P-hacking in clinical trials and how incentives shape the distribution of results across phases

Authors: Jérôme Adda,[1] Christian Decker,[2] Marco Ottaviani[1]*

Affiliations:

[1]Bocconi University and IGIER.

[2]University of Zurich.

*Correspondence to: marco.ottaviani@unibocconi.it

Replication Package

This Stata replication package reproduces the preparation of the data and our econometric analysis completely and saves as output the graphs and tables displayed in the paper in the respective sub-folders. Stata version 15 or 16 is required for proper creation of the graphs; all other parts of the program run also on Stata 14.

Instructions

1. Unpack the zip-folder, maintaining the folder structure.
2. In section 1 of *0_globals.do*, set your system user name and insert the paths to the data and analysis folders on your local machine.
3. In section 0.1 of all other do-files, set your system user name and insert the path to *0_globals.do* on your local machine.
4. Run the master file *1_master.do*.

Organization

The data-subfolder ***raw*** contains the raw data which the program uses as input. They are gathered from the following sources:

- **AACT database** (as of August 15, 2019), freely available for download at http://aact.ctti-clinicaltrials.org/, where also descriptions of the variables can be found.

- *browse_conditions.csv*
- *interventions_other_names.csv*
- *interventions.csv*
- *mesh_thesaurus.dta*
- *outcome_analyses.csv*
- *outcomes.csv*
- *sponsors.csv*
- *studies.csv*

- **Hand-collected data by the authors**

  (for details on the construction see Materials and Methods section of the paper and the *SI Appendix*)

  - *phase2_interventions_hand_collected.dta* contains the main experimental intervention of applicable phase II trials, hand-collected from information provided in the AACT database and on the online registry https://clinicaltrials.gov/ .
  - *medicareD_spending_collapsed.dta* contains the market sizes by MeSH groups for conditions obtained by classifying drugs manually and summing the total Medicare D spending. These calculations are based CMS medicare drug spending data, freely available for download at https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Information-on-Prescription-Drugs/2015MedicareData.html
  - *mht_adjustent_hand_collected.dta* contains dummies indicating adjustment procedures for multiple hypothesis testing, codified manually from the information provided in the AACT database (especially the variable *p_value_description*) and on the online registry https://clinicaltrials.gov/ .

All the do-files of the different parts of the analysis are gathered in the analysis-subfolder ***doflies***. You can start the whole program by running the master file or you can run individual parts of the analysis separately (note, however, that some files require datasets produced in previous steps). Log-files are saved in the analysis-subfolder ***log,*** graphs in ***graphs***, and tables in ***tables***.

The program saves the processed data in the data-subfolders ***intermediate*** and ***final***. Files with no further indication contain data on primary outcomes of trials, which are the main focus of our study. Files with the postscript '*_sec*' contain data on secondary outcomes. Intermediate files will be delated at the end of the master code. For further analyses, the following files, saved in the ***final*** data folder, might be especially useful:

- *linked_phases.dta* contains the constructed linking of phase II and phase III trials (for details on the construction, see *SI Appendix*).
- *data_selection_functions.dta* contains the restricted dataset of phase II trials used for the estimation of the selection functions.
- *data_counterfactual_analysis.dta* contains the full dataset of drug trials with results used for density estimations, discontinuity tests, and our counterfactual analysis.