Replication Package README for "Race to the Bottom: Competition and Quality in Science"

Ryan Hill and Carolyn Stein

January 22, 2025

Abstract

This replication package provides data and code to produce the analysis in "Race to the Bottom: Competition and Quality in Science." The data contains a mix of publicly-available and proprietary data sources. All code is provided, and the vast majority of tables and figures can be fully reproduced. However, some results cannot be reproduced without access to the proprietary data. This document provides step-by-step instructions for researchers seeking to reproduce our results. A data and code schematic is provided at the end of this document.

1 Quick-Start Guide

This replication package is subdivided into four folders:

- Code: contains all of the code for the project, including comp_qual_master.do.
- Data: contains all raw data for the project. Some data has been stripped of certain variables due to data sharing restrictions.
- Figures: figures will populate in this folder upon running the code.
- Tables: contains an Excel workbook that will automatically populate. Each table in the paper is on its own tab.

To reproduce our results, open comp_qual_master.do. Uncomment any packages you need to install. Replace the global user_root with the file path where the replication package is stored on your local machine. Run the master file, and all tables and figures should be generated automatically[1] (except for those relying on proprietary data).

2 Data

The data folder has three subfolders.

1. 01_Raw contains the raw data.
2. 02_Clean is empty, but will populate with the clean data once the code is run.

[1] Note: the Excel file of tables cannot update if it is currently open. Make sure to close it before hitting run.

3. 03_Built is (almost) empty, but will populate with the built data once the code is run.
    - There are is one exception: we are providing phat.dta in the 03_Built folder. See Section 3 below for details.

The 01_Raw folder contains a mix of publicly available and proprietary data, which we will discuss below.

2.1 Publicly Available Data

PDB

This is the Protein Data Bank data. It was primarily 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_Citation.csv
- pdb_ClusterEntity.csv
- pdb_DataCollectionDetails.csv
- pdb_RefinementDetails.csv
- pdb_RefinementParameters.csv
- pdb_Sequence.csv
- pdb_StructureSummary.csv
- pdbe_Validation.csv
- scooped_groups.dta[2]

UniProt

The UniProt data consists of a crosswalk between UniProt IDs and PubMed IDs.

- uniprot_pubmed.dta

Survey

Survey data from our survey experiment conducted in May 2023 can be found here. The data contains no personally-identifiable information.

- survey_data_050623.csv

[2]Note: this is clean data file from another project that we merged into this project. Code and details on its construction can be found here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TJ5VCW

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 are easily available, but we cannot provide them without violating the terms of the use agreements. All of these files can be found in 01_Raw/Restricted. We have provided the datasets so that readers can understand the structure of the data, but key variables have been censored.

Web of Science

We used proprietary data owned by Stanford University (curated and provided by Heidi Williams in 2020) that provides panel data on citations counts received by each paper each year. Data can be accessed and purchased 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/). Similar data can be constructed from the open-access resource OpenAlex, though we have not attempted to use these data in this project. More information available here: https://openalex.org/.

- wos_pubmed2pubmed.dta

DrugBank

DrugBank data was downloaded on February 18, 2020. Academic users can download the full dataset for free here: https://go.drugbank.com/releases/latest. However, we cannot post the data publicly as part of the terms of agreement.

- db_drug_index.csv
- db_target_all.csv
- db_target_pharmacologically_active.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 in various subfolders. In addition, we have a subfolder called Restricted. This contains files where we have stripped some (or all) of the variables due to data-sharing restrictions. Running the code will populate the 02_Clean and 03_Built folders within the Data folder. However, one file (phat.dta) in the 03_Built folder is pre-populated. This is because we could not share the underlying raw data for this file, but we can share the (aggregated) built data.

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 insert the file path of the replication package on their local machine where the global user_root is defined. 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 3.3 below on user-written packages).

The do-files are arranged in three folders, and each do-file is called by master.do. The order that the files are run in matters, so users should use the master file to run files in the correct order. 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 files turned off so that the code can run with the available data. Other files will still run, but will output incomplete results due to missing proprietary data. These are marked in blue, and remain turned on in the master file.

The analysis will output figures into the /Output/Figures/ subfolder. The regression estimates are automatically output into sheets in /Output/Tables.xlsx.[3] 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.

- 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.
- 01_Clean
  - 01_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_validation.do
  - 08_clean_drugbank.do
  - 09_clean_survey.do
- 02_Build
  - 01_build_pdb.do
  - 02_build_entities.do
  - 03_build_structures.do
  - 04_build_papers.do
  - 05_define_sample.do

[3]Note that Tables.xlsx must be closed in order to update. Ensure you close this file before running the code.

  - 06_generate_p.do

- 03_Analysis
  - 01_summary_stats.do
  - 02_potential_regressions.do
  - 03_structural_genomics.do
  - 04_competition_regressions.do
  - 05_welfare.do
  - 06_welfare_calculations.do
  - 07_appendix_misc.do
  - 08_phat_bootstrap.do
  - 09_survey_analysis.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.

One exception in our past experience is that lassopack can lead to result sensitivities depending on the version. Fortunately the authors of this package have saved older versions of their packages on github. Therefore, we use the net install command to install this package and specified the version. We use version 13 of lassopack. Using other versions of these packages may lead to errors or slightly different results.

Race to the Bottom: Competition and Quality in Science

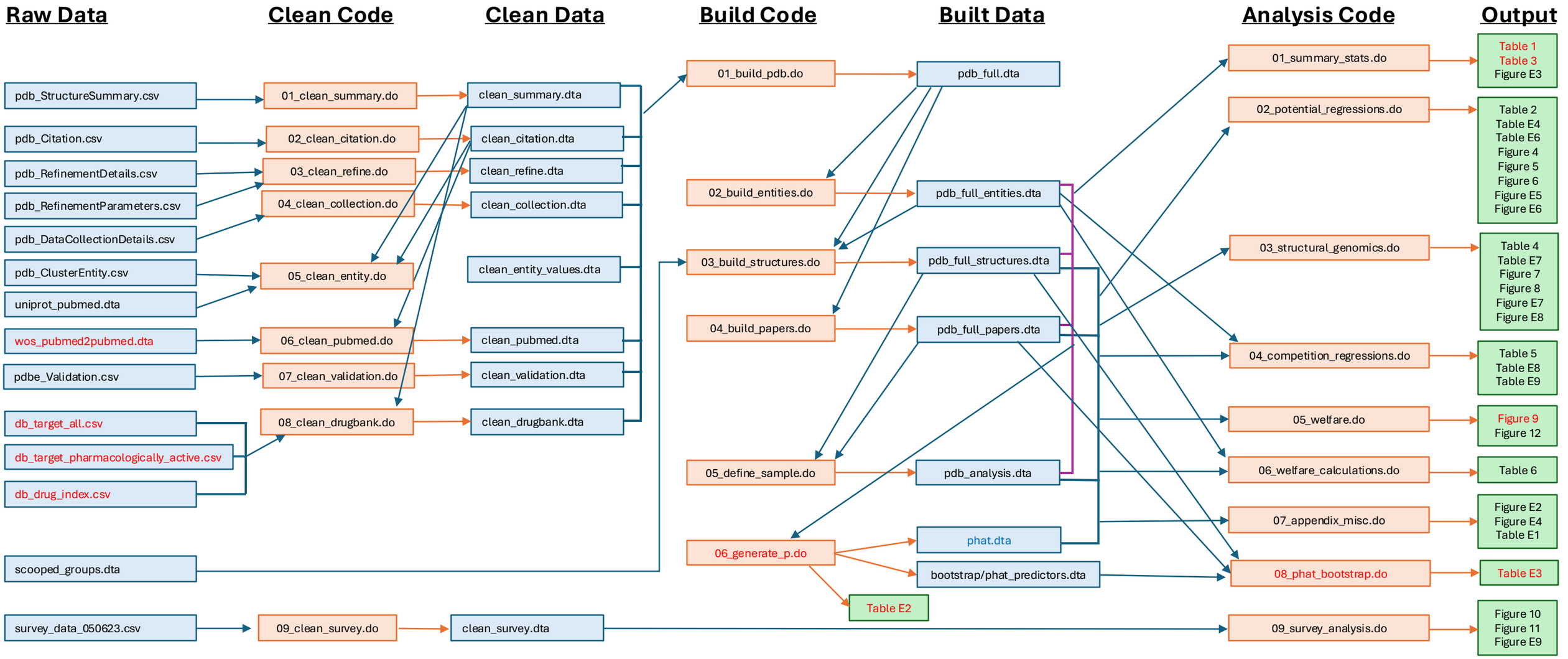
Data and Code Schematic

Ryan Hill and Carolyn Stein, 2025

Notes:

raw data files in red are missing key variables

code files in red cannot run due to missing variables

intermediate data files in blue are provided because they cannot be built from the raw data due to missing variables

output files in red cannot be created in part or in whole due to missing variables