Only Section I of this documentation is completed at this point. We are working on the rest and the corresponding folders and expect to finish everything by Q3 2021. Please contact Rembrand M. Koning (rkoning@hbs.edu) if you have questions.

Replication Package Documentation (Data_Cleaning Folder)

I. build_clean_data Folder

This folder contains the code files that process all raw data and produce the clean datasets to be used in our main paper and appendix analyses. Specifically, they are:

- *xml_scrape.R*
  Rscript to scrape data from all XML files including PubMed and the clinical trials (at least 350 GB RAM needed)

- *pubmed_clean.R*
  Rscript cleaning the scraped data for analyses (at least 200 GB RAM needed)

- *prediction_model.R*
  Rscript file cleaning the patent and PubMed text and fitting the Fasttext prediction model.

- *build_data_for_figures_123.do*
  STATA do file creating the datasets for making Figures 1, 2, and 3 in the main paper.

II. raw_data Folder

III. intermediate_data Folder

IV. Raw and Processed Data from Other Sources

Our final datasets are built from several resources. Since some of them are large files and publically available, we do not include them in our replication package and specify their origins below.

- Publication/PubMed Data Files
  - Original XML files
    Can be obtained from the annual baseline bulk download option provided by the National Library of Medicine. (Link: https://www.nlm.nih.gov/databases/download/pubmed_medline.html)
    Note that the raw data are indexed from1 to 1015. We only used the post 300 XML files. Also note that we use the year 2020's version.
  - desc2020.xml
    MeSH tree number information corresponds to each MeSH term provided by the National Library of Medicine (Link: https://www.nlm.nih.gov/databases/download/mesh.html)

Only Section I of this documentation is completed at this point. We are working on the rest and the corresponding folders and expect to finish everything by Q3 2021. Please contact Rembrand M. Koning (rkoning@hbs.edu) if you have questions.

- Patent Data Files
    - Detailed patent description text
      This data source is also from PatentView and contains the detailed description text of all patents we analyzed. We mainly used these datasets for fitting our Fasttext prediction models. The documentation and download can be found here: https://patentsview.org/download/detail_desc_text
    - GBD

- Clinical Trial Related Data