Replication Package Documentation (Analysis Folder)

This folder contains the code, datasets, and Fasttext models we build to produce the tables and figures in our main paper and appendix.

I. build_tables_figures Folder

This folder contains the code files that produce all tables and figures in our main paper and appendix. Specifically, there are three code files in it:

- *analysis_MainPaper.R*
  Rscript file producing Figures 1-4 in the main paper.

- *analysis_Appendix.do*
  STATA do file producing results in appendix section S-13, appendix tables S-1 to S-4, and appendix figures S6 to S8.

- *analysis_Appendix.R*
  Rscript file producing the rest of the appendix results

- *spec_chart_no_dot.R* and *spec_chart_only_dots.R*
  R source function files to make the coefficient plots. These functions are modified per our specific needs. Documentation of the original function is here: https://github.com/ArielOrtizBobea/spec_chart

II. clean_data Folder

This folder contains all datasets and Fasttext models that are directly used in the code files mentioned above. Some datasets are cleaned and derived from raw data sources, and the Fasttext models fitted at the backend. For the detailed cleaning/Fasttext-modeling process, see our data_cleaning folder and the corresponding documentation file *readme_data_cleaning.pdf*.

Below specifies this folder's content:

- **article_data Subfolder**

It contains all datasets used in the code files mentioned above to produce our article-level results. Specifically, they are:

  - *pubmed2002_reg_meshC_types.csv*
    Our full sample article-level regression dataset with 5,529,360 research publications. It contains all publications that are classified as original research. See Appendix S-10 for further information.

- *jcif_1k_regdat.csv*
  Article-level regression dataset only with research articles published on the top 1,000 journals by their journal commercial impact factors (JCIF). This is a subset of *pubmed2002_reg_meshC_types.csv*. See Appendix S-10 for further details.

- *jcif_500_regdat.csv*
  Similar to above, but this only contains articles in the top 500 journals by their JCIF. It is mainly used for producing Table S-40 and is specifically described in Appendix S-12.3.

- *jcif_100_regdat.csv*
  Similar to above, but this only contains articles in the top 100 journals by their JCIF.

- *dm_all_f_1k.csv*
  The exact-matching dataset based on research articles published on the top 1,000 journals by their journal commercial impact factors (JCIF). The matching is for the all female team dummy variable and by each article's publication year, research team size, and focused disease areas. See Appendix S-12.3 for further details.

- *dm_maj_f_1k.csv*
  Similar to above, but this is the matched dataset for the majority female team dummy.

- *dm_min_f_1k.csv*
  Similar to above, but this is the matched dataset for the minority female team dummy.

- *dat_pred.csv*
  Dataset for all article gender focus prediction regressions and tables. It is made from our Fasttext supervised models and *jcif_1k_regdat.csv*. For the detailed cleaning process, see code file *pubmed_clean.R* in the data_cleaning folder. Its specific usage is described in Appendix S-12.4.

- *incidence_reg1k.csv*
  The article-level dataset for all of our disease incidence level analyses. See Appendix 12.1 for more information.

- **patent_data Subfolder**

It contains all datasets used in the code files mentioned above to produce our patent-level results. Specifically, they are:

- *dump_sep17.csv*
  MeSH terms applied to 480 stratified and then randomly selected patents by two trained medical indexers/coders. We use this data for making Appendix tables S-1 to S-4. See Appendix S-2 for further details.

- *female_net_sums.csv*
  Dataset made in the cleaning process and is used for making Figure 3 in the main paper.

- *indexing_sample_weights_and_strata.dta*
  As described in Appendix S-2, we used stratified sampling to select 480 patents for human medical indexers to apply MeSH terms to. This file has the sampling weights that we use to adjust our estimates to reflect the full population of patents in our data.

- *inventor_gender_counts.dta*
  The counts of female and male inventors for each patent. See Appendix S-4 for detail.

- *inventor_invention_counts.csv*
  The annual amount of patents produced by the inventor team's gender composition. See Appendix S-9 for detailed explanations.

- *pat_predicted_female_trial.csv*
  Probability estimates of patent's female focus from our Fasttext supervised model. The Model fitting detail can be found in Appendix S-3.4.

- *pat_predicted_male_trial.csv*
  Probability estimates of patent's male focus from our Fasttext supervised model. The Model fitting detail can be found in Appendix S-3.4.

- *patent.tsv*
  Patent's basic information from PatentView. We use this to obtain each patent's title. See Appendix S-3.2 for its specific usage. (Data and documentation can be found here: https://patentsview.org/download/data-download-tables.)

- *patent_gbd_level.dta*
  Our main patent-level regression dataset. We obtain this dataset by merging our patent information with the Global Burden of Disease database (GBD). We combined the data based on patents' MeSH-identified diseases. In this particular dataset, the MeSH terms are produced by the default MTI. More detail can be found in Appendix S-3.3.

- *patent_gbd_level_ta2500.dta*
  Similar to above but the data are merged based on the MTI MeSH terms generated from the first 2,500 characters of patents' title-abstract text. The dataset is used for building Table S-22. See Appendix S-7 for more details.

- *patent_gbd_level_tas_2500 strict.dta*
  Similar to above but the data are merged based on the MTI MeSH terms generated from the first 2,500 characters of patents' title-abstract text with MTI's strict filtering option.

- *patent_gbd_level_tas_10000.dta*
  Similar to above but the data are merged based on the MTI MeSH terms generated from the first 10,000 characters of patents' title-abstract text.

- *patent_gbd_with_paper_pair_variables.dta*
  This dataset includes our patent data along with measures for the sex-focus of the papers the patent cites. It also includes the sex-focus of the undelrying research in cases where the patent is part of a patent-paper pair where the researchers are listed as inventors on the patent. See S-12.5 for further information.

- *patent_mesh_and_tree.dta*
  Contains the patent-level MeSH terms identified by the MTI and each MeSH term's corresponding tree structure based on the NIH MeSH classifications. More tree structure information can be found here: https://meshb-prev.nlm.nih.gov/treeView. This dataset is used for creating our disease area fixed effects in the patent-level regressions.

- *patent_mesh_and_tree_ta_2500.dta*
  Similar to above but the MeSH terms are identified based on the first 2,500 characters of patents' title-abstract text. This dataset is mainly used for building Table S-22 and is explained in Appendix S-7.

- *patent_mesh_and_tree_tas_2500 strict.dta*
  Similar to above but the MeSH terms are identified based on the first 2,500 characters of patents' title-abstract text with MTI's strict filtering option.

- *patent_mesh_and_tree_tas_10000.dta*
  Similar to above but the MeSH terms are identified based on the first 10,000 characters of patents' title-abstract text.

- *patent_mesh_raw.dta*
  Raw MeSH terms and other results directly produced by the MTI based on patents' text. See Appendix S-1 for further details.

- *patent_mesh_raw_ta_2500.dta*
  Similar to above but the text used is the first 2,500 characters of patents' title-abstract text. This dataset is mainly used for building Table S-22 and is explained in Appendix S-7.

- *patent_mesh_raw_tas_2500 strict.dta*
  Similar to above but the text used is the first 2,500 characters of patents' title-abstract text with MTI's strict filtering option.

- *patent_mesh_raw_tas_10000.dta*
  Similar to above but the text used is the first 10,000 characters of patents' title-abstract text.

- *patent_tas_2500.tsv*
  A dataset cleaned from PatentView's raw data. We keep the first 2,500 characters of each patent's title-abstract text. We use this dataset as the input of our supervised models to predict each patent's gender focus. See Appendix S-3.4 for further details.

- *patents.csv*
  A dataset containing patent description text for making appendix tables S1 to S4. See Appendix S-2 for further details.

- *pats_human_sex_review_team_gender.csv*
  This file contains the authors' coding of whether patents are "actually" male- or female-focused for patents with different sex-tag combinations. This data is used to generate Table S-7 and the data generating process is described in detail in S-3.2

- **fasttext_models Subfolder**

This folder contains the Fasttext models and word vectors we fit based on the patent detailed description text and the PubMed abstract text. See S-3 for further details. The .bin and .vec files represent Fasttext model and word vector files, respectively. All modeling processes can be found in our code file *prediction_model.R* in the data_cleaning folder. These Fasttext results are used to build our gender prediction regressions. Specifically, the content includes:

- *fasttext_model_segment1, 2, and 3*
  Because of the file size, we split our main Fasttext model into three segments. To retrieve the model, one can combine the 3 segments using the cat command via terminal/command line. The command, specifically, is

cat fasttext_model_segment1 fasttext_model_segment2 fasttext_model_segment3 > fasttext_model.bin

- *fasttext_model.vec*
  Main model's word vector file for fitting the supervised models mentioned below.

- *supervised_model_female.bin* and *supervised_model_male.bin*
  The supervised models built based on *fasttext_model.vec* and the clinical trial text for obtaining each article's/patent's female and male focus probabilities.

III. tables_figures Folder

This is an empty folder to be used for storing the tables and figures generated by our code.