Clustering Healthcare Costs by Disease

November 10, 2017

1 Project Purpose

This thesis will find groupings of medical issues, whether expected or unexpected, that our sample population members have claimed through their insurance.

2 Project Importance

All insurance companies, regardless of the kind of insurance they offer, do their best to predict the future by comparing current clients to historical clients' information. Any statistically significant connection, regardless of expectations and hidden factors, can help to actuarially model future behavior. Any suggested clusters, or groupings, will be explored in light of current medical knowledge and otherwise statistical significance.

This project will be an innovative application of existing Bayesian methods of cluster analysis to known healthcare information. The majority of existing statistical literature focuses on healthcare performance, not the maladies that demand healthcare in the first place.

3 Project Overview

3.1 Data Overview

This data comes from an undisclosed healthcare insurance provider and includes all paid claims from 1 July 2012 to 30 June 2013. In that time, the company paid claims for 6,113,838 people. Each claim includes the person's age, gender, claim amount paid, kind of treatment, and episode start date. The first few rows, as an example of the data, are shown in Table 1.

ETG, or "Episode Treatment Group", is based on a commercial classification of possible medical treatments. Technically there are six digits, not four, but we remove the last two (more specific) digits to protect patient privacy. For example, ETG 711901 (Major joint inflammation - foot & ankle) becomes, for us, a more general ETG 7119 (Major joint inflammation). One book[1] explains,

[1]Bottle, Alex, and Paul Aylin. *Statistical Methods for Healthcare Performance Monitoring.* CRC Press, 2016.

Table 1: Persons 1-3

ID	Age	Sex	ETG	Total Amount	Episode Start Date
1	41	F	1622	927.85	01JUL2012
1	41	F	4383	2145.76	21MAY2012
1	41	F	6351	7676.34	01JUL2012
2	31	M	4388	543.60	01JUL2012
2	31	M	7791	628.39	05JUN2013
2	31	M	7794	598.89	05JUN2013
3	40	M	3999	362.00	16APR2013

An episode can belong to only one patient, but a patient can be in multiple episodes at the same time. Episodes are defined and grouped by proprietary software in the United States for purposes such as value-based purchasing initiatives to improve the quality of care while avoiding unnecessary costs. The market-leading method for grouping such episodes is the Optum™ Symmetry® Episode Treatment Group® (ETG) for medical and pharmacy claims data. It adjusts for disease severity and associated costs and has some similarities with diagnosis-related groups such as its ability to create resource-homogeneous groups. One key difference is that it cuts across healthcare sectors, so it covers inpatients, outpatients and ancillary services. An ETG captures medications, diagnostic information (comorbidities and complications) and procedures.

So the data indicates that in the recorded year, person 1, a 41-year-old woman, had three paid health insurance claims, two of which began before the recording date. Her insurance paid $927.45 toward her hyper-functioning thyroid gland treatment, $2,145.76 toward her acute bronchitis treatment, and $7,676.34 toward her treatment for conditions associated with infertility. We have similar data for 6,113,837 other people.

These ETGs are grouped into 22 MPCs, or Major Practice Categories. Aggregate claim information, grouped by MPC, is shown in Table 2.

3.2 Methodology Overview

We will first sort the data into a matrix with individual people on the rows, and ETGs as columns. Each person will have a 2 in a ETG column where they received at least one claim payment, and a 1 in every other ETG column. Each person will represent a single Monte Carlo draw, with natural clusters where they filed claims. We will then use several methods of finding the Bayesian posterior cluster models, i.e. identifying final clusters based on given information. These methodologies have been aggregated and explained by Fritsch and Ickstadt[2], and implemented in the R package mcclust. The following subsections

[2]Fritsch, Arno; Ickstadt, Katja. Improved criteria for clustering based on the posterior similarity matrix. Bayesian Anal. 4 (2009), no. 2, 367–391. doi:10.1214/09-BA414. https://projecteuclid.org/euclid.ba/1340370282

Table 2: Claim Aggregates by MPC

Major Practice Category	Claim Count	Average	Maximum
Cardiology	1,633,554	$3,445.24	$2,533,176.32
Chemical dependency	123,686	$1,669.96	$1,645,904.58
Dermatology	2,496,394	$7,579.95	$4,072,522.25
Endocrinology	2,026,801	$1,697.56	$766,141.35
Gastroenterology	1,219,131	$3,328.24	$491,260.35
Gynecology	834,650	$5,437.56	$2,375,594.92
Hematology	187,517	$956.20	$749,383.32
Hepatology	100,151	$4,071.38	$4,357,839.25
Infectious diseases	278,211	$740.40	$1,740,124.22
Isolated signs & symptoms	387,158	$2,429.96	$3,443,890.40
Late effects, environmental trauma & poisonings	93,615	$4,460.83	$2,846,706.29
Neonatology	53,307	$12,455.76	$2,014,169.82
Nephrology	73,389	$34,142.21	$3,244,267.32
Neurology	581,348	$3,024.77	$1,080,813.86
Obstetrics	100,783	$14,624.49	$2,878,626.48
Ophthalmology	1,155,398	$4,385.07	$1,494,069.63
Orthopedics & rheumatology	2,612,979	$882.96	$2,080,023.16
Otolaryngology	3,230,160	$3,820.86	$3,558,499.99
Preventive & administrative	3,908,909	$11,320.53	$2,968,575.27
Psychiatry	773,879	$485.96	$507,713.88
Pulmonology	1,072,092	$3,378.84	$4,774,041.97
Urology	751,337	$493.49	$287,014.67

describe these methodologies, labeled by the package function name as released by Fritsch and Ickstadt.

3.2.1 MCLUST criteria

A benchmark criterion that focuses finding each individual cluster probability π_{ij}, given in Eq. 1:

$$\pi_{ij} = P(c_i = c_j|y) \approx \frac{1}{M} \sum_{m=1}^{M} I_{c_i^{(m)}=c_j^{(m)}} \tag{1}$$

3.2.2 MAP criteria

MAP clustering, or Maximum a posteriori, finds cluster estimate $\hat{c}$ by maximizing the posterior density.

3.2.3 MedvComp criteria

Developed by Medvedovic et al., this is a method to obtain $\hat{c}$ by finding $1 - \pi_{ij}$, i.e. the probability that observations i and j are not clustered together.

3.2.4 MinBinder criteria

This criteria minimizes Binder's loss, the posterior expectation of which is given in Eq. 2.

$$E(L(c^*, c)|y) = \sum_{i<k} \left| I_{c_i^*=c_j^*} - \pi_{ij} \right| \tag{2}$$

i.e. the sum of absolute deviations of the estimated similarity matrix to the posterior similarity matrix, or a matrix that containsthe pairwise probabilities that two observations belong to the same cluster.

3.2.5 MPEAR criteria

MPEAR maximizes the Posterior Expected Adjusted Rand, shown in Eq. 3.

$$\frac{\sum_{i<j} I_{c_i^*=c_j^*} \pi_{ij} - \frac{\sum_{i<j} I_{c_i^*=c_j^*} \sum_{i<j} \pi_{ij}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_{i<j} I_{c_i^*=c_j^*} + \sum_{i<j} \pi_{ij}\right] - \frac{\sum_{i<j} I_{c_i^*=c_j^*} \sum_{i<j} \pi_{ij}}{\binom{n}{2}}} \tag{3}$$

where π_{ij} is estimated from the MCLUST Eq. 1.

3.3 Analysis of Resulting Clusters

As previously stated, once we have analyzed the data for any groupings, we will compare the correlations to existing medical knowledge. We might reasonably expect, for example, that obesity would be linked with diabetes, but current medical knowledge may not link Alzheimer's disease with viral pneumonia. These are hypothetical examples of connections we may find when we analyze the health insurance data.

4 Qualifications of Thesis Committee

Faculty Advisor: Brian Hartman
Dr. Hartman received his B.S. in Actuarial Science from BYU and his Ph.D. in Statistics from Texas A&M. He is an Assistant Professor at BYU and the Actuarial Program Director in the Department of Statistics. His research interests include Bayesian methods and their applications in actuarial science and risk. He has worked in various capacities with companies in property-casualty, health, and long-term care insurance,[3] which is why we have the healthcare data available. I took Stat 274 and will take Stat 475 from Dr. Hartman, both

[3]from hartman.byu.edu

of which are actuarial courses intended to prepare students for upcoming actuarial examinations.

Faculty Reader: Robert Richardson
Dr. Richardson received his B.S. in Statistics from BYU and his Ph.D. in Applied Math and Statistics from UC Santa Cruz. He is an Assistant Professor at BYU and researches, among other things, Bayesian nonparametrics with applications in actuarial science.[4] I took Stat 377 and Stat 477 from Dr. Richardson, both of which are also actuarial courses.

Honors Coordinator: Del T Scott
Professor and Undergraduate Advisor in the BYU Department of Statistics

5 Project Timeline

Because I intend to finish classes in April and begin working, though officially graduating in June, I aim to finish all thesis requirements before April 25, 2018. I will likely defend my thesis in the beginning of April.

6 Funding

No funding is required for this research.

7 Culminating Experience

I intend to present our findings in a professional conference setting. If the results of our analysis prove to be sufficiently groundbreaking, we may be able to publish them in actuarial literature.

[4] from richardson.byu.edu