This archive contain replication files for “**Predictive modeling of US healthcare spending in late life**” by Einav, Finkelstein, Mullainathan, and Obermeyer.

The archive only includes the program code, but not the underlying data files that are needed to run the code. To obtain access to the data files, interested scholars would need to apply to the Centers for Medicare and Medicaid Services (CMS). The included code is designed to access these data via a Subversion repository. Scholars wishing to replicate our results will need to obtain their own copy of the data via an independent Data Use Agreement with CMS. The code will need to be updated to reflect any differences between ours and the researcher’s own copy of the raw data, and to point to the appropriate Subversion repository or storage location.

Interested scholars should feel free to e-mail us for any additional assistance with replication.

## 1. How to Use.

Our code directory is divided into a handful of main directories:

- Directories in /raw/ store raw data (or symbolic links to raw data), with no code or code that performs minimal cleaning
- Directories in /derived/ take data as inputs (often from /raw/, but also from /derived/ or /derived_local/) and produce data as output
- Directories in /derived_local/ also take data as inputs (again, often from /raw/, but also from /derived/ or /derived_local/) and produce data as output
- Directories in /analysis/ take data as inputs (from /derived/ and /derived_local/) and produce results as output
- Directories in /lib/ store code libraries that are loaded and used by other directories

Within each of these main directories, there are additional subdirectories, each of which is self-contained. That is, *after updating the code to reference the researcher’s own Subversion repository and storage locations*, the code can be run without inputs from outside the directory. This is accomplished by constructing symbolic links or version-controlled references to datasets or files necessary for analysis.

Most of these directories have a similar structure:

*/code/*

Contains all the scripts necessary to execute the directory. In most cases, there will be a Stata do file or a SAS script with macros that perform the main functionality.

*/code/externals.txt*

Lists the inputs that are needed to run the directory, calling versions of data or programs saved in a Subversion repository. Items referenced in this file will be populated in /externals/. This file can be used to locate other portions of code that create datasets used in an analysis.

*/code/links.txt*

References datasets that are too large for version control and so are stored locally on the machine. Items referenced in this file will be populated as symbolic links under /external_links/ in the directory. This can also be used to locate other portions of code that create datasets used in an analysis.

*/code/make.py*

Executes all the code in the directory. Other researchers will not be able to successfully run this code without (1) access to the data and (2) updating externals.txt and links.txt to point to the researcher's own Subversion repositories and storage locations. However, one can use it as documentation for the correct order in which the other scripts are being executed.

*/externals/*

Contains data, programs, and other inputs that are necessary to execute the directory. The file /code/externals.txt lists its contents, and it is created and populated by make.py.

*/external_links/*

Contains symbolic links to large datasets that are necessary to execute the directory. The file /code/links.txt lists its contents, and it is created and populated by make.py.

*/output/* and */output_local/*

Contains log files, txt files with tables, pdf files with figures, csv files with data, or dta files with Stata datasets. This is created and populated by make.py.

In this replication archive, the directories analysis/, derived/, derived_local/, lib/, and raw/ live under the /EOL_code/ directory. When parsing the list of externals or links under /code/externals.txt or /code/links.txt, you will see, for example, /trunk/analysis, /trunk/derived/, or /shared_data/derived_local/. These will be found under the corresponding folder under /EOL_code/ in the replication archive.

## 2. Building from the raw data.

The project is built from a series of raw Medicare claims files and the Medicare denominator file. We provide the code that unzips and organizes the Medicare denominator data from the format given by CMS obtained under DUA 22559. This is the code found under /raw/Medicare Denominator (20 pct)/.

Many links.txt files also contain paths to the raw Medicare claims files. These are the paths that begin /disk/aging/medicare/data/20pct/*XXX*. To replicate our results, these paths would need to be updated to the researcher's own path to the raw Medicare claims files.

## 3. Constructing Predictors.

There is a parallel set of directories that create the predictors for, separately, the January 1st, 2008, sample and the Inpatient Event sample. The directories that create the predictors for the January 1st sample are all found under the /derived/ directory:

- /Jan 1 Chronic Conditions/
- /Jan 1 Gagne/
- /Jan 1 HCC Scores/
- /Jan 1 Utilization Predictors/
- /Jan 1 ZO Cats/
- /Jan 1 Physician Visits/

In some cases, other inputs will need to be created first (in other directories) before a given directory can be run successfully. The inputs that a given directory requires can be found in externals.txt and links.txt.

The Inpatient Event sample predictors are generated by a parallel set of directories with the prefix "Event" in place of "Jan 1".

## 4. Constructing Outcomes.

There is a parallel set of directories that create the outcomes for, separately, the January 1st, 2008, sample and the Inpatient Event sample. These directories are found under /derived/:

- /Jan 1 Backfill Spending/
- /Jan 1 Backfill Visits, Days, Procs/
- /Jan 1 Combine Spending w Backfill/
- /Jan 1 Combine Visits, Days, Procs w Backfill/
- /Jan 1 Spending/
- /Jan 1 Visits, Days, Procs/

In some cases, other inputs will need to be created first (in other directories) before a given directory can be run successfully. The inputs that a given directory requires can be found in externals.txt and links.txt.

The Inpatient Event sample outcomes are generated by a parallel set of directories with the prefix "Event" in place of "Jan 1".

## 5. Samples.

There is a parallel set of directories to combine the constructed predictors into a single data set and split the data into separate samples for training, calibrating, and testing. These directories are found under /derived_local/:

- /ML Samples Jan 1/

## 6. Tuning Random Forest and Xgboost.

The code for tuning the Jan 1 random forest and the linear gradient boosted regression trees is found under /analysis/:

- /RF Tune Jan 1/
- /Xgboost Tune Jan 1/

## 7. Prediction Algorithm.

There is a parallel set of directories for estimating our ensemble probabilities. These directories are found under /analysis/:

- /Full Ensemble/

## 8. Exhibits Crosswalk.

The directories in /analysis/ populate a spreadsheet under /analysis/Liran/ named exhibits.xls. The following reports which directories in /analysis/ produce which exhibits, as well as the name of the sheet in exhibits.xls that contains the data for that exhibit.

### *a. Main Paper*

Figure 1 – /Spending shares/, sheet E#1 in exhibits.xls

Figure 2 – /Spending by phat/, sheet E#2 in exhibits.xls

Figure 3 – /Share of total spending by phat/, sheet E#3 in exhibits.xls

Figure 4 - /Spending by phat/, sheet E#4 in exhibits.xls.

### *b. Appendix*

Figure 1 – Schematic (no code)

Figure 2 – Theoretical relationship from equation (no code)

Figure 3 – /Distribution of phat/, sheet E#B4 in exhibits.xls

Figure 4 – /Share of total spending by phat/, sheet E#E1 in exhibits.xls

Figure 5 – /Spending shares/, sheet E#C3 in exhibits.xls

Figure 6 – /Spending by phat/, sheet E#C4 in exhibits.xls

Figure 7 – /Share of total spending by phat/, sheet E#C5 in exhibits.xls

Figure 8 – /Spending by phat/, sheet E#C6 in exhibits.xls

Table 1 – /Summary Statistics/, sheet E#A1 in exhibits.xls

Table 2 – /Descriptive statistics on phat/, sheet E#B5 in exhibits.xls

Table 3 – /Variable Importance/, sheet E#B6 in exhibits.xls

Table 4 – /Summary Statistics/, sheet E#C1 in exhibits.xls

Table 5 – /Descriptive statistics on phat/, sheet E#C2 in exhibits.xls

Table 6 – /Ex ante perspective/, sheet E#D1 in exhibits.xls