# Replication Package Documentation (General)

## I. patent_gender.csv and paper_gender.csv

They are the datasets that have the primary variables we used in the patent and paper analyses. Variables include:

- patent_id/PMID: unique patent and PubMed article IDs
- patent_female/paper_female: dummies that classify if a patent/paper is linked to the MeSH term “female”.
- patent_male/paper_male: dummies that classify if a patent/paper is linked to the MeSH term “male”.
- female_trial_est/male_trial_est: the confidence level that a paper/article is female-focused/male-focused based on our Fasttext model.
- incidence_female/incidence_male: average female/male-related disease incidence count associated with a given patent/paper based on its MeSH term(s).

## II. data_cleaning Folder

This folder contains our code to clean the raw data and the cleaned data that are used in our main paper and appendix analyses. For detailed descriptions of this and its subfolders, see the folder’s documentation readme_data_cleaning.pdf.

## III. analysis Folder

This folder contains all datasets and code files we built to produce our main paper and appendix results. With this folder alone, one should be able to replicate all our presented findings. For detailed descriptions of this and its subfolders, see the folder’s documentation readme_analysis.pdf.