Supplementary Material for "Beyond Statistics: The Economic Content of Risk Scores"

Liran Einav, Amy Finkelstein, Raymond Kluender, and Paul Schrimpf

July 10, 2015

This is a readme file for the supplementary material that accompanies "Beyond Statistics: The Economic Content of Risk Scores" by Liran Einav, Amy Finkelstein, Raymond Kluender, and Paul Schrimpf. Please contact any of us (leinav@stanford.edu; afink@mit.edu; kluender@mit.edu; schrimpf@mail.ubc.ca) with any questions.

The enclosed material includes program files that generate all the results reported in the paper. Note that the data used for the paper is confidential and cannot be made publicly available. They were obtained under DUA #22559 through NBER (PI: Amy Finkelstein). You must contact CMS (http://www.resdac.org) to apply for access to the data. If you are an NBER affiliate, contact Jean Roth at NBER: jroth@nber.org.

The program files are organized in two subfolders, the content of which is described in more detail below. In the Descriptive Analysis subfolder, we include Stata do files that clean the data and produce the results reported in the paper. In the Risk Scores folder we include all the files that are used to generate the risk scores used in the baseline sample.

Descriptive Analysis Subfolder

DataPreparation_final.do

This file reads in the raw beneficiary, plan, and claim files for years 2007-2009 from CMS and ResDac and creates the baseline file that we use for analysis in our project. It goes through the following steps:

1) Cleaning Beneficiary files - it takes in raw beneficiary files, and limits the file to beneficiaries enrolled in Medicare Part D. It then cleans and labels various variables. We also generate variables for beneficiary birth month, join month, and dummies for whether a beneficiary switches plans during each year, and for whether he or she receives any cost-sharing.

2) Cleaning claims files - it takes in raw claims files with observations at the claims level. It collapses these files at the beneficiary level, creating variables for the dollar amount and number of claims at each phase of the standard Part D plan and for categories of drugs. We also generate monthly claim totals, average fill size, weekly spending variables, and the final phase that each beneficiary ends up in each year.

3) Cleaning plans files - Renames and cleans various variables within the plan data.

4) Merging 1), 2), and 3) above - we merge the beneficiary, claims, and plan files that have been cleaned above, and drop observations corresponding to claims unmatched to beneficiaries, claims not matched to plans, and plans without any claims. We drop any beneficiaries not in the strict 20% Medicare sample, and beneficiaries not in the US.

5) Appending each of the year-level datasets - We append the 2007-2009 year-level datasets generated by 4) above, then drop any duplicate beneficiaries randomly. We then drop beneficiaries who join for a reason other than turning 65, dual Medicaid eligibles, beneficiaries with listed 3rd party cost-sharing, non-stand alone Medicare Part D plans, beneficiaries who switch plans within the year, and beneficiaries who die within the year.

6) Creating baseline analysis variables - we take the dataset created by 5) and create variables for plan type, the yearly standard ICL, the yearly standard deductible, the difference between beneficiary spending and the ICL, deductible, and categorical spending amounts. We calculate beneficiary cost-sharing averages within each phase, and merge in risk scores. This is the baseline file, in which each observation is a beneficiary-plan-year for the years 2007-2009. We also generate a wide version of the baseline file for use in the cross-year substitution analysis.

The final dataset which is prepared by this program is called "bene070809_baseline.dta", a long dataset in which an observation is a beneficiary-plan-year for years 2007- 2009 but also includes claims and mortality for the first 6 months of the next year (through 2010).

MultidimensionalAnalysis_final.do

This file reads in the baseline data file created above and generates Figures 2, 3, and 4, and Table 1. It begins by estimating some descriptive summary statistics. It then generate predicted incidence rates through the kink of age, gender, risk score, and hierarchical condition categories, then compares and plots the predictions against the actual incidence rates through the kink region.

Risk Scores Subfolder

These files must be run in the order they are listed below:

Contents.sas, DataMacros.sas, UtilityMacros.sas

Various macros for reading in data that are used to facilitate programming in the SAS files below. These files are not independently run.

BeneDataPrep.sas

This file reads in raw beneficiary data from Medicare from 2006-2010 and cleans the demographic and enrollment information, to input into SAS risk score programs.

DiagnosesDataPrep.sas

This file reads in raw diagnoses from Medicare from 2006-2010 and cleans the diagnosis names and ICD- 9 code information, to input into SAS risk score programs.

RunRxRiskScores20xx.sas

These files create RxHCC risk scores in SAS files for each respective year, based on CMS programs and formulas that are in subfolder RxHCC_2012software. Please see the CMS documentation in this folder for more detail on the risk score software.

RunMARiskScores20xx.sas

These files create MA HCC risk scores, based on CMS programs and formulas that are in subfolder CMSHCC_2012software_MA_Cost. Please see the CMS documentation in this folder for more detail on the risk score software.

Specialists.sas & Specialists_Collapse.do

These files create a crosswalk between the beneficiaries in baseline sample and the specialties of the physicians that they are seeing, for use in the Medicare Parts A and B utilization file.

CodingMedicareUtilization.sas

This file reads in Medicare Part A and B data, and data on specialists that our beneficiaries see, to create variables on the utilization of Parts A and B and the death dates of our beneficiaries who die during 2006- 2010, to merge into the baseline data.

cleaning_scores.do

This .do file reads in the risk score and utilization files generated above for each year, and then appends the yearly files to create 4 output files:

A) MA_riskScores_baseline – HCC risk scores for our baseline beneficiaries in 2006-2009
B) RiskScores_baseline –RxHCC risk scores for our baseline beneficiaries in 2006-2009
C) utilization_tomerge – Utilization of Medicare Parts A and B for our baseline beneficiaries in 2006-2010.

D) death_dates_full – Death dates of beneficiaries that died in our sample.

These four files are produced and are merged into the baseline sample at the end of the "DataPreparation_final.do" program.