Data and Method

Huanye Liu

Model

First, we introduce the Bayesian network model used in this project. To detect the over-prescription behavior, we need to first operationalize the concept of over-prescription. In this project, it is defined as the relatively large number of a target drug prescribed by one doctor compared with the normal number or expected number of the target drug the doctor should have prescribed, and it can also be defined as the relatively large Medicare payment on procedures. We detect the over-prescription or the outliers from data based on the small posterior probabilities using Bayesian statistics to capture the probabilistic dependency among different variables. Among many outliers/anomaly detection computational tools, Bayesian network model can directly map the relationship between multiple variables to a graph, and there are also ready-to-use algorithms for model learning and model parameter inference.

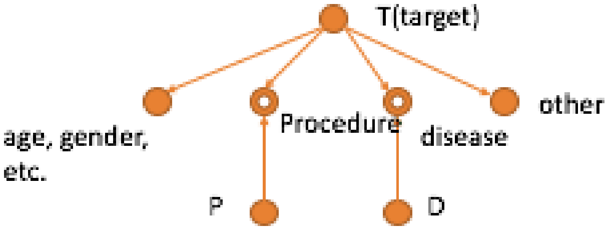


Figure 1: Bayesian network

From the above figure, we can see that what we need to compute is the posterior probability of P(T|age, gender, Procedure,disease, other). Here the target variable can be the target type of drug, or it can also be the medicare payment on a medical procedure, depending on how to define the overprescription behavior. If the posterior probability is below the threshold we choose, then we say this is an anomalous record.

In addition, if the target variable is the drug type, we also use the dynamic programming to accumulate each individual posterior probability of using a target drug prescribed by one doctor to detect the doctor???s over-prescription of this drug. More specifically, we need to calculate the probability that the expected number of prescriptions for the target drug based on the model inference is greater that the number of the observed ones, and if that probability is small, we would classify this particular prescription as over-prescription. ## Data

The data we use for this project can be downloaded from the Centers for Medicare and Medicaid Services website: https://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems.html, which are stored in the Basic Stand Alone Outpatient Procedures Public Use Files aggregating information from 2010 Medicare outpatient claims. These files are procedure-level files in which each record corresponds to a procedure in an outpatient claim incurred by a 5% sample of Medicare beneficiaries. These files also include some demographic and claim-related variables which will be explained in detail later.

The original data source is the 100% Beneficiary Summary File for the reference year, and CMS Basic Stand Alone Outpatient Procedures Public Use Files consists of records of 5% simple random sample of beneficiaries drawn without replacement from the original data. There is no overlap between the sample used for the CMS Basic Stand Alone Outpatient Procedures Public Use Files and the existing 5% CMS research sample in terms of the beneficiaries in the CMS Basic Stand Alone Outpatient Procedures Public Use Files and the 5% CMS research sample. There is also no overlap between the sample dataset used for the CMS Basic Stand Alone Outpatient Procedures Public Use Files and other Basic Stand Alone Public Use Files such as CMS 2008 Inpatient Claims PUF, CMS 2008 PDE PUF, CMS 2008 DME Line Items PUF, and CMS 2008 Hospice

Beneficiary PUF, CMS 2008 SNF Beneficiary PUF, CMS 2008 HHA Beneficiary PUF, and CMS 2008 Carrier Line Items PUF, all of which have already been released so far.

More specifically, there are approximately 2.5 million beneficiaries included in the 5% sample data, and all medicare claims with related information about those 2.5 million beneficiaries are collected, and the are further transformed to represent one procedure of a claim for one record instead of one claim for one record. About 1.2 million had claims out of those 2.5 million beneficiaries in the 5% sample, resulting in a public use file of 33,600,194 procedures. The file contains six anaytic variables in addition to the procedure key: patient's age, patientls gender, ICD-9 primary diagnosis code, HCPCS procedure code, count of number of services, and Medicare payment for procedures. To prevents users from linking data across multiple files for identification purposes, procedures cannot be linked by claim or beneficiary, and they cannot be linked to any external data source by means of the procedure ID which is a cryptographic key only applying to this Outpatient Procedures public use file but not available elsewhere.

Here are brief descriptions of each analytic variables of the public use file dataset used in this project:

1) Gender (BENE_SEX_IDENT_CD): The beneficiary's gender, (1) male or (2) female.

2) Age (BENE_AGE_CAT_CD): The beneficiary's age, reported in six categories: (a) under 65,(b) 65-69, (c) 70-74, (d) 75-79, (e) 80-84, (f) 85 and older.

3) ICD-9 primary diagnosis code (OP_CLM_ICD9_DIAG_CD): "International Classification of Diseases" version 9.3 This is a three-digit code. 869 such codes are observed in the dataset.The variable is blank (or missing) when there does not exist a primary diagnosis on the procedure. There are only 2 such records in the PUF.

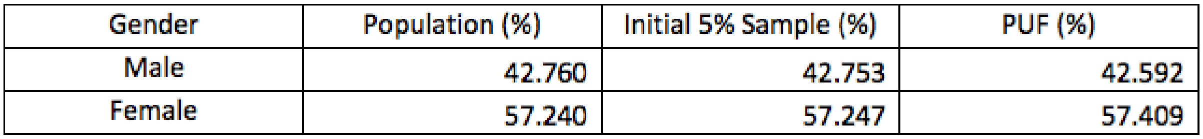
4) HCPCS procedure codes (OP_HCPCS_CD): These are HCPCS codes (HCPCS Level I and Level II) and take on 3,867 possible values in the dataset.

5) Count of services (OP_HCPCS_UNIT_CNT): This is the count of the total number of services associated with the procedure. This variable is not rounded. There are some records in the dataset for which count of services is equal to zero (0).

6) Medicare payment on procedures (OP_HCPCS_PMT_AMT): The amount paid by Medicare for the associated procedure.

7) Procedure count (PROC_CNT): Total number of procedures associated with the record(i.e., combination of the analytic variables).

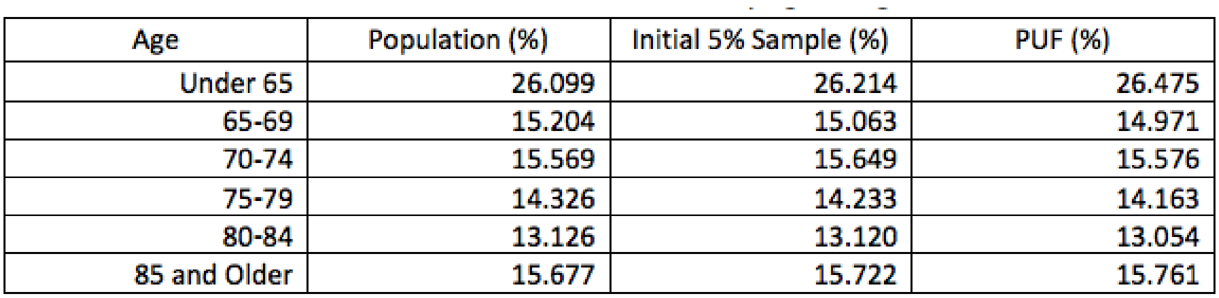
Here are some summary statistics for key variable mentioned above.Figure 3 and Figure 4 compare the distribution of procedures by gender and age of beneficiaries in the CMS 2010 Basic Stand Alone OP Procedures Public User File, the initial 5% sample and the entire population. The figures presented in those tables indicate that the Public User File, despite only 5% of the procedures in the original sample, provides good estimates of population parameters:



Gender	Population (%)	Initial 5% Sample (%)	PUF (%)
Male	42.760	42.753	42.592
Female	57.240	57.247	57.409

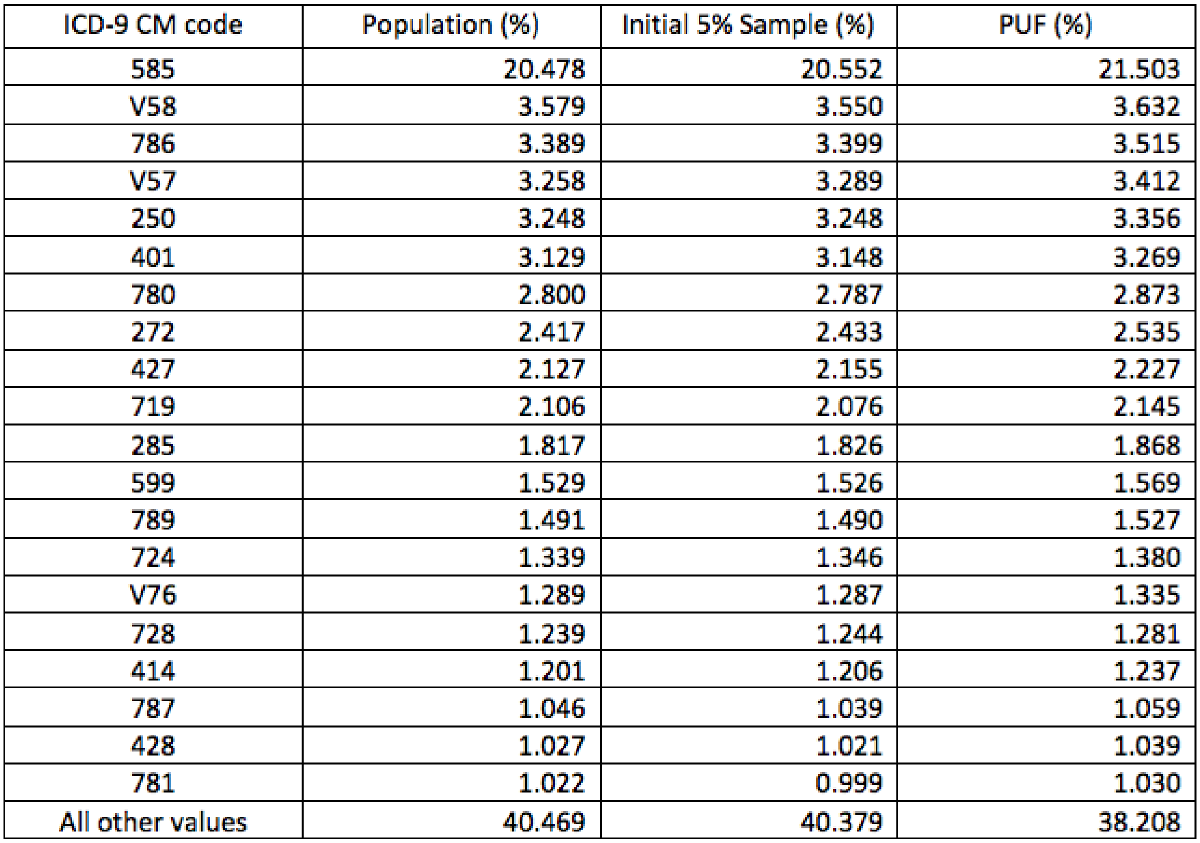
Figure 2: Distribution of Procedures by Gender of Beneficiary

and Figure 4 and Figure 5 provide the same comparison for the ICD???9???CM diagnosis codes and HCPCS codes:



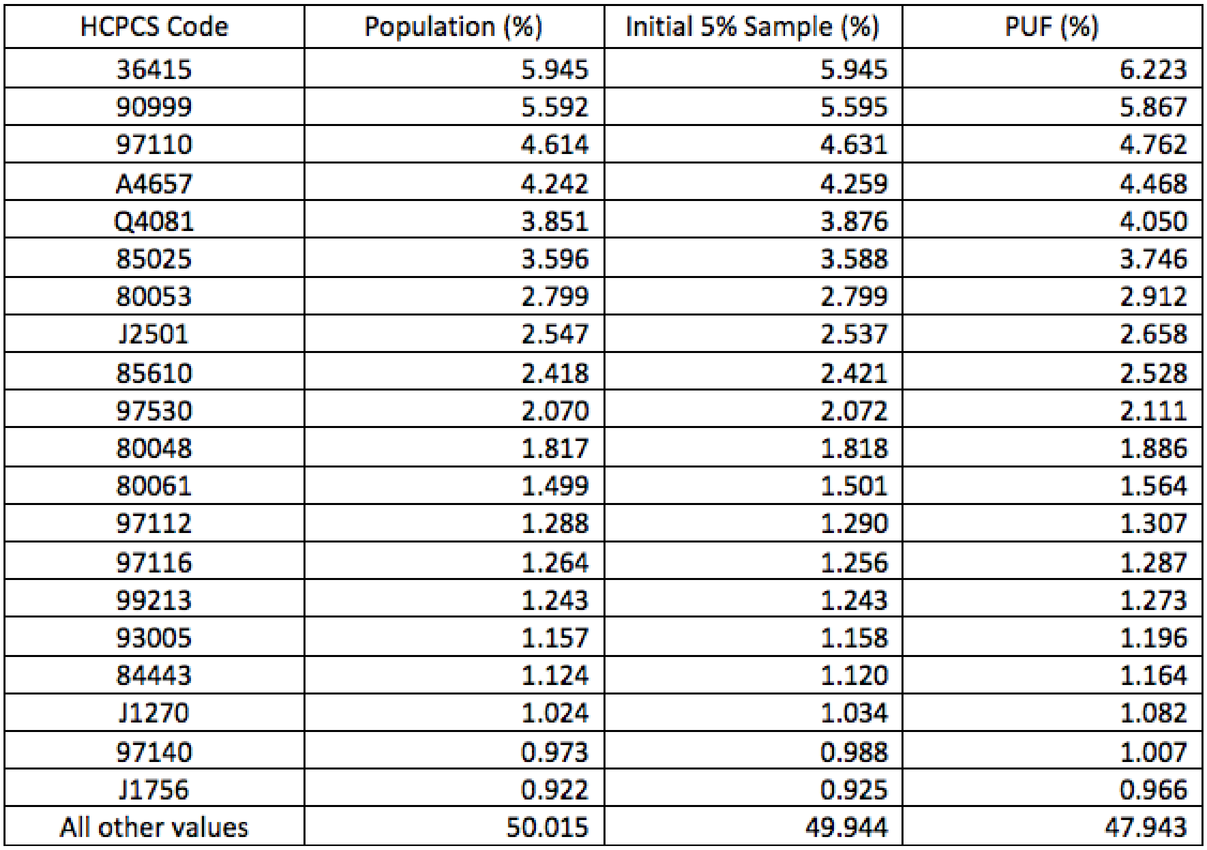
Age	Population (%)	Initial 5% Sample (%)	PUF (%)
Under 65	26.099	26.214	26.475
65-69	15.204	15.063	14.971
70-74	15.569	15.649	15.576
75-79	14.326	14.233	14.163
80-84	13.126	13.120	13.054
85 and Older	15.677	15.722	15.761

Figure 3: Distribution of Procedures by Age Categories



ICD-9 CM code	Population (%)	Initial 5% Sample (%)	PUF (%)
585	20.478	20.552	21.503
V58	3.579	3.550	3.632
786	3.389	3.399	3.515
V57	3.258	3.289	3.412
250	3.248	3.248	3.356
401	3.129	3.148	3.269
780	2.800	2.787	2.873
272	2.417	2.433	2.535
427	2.127	2.155	2.227
719	2.106	2.076	2.145
285	1.817	1.826	1.868
599	1.529	1.526	1.569
789	1.491	1.490	1.527
724	1.339	1.346	1.380
V76	1.289	1.287	1.335
728	1.239	1.244	1.281
414	1.201	1.206	1.237
787	1.046	1.039	1.059
428	1.027	1.021	1.039
781	1.022	0.999	1.030
All other values	40.469	40.379	38.208

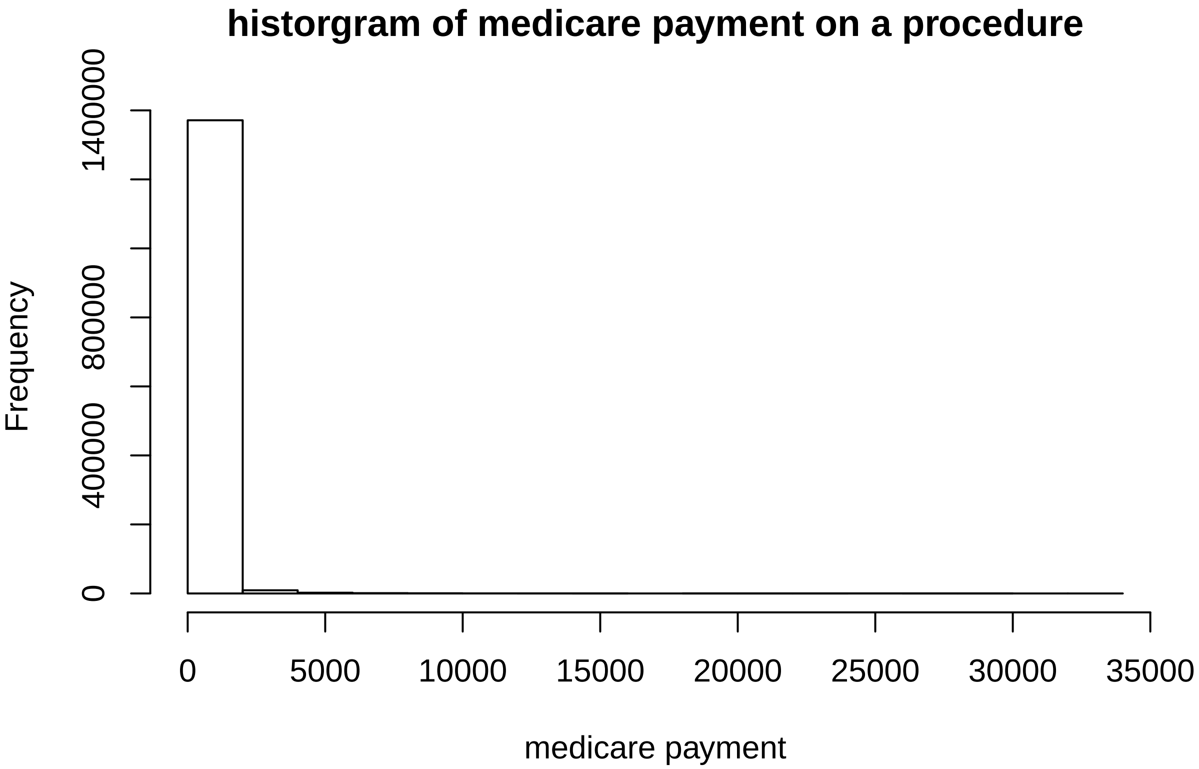
Figure 4: Distribution of Procedures by ICD???9???CM Diagnosis Codes



HCPCS Code	Population (%)	Initial 5% Sample (%)	PUF (%)
36415	5.945	5.945	6.223
90999	5.592	5.595	5.867
97110	4.614	4.631	4.762
A4657	4.242	4.259	4.468
Q4081	3.851	3.876	4.050
85025	3.596	3.588	3.746
80053	2.799	2.799	2.912
J2501	2.547	2.537	2.658
85610	2.418	2.421	2.528
97530	2.070	2.072	2.111
80048	1.817	1.818	1.886
80061	1.499	1.501	1.564
97112	1.288	1.290	1.307
97116	1.264	1.256	1.287
99213	1.243	1.243	1.273
93005	1.157	1.158	1.196
84443	1.124	1.120	1.164
J1270	1.024	1.034	1.082
97140	0.973	0.988	1.007
J1756	0.922	0.925	0.966
All other values	50.015	49.944	47.943

Figure 5: Distribution of Procedures by HCPCS Codes

Initial Results



```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      0.0    10.0    30.0   141.1    80.0 33000.0
```

If we quantify the overprescription behavior as the relatively large medicare payment on procedures, we could first of all check the distribution of medicare payment ignoring all other related variable. We can see from the histogram and the summary of medicare payment variable above that there are some unusual large values, the max of which is 33000, way higher than the median and mean.

Therefore, next step we need to use the Bayesian network to incorporate all other ralated variables to capture the relationship between those variable, to see that if those extremely high medicare payment are normal or abnormal conditioning on all other related variables.