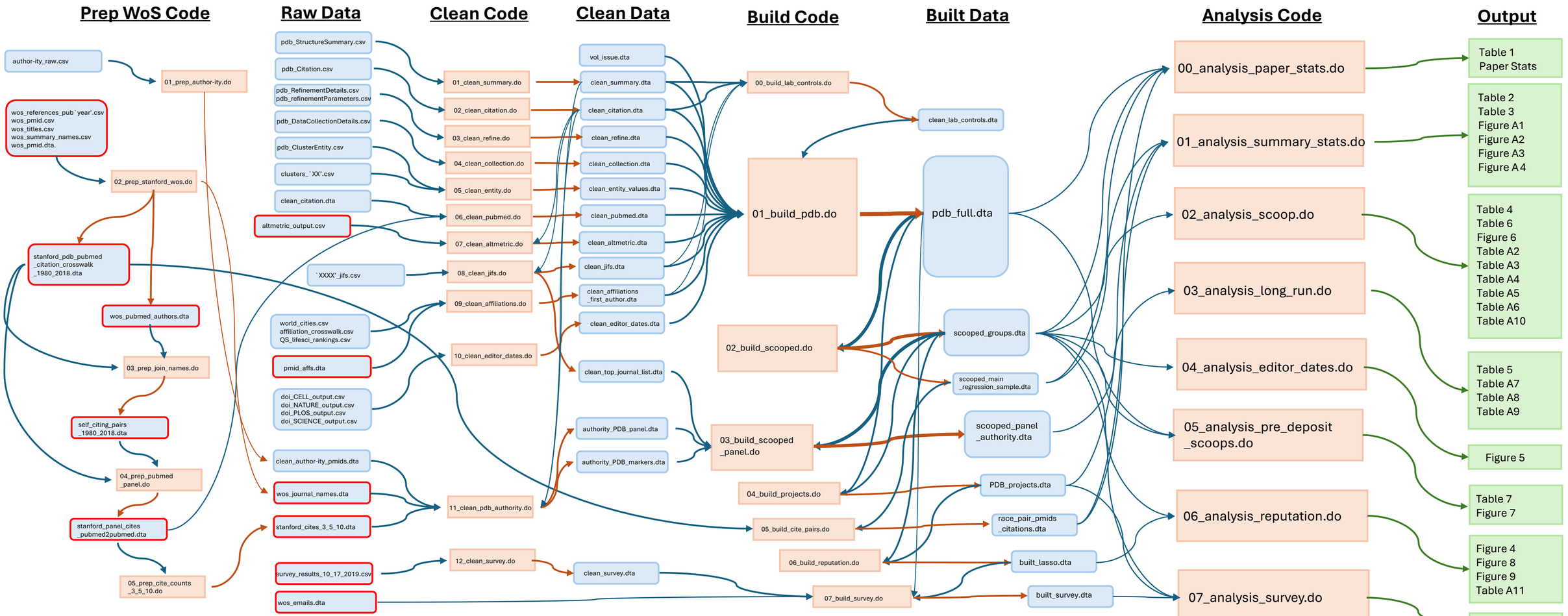
Statement about Rights

We certify that we have legitimate access to and permission to use the data used in this manuscript.

Scooped! Estimating Rewards for Priority in Science

Data Build Schematic

Ryan Hill and Carolyn Stein. 2024



* Red border represents restricted data not provided in replication package