DSS DATAWAREHOUSING ETM 538

Winter 2017

FINAL PROJECT

DEVELOPING REQUIREMENTS FOR THE OREGON ALL-PAYER-ALL-CLAIMS (APAC) DATABASE

Instructors: Dr. Mike Freiling, Dr. Daniel Sagalowicz

Team:

Marek Szumowski

Lei Wei

Qing Lai

Devdeep Aikath

ETM OFFICE USE ONLY

Report No.:
Type: Student Project
Note:

Introduction

Project selection and Interview

For our group project, we sought a database that was sizeable enough to necessitate advanced Data mining techniques while being common enough that the implications of the data and the business questions are clear to the audience. Specifically, two of our team members have past or current links to the healthcare industry and that field seemed appropriate.

One of our team members, working as a Graduate Intern within the Data and Information Management Enhancement (DIME) Department at Kaiser Permanente (KP), arranged for an interview with Mr. Jeff Emch, MBI , a senior analyst at KP responsible for collecting and sending claims data to the State health agency, the Oregon Health Authority (OHA). He offered the perspective of both the preparer of this data as well as the recipient of this data, having considerable experience understanding the latter's needs. During the interview, he role-played both the persona as a KP analyst and an OHA analyst- which helped us develop an all-round understanding of the project.

The sponsor OHA, their business goals, and the APAC Database

For the purpose of our project, the sponsor is the OHA and their branch organization Office of Health Analytics (HA). The OHA's mission of "Helping people and communities achieve optimum physical, mental and social well-being through partnerships, prevention and access to quality, affordable health care" informs the higher level questions that they ask and answer using the claims data they receive and analyze through the HA. In summation, the OHA's goals are to improve quality of healthcare while optimizing costs. Consequently, it examines all the claims data from the healthcare providers and insurance agencies to examine all the medical cases that were diagnosed, treated and billed.

Previously, claims data used to be separately housed in a more needs-based, disorganized manner. In 2009, the Oregon State Legislature established the Oregon All Payer All Claims Database through a House Bill, "authorizing the formation of a healthcare data reporting program to measure the quality, quantity, and value of healthcare in Oregon"[1]. It includes medical and pharmacy claims, enrollment data, premium information and provider information for Oregonians receiving coverage through both commercial insurance agencies as well as public agencies such as Medicare and Medicaid. The database contains information for about 80% of Oregonians, roughly 3.2M individuals.

While OHA maintains oversight and management of the APAC, an external agency, Milliman, Inc.. is contracted to collect and process the data. The data is secured with the provisions of the HIPAA and restricted subsets are released for public usage and research.

Key business questions of the OHA

Prior to the passing of the APAC program by Oregon State Legislature, OHA needed to better understand the type of questions they needed to answer in order to determine the type data to be collected from the insurers. Through our interview, we were able to pyramid these questions from high level issues to be addressed, to lower level

[1] https://www.oregon.gov/oha/analytics/APACPageDocs/APAC-Overview.pdf

questions reached by drilling further down into the details of the data.

At the highest level, the questions are:

1. **Quality of care:** Are the Oregon residents receiving good quality of care? Can that level of quality be improved? How can that improvement be achieved?
2. **Cost of care:** Is the health care in Oregon cost effective? Can this cost efficiency be improved? How can efficiency that be implemented?

If we go deeper, specific issues are covered at the next highest level, and the questions become specific to a disease condition or diagnosis or a population subset, finally narrowing down to a question that can directly answered by mining the dataset. We have used this hierarchical classification to tie specific questions to the broader business goals, since it is beyond the scope of this report to produce an exhaustive list of business questions that the OHA asks on a regular basis. The following list is merely a representation that we will use, first to find the relevant attributes and bucketing suggestions, and then to test whether our algorithms can find the corresponding answer.

- Are children being treated to maximize long-term health?
    - Are children <2 yrs being vaccinated?
        - Are children <2yrs with low-income families being vaccinated
            - Are <2yr-children from minority population being vaccinated on first visit?
- Are pre-diabetic patients getting proper preventive therapy?
    - Are patients with high HBA_{IC} levels getting proper preventive therapy
        - Are high-HBA_{IC} patients receiving insulin?
            - Are high-HBA_{IC} patients being counselled re insulin use on each visit?
- Which diseases are most costly for the State?
    - Which are the costliest methods of intervention? (= ER visits)
        - Which diagnoses are associated with maximum state (Medicaid and Medicare) payout?
            - Which diagnoses require most ER visits?
                - Trend of ER visits (recurrence, seasonality, clustering)
- Can we better optimize the treatment coding for cost-effectiveness?
    - Which treatments are always grouped together?
        - How many times are all the component of these groups necessary?
            - Will suggested specific alternative grouping reduce costs?
- Can we improve treatment outcome and save costs by preventive treatment?
    - Which diagnoses have specific pre-diagnostic conditions/tests?
        - What are the compounding factors for the diagnosis? (=smoking:lung ca)
            - Can these compounding factors be prevented by pharmacological/behavioral intervention?
                - What are the costs of non-smoking counseling/how much can be invested for that purpose?

Concepts the OHA uses/requires

The premise of the APAC data and its uses lie in Association learning. There is no specified class, and the goal is to find "interesting structures"[2] in the data. We can expect to find some association rules in the APAC data without even examining the actual dataset. For example,

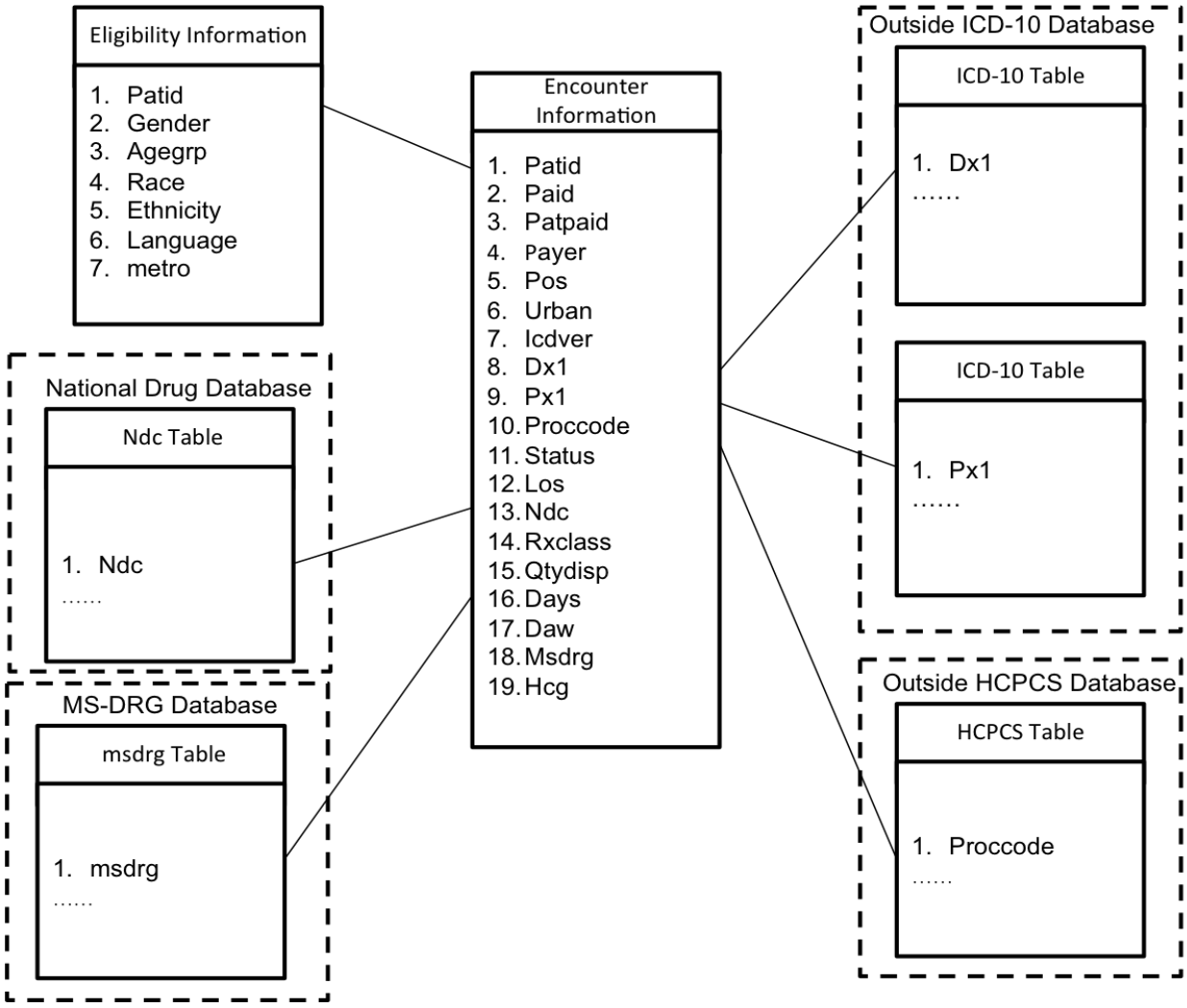
If pos (Place of service) = 20 (ER), then paid (payment amount) is likely to be high.

[2] Witten, Frank and Hall (2010), Data Mining

Here, we are trying to predict more than one attribute (as addressed in Table 2 below).

In terms of relational concepts, the APAC database heavily uses foreign keys to link with external databases like the ICD-10. In fact, prior to the APAC implementation, the distributed claims data prevented the OHA from being able to link spending and member volume to specific outcomes.

Figure 1 The inter-relationship of various internal tables within the APAC database and external databases (dotted lines)



Attributes to answer the questions and suggested bucketing:

This a comprehensive list of all the attributes that are included in the APAC Database. The subsections designate the table/data section the attributes belong to (also shown in figure 1). The suggested bucketing for the continuous variables (except the foreign keys) are also included. The specific attributes to answer individual questions is given in table 2.

Table 1: Attributes and bucketing

Eligibility information (pertains to the resident's demographic information)		
Code	**Description**	**Bucketing**
patid (FOREIGN KEY)	Unique member key for a person who is or was enrolled in a health insurance plan. The member is the person who has received the service. The same individual has a unique patid for each insurance plan.	None
gender	Member's gender.	Done
agegrp	Member's age range in years.	5 yr buckets (e.g., 1-5)
race	Member race.	None
ethnicity	Member ethnicity.	None
language	Member language.	None
metro	Indicates if the member's address is within a Metropolitan Statistical Area or a non-Metropolitan Statistical area as defined by the Office of Management and Budget.	Categorical
Encounter information (pertains to a specific visit to a provider facility)		
Code	**Description**	**Bucketing**
patid (FOREIGN KEY)	Unique member key for a person who is or was enrolled in a health insurance plan. The member is the person who has received the service. The same individual has a unique patid for each insurance plan.	NONE
paid	Amount in dollars that was paid by the payer to the service provider for the service.	$100s up to $1000 $1000s up to $10,000 $10,000s up to $1M
patpaid	Amount patient paid.	$100s up to $1000 $1000s up to $10,000 $10,000s up to $1M
payer	Line of business category of the payer that paid the claim, e.g. Medicaid fee-for-service, Medicare Advantage, etc.	Categorical
pos	Industry standard place of service code. i.e. 20 = Urgent Care Facility, 21 = Inpatient Hospital, 34 = Hospice, etc.	None

urban	Identifies whether the member's ZIP code is associated with a list of Oregon urban ZIP codes.	Into counties
icdver	Indicates whether or not the claimline has ICD 10 (or higher) codes.	None
status	Two-character code that represents the disposition of the patient upon leaving the facility. If the patient died this event may be indicated here.	Died/Survived
los	Length of stay as reported by data submitter	Weeks up to 4 wks Months up to 1 year Years above that
qtydisp	Quantity of the prescription that was dispensed.	Weeks or Months
days	Number of days that the drug will last if taken at the prescribed dose.	Weeks
daw	Indicates if the physician has or has not authorized a substitution for the prescribed drug. 'Y' indicates the drug is to be dispensed as written; 'N' indicates a substitution is permissible.	None

Foreign keys (links to an external database)	
Code	**Description**
ndc	National Drug Code is a unique product identifier for drugs, e.g. i.e. 00006011731 = SINGULAIR, 00025152531 = CELEBREX, etc. NDCs are assigned by the US Food and Drug Administration.
rxclass	Grouping of drugs with the same therapeutic properties as defined by Medi-Span. It is the first 10 characters of Medi-Span's Generic Product Identifier (GPI), e.g. 4927002510 = *ULCER DRUGS* , 0120001010 = *PENICILLINS*, or 2810001010 = *THYROID AGENTS*.
msdrg (FOREIGN KEY)	MedInsight Medicare Severity Diagnostic Related Group Code is the MedInsight derived MS-DRG code. The MS DRG is a Medicare grouping system that classifies inpatient hospital services into one of approximately 750 groups. The codes in this column are for MS DRG version 25 and above. i.e. 864 = FEVER.
dx1 (FOREIGN KEY)	Main or principal diagnosis ICD code associated with the service. ICD is the International Statistical Classification of Diseases and Related Health Problems that classifies diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease, e.g. 95909 = FACE & NECK INJURY, 78652 = PAINFUL RESPIRATION, etc.

px1 (FOREIGN KEY)	Main or principal surgery ICD code associated with the service. ICD is the International Statistical Classification of Diseases and Related Health Problems that classifies diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease, e.g. 0331 = SPINAL TAP, 9921 = INJECT ANTIBIOTIC, etc.
ecode (FOREIGN KEY)	Supplies the first diagnosis code that begins with an "E". Reviews up to 13 diagnoses codes on each claim to determine if an E code exists.
proccode (FOREIGN KEY)	American Medical Association's Current Procedural Terminology (CPT) code, the Healthcare Common Procedure Coding System (HCPCS) code, or the Common Dental Terminology (CDT) code for the service. i.e. 90471 = IMMUNIZATION ADMIN, 80061 = LIPID PANEL, or 74170 = CT ABDOMEN W /O & W /DYE.

Strategies and Algorithms used ms to Answer Questions:

Here, we try to address each of the 5 high-level questions stated earlier and the detailed lower-level questions they generate using the APAC database. In the table below, we give the attributes and algorithms that are needed to answer each of the question cohorts.

The guidelines we developed and adopted to make decisions re algorithm of choice are given in table 3.

Table 2: Answering specific questions (note that the highest level question is non-specific and cannot be answered using the data without specifying further).

Questions	Strategies and Algorithms used
● Are children being treated to maximize long-term health? ● Are children <2 yrs being vaccinated? ● Are children <2yrs from minority populations being vaccinated? ● Are <2yr-children from minority populations being vaccinated on first visit?	● Query related attributes among different dataset by foreign keys. Here: Query attributes Age group (agegrp) and race from Eligibility Information table and Proccode from Encounter Information table by foreign key Patid. ● Do the calculations by using “group”, “if” and “and” function.
● Are pre-diabetic patients getting proper preventive therapy? ● Are patients with high HBA_{IC} levels getting proper preventive therapy ● Are high-HBA_{IC} patients receiving insulin? ● Are high-HBA_{IC} patients being counselled re insulin use on each visit?	● Query related attributes from Outside ICD-10 Database by foreign key Dx1 and related attributes from Outside HCPCS Database by foreign key Proccode ● The above attributes inform the diabetic/pre-diabetic status based on normal vs high HBA_{IC} levels from the Dx code as well as insulin treatment information from Proccodes ● Do the calculations by using group, if and and function.
● Which diseases are most costly for the State?	● Query related attributes from Table Encounter Information and connect them with Outside ICD-10

● Which are the costliest methods of intervention? (= ER visits) ● Which diagnoses are associated with maximum state (Medicaid and Medicare) payout? ● Which diagnoses require most ER visits? ● Trend of ER visits (recurrence, seasonality, clustering)	Database by foreign key Dx1 and related attributes from Outside HCPCS by foreign key Proccode ● The information above pulls the place of service (pos) and the diagnosis (Dx1) and looks for patterns. ● Logistical regression can be used to examine probability of ER visit (DV) for the diagnosis type (factor or IV) ● Do the calculations by using group, if and and function.
● Can we better optimize the treatment coding for cost-effectiveness? ● Which treatments are always grouped together? ● How many times are all the component of these groups necessary? ● Will suggested specific alternative grouping reduce costs?	● Query related attributes from Table Encounter Information and connect them with Outside ICD-10 Database by foreign key Dx1 and related attributes from Outside HCPCS by foreign key Proccode ● Again, here Linear modeling can find the least cost achieved by examining and comparing treatment costs resulting from different types of grouping. ● Use unsupervised Learning K-Means to find appropriate groups for treatments. We should note that the contracted database analysts, Milliman, Inc. already uses a grouping algorithm called the Medical Episode Grouper. More information is given in Appendix (table A1)
● Can we reduce treatment costs by preventive treatment? ● Which diagnoses have specific pre-diagnostic conditions/tests? ● What are the compounding factors for the diagnosis? (=smoking:lung ca) ● Can these compounding factors be prevented by pharmacological/behavioral intervention? ● What are the costs of non-smoking counseling/how much can be invested for that purpose?	● Query related attributes from Table Encounter Information and connect them with Outside ICD-10 Database by foreign key Dx1 and related attributes from Outside HCPCS by foreign key Proccode ● Use association rule to guess missing data and eliminate dependent variables. ● Use Linear modeling to explore relationship between IVs and Dv and develop prediction models.

Algorithms Evaluation:

We compared the eight different kinds of algorithms learned in class (see Table 3: Algorithm Comparison) to preliminarily evaluate which algorithms we may use. Based on the analysis of Table 3, we chose the combination of three algorithms. We use Association Rule to guess missing data and eliminate dependent variables, use unsupervised Learning K-Means to find appropriate clusters, and use Linear/Logistic regression to build models based on the IVs and DV.

Table 3: Algorithm Comparison

Algorithms	Brief Introduction	Choose (Y/N)
1R	1. Simple. Use a single attribute with the most predictive power 2. Null values can be treated as values	Probably. Use it to preliminarily explore the important 1 single variables that influence the prediction of dependent variables.
Entropy and Iterative Algorithms	1. Based on Information of each node to develop Decision Trees 2. For highly branching useless attributes: Choose attributes that maximizes gain ratios.	No. Different split of the training data can lead to different trees.
Naïve Bayes	1. Values are (probabilistically) independent of each other. 2. All attributes are equally important 3. Missing data can be removed from calculation 4. Careful attribute selection makes it more reasonable by eliminating attributes that show too much dependency	Probably. This algorithm in combination with Association Rule can avoid using highly mutually dependent attributes.
Instance Based Classification	1. No rules or trees are created. Key Metric: Distance and Neighborhood 2. Similar instances are "combined" to form answer. 3. Fast setup time. Very slow run time. Works well on complex cases. Highly dependent on judgment	No. Long run time. Even use K nearest neighborhood method to help reduce run time. The choice of K will influence prediction and also subjective. Categorical variables don't work well.
Covering	1. Key metrics: Support and Confidence (probability). Maximum confidence and Maximum support (when confidence is equal). 2. Iterative Step (eliminated chosen cases). 3. Each rule stands on its own – independent of order.	Probably. Use it to explore the best rules although some of the rules might be overfit.
Association Rule	1. Identify relationships between attributes 2. Help to guess missing values, and eliminate variables that can be inferred from others (e.g.,Beer and diapers) 3. Key Metrics: Same as for covering algorithms	Yes. Use it to guess missing data and eliminate dependent variables.
Linear/Logistic regression	1. Linear regression is generally used for continuous variables. 2. Assumptions of LR: errors are statistically independent, errors are normally distributed, and error distributions have same standard deviation. 3. Logistic regression works well for categoric variables	Yes. These two algorithms can well supplement each other and use for categorical and continuous variables.

	by changing them into boolean/binary variables.	
Unsupervised Learning K-Means	1. Find the number of clusters by finding their centers and variances. 2. Need to know k 3. Local minima: 4. High dimensionality of the space 5. Lack of mathematical basis	Yes. Use this method to find appropriate clusters.

APPENDIX

Table A1: Medical Episode Grouper codes used by Milliman, Inc.

Code	Long name	Description
megcode	MEG code	Medical Episode Grouper (MEG) episode code. MEG is a proprietary grouping algorithm that creates episodes that describe a patient's complete course of care for a single illness or condition.
megdesc	MEG description	Medical Episode Grouper (MEG) episode description.
megnum	MEG episode number	Medical Episode Grouper (MEG) unique identifier for a single episode.
megdays	MEG episode duration in days	Medical Episode Grouper (MEG) duration of episode in days.
megprorate	MEG prorated episode count	Medical Episode Grouper (MEG) prorated episode allowed amount allocation for the given service line. This field allows a user to sum detail lines for an overall episode count. Summing this field over all related service lines for a given episode will yield a result of 1.
megoutlier	MEG outlier status	Medical Episode Grouper (MEG) indicator for an outlier episode.
megsys	MEG body system	Medical Episode Grouper (MEG) system in the body.
megstg	MEG stage	Medical Episode Grouper (MEG) stage of the given episode.
megtype	MEG type of care description	Medical Episode Grouper (MEG) type of episode, e.g. Acute, Chronic and Well Care
megcomp	MEG episode completion	Medical Episode Grouper (MEG) indicator that episode is complete.