SOCIETY OF ACTUARIES

Predictive Analytics 2014 Call For Articles

Predictive Analytics: It's the process of using modeling and data analysis techniques on large data sets to discover predictive patterns and relationships for business use. From all corners of the world, across many different types of companies and practice areas, predictive models have emerged to help guide business decisions and opportunities. The techniques have also quickly developed far beyond the well-known examples of technology-dependent business models like Amazon and Netflix. More than ever, it has been financial services firms, healthcare providers and all levels of government agencies getting into the predictive analytics mix. Yes, even today, ESPN has on air their own predictive analytics expert.

Over the past years, the actuarial profession has been actively advancing the use of predictive analytics methods in its work. Actuaries have been fast to collect and transform big data into useful information to glean future tendencies and patterns. The results have had a strong impact as companies are able to use actuarial predictive modeling results to optimize their business practices. In addition, actuaries bring to the table a unique skillset that can understand the importance and complexities of preparing the data, combined with the business insights to use the resulting models.[1]

In the following pages, we've selected examples of predictive analytics from a recent Call for Articles issued by the Society of Actuaries. The articles give insights on the growing number of ways the actuarial profession is using these methods in affecting business decisions. While just the tip of the iceberg, these tangible examples demonstrate the expanding ways that actuaries are putting predictive modeling techniques into practice.

Enjoy the articles that follow, let us hear about additional ideas for future exploration, and prepare for the continuing evolution of actuarial predictive analytics.

R. Dale Hall, FSA, MAAA, CERA, CFA
Managing Director of Research
Society of Actuaries

1 For more information on the importance of preparing data, see Steve Lohr: "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights." *New York Times.* Aug. 18, 2014

SOCIETY OF ACTUARIES

Predicting Emergency Room Frequent Flyers

Joseph Randazzo
Vice President,
Healthcare Practice
d-Wise Technologies, Inc.

J. Patrick Kinney,
FSA, MAAA
VP, Enterprise
Financial Planning
Excellus Health Plan

Predictive modeling can help identify people who are more likely to utilize the emergency room many times per year. These patients are often referred to as "frequent flyers". For this project, frequent flyers are defined as three or more ER visits in a calendar year. We analyzed health plan membership, claims, and care management data to determine predictive factors for such frequent emergency room utilization. Increased awareness of these factors among primary care clinicians may offer opportunities to provide needed care in more appropriate and cost-effective settings. This type of analysis may find useful application in Accountable Care Organizations and similar quality- and cost-sensitive arrangements.

Data

We used member-specific demographic, product, medical and pharmacy claims data to create models to identify health plan members with higher risk of becoming a frequent flyer. Commercial, Medicare, and Medicaid lines of business were compared to see if each had similar or different factors contributing to a member's high ER utilization. Additionally, care management data (assessments, care plans, personal health profiles, etc.) was integrated into the Medicare and Medicaid models to study the impact on model performance.

This study is based on 2009-2012 data extracted as of December 2013. Frequent flyers are flagged based on their total number of emergency room claims in a calendar year. Demographic data is tied to a member's last month of eligibility in 2009 through 2012.

Overall, between 2-2.5% of the health plan's members were frequent flyers each year. However, close to 5.5% of Medicaid members become an ER frequent flyer each year. Frequent flyers are expensive, contributing to over 25% of all emergency claims expenses each year.

Model

We used a limited dependent variable model with a binary flag for Frequent Flyers as the target (i.e. dependent) variable to determine which cost, utilization, demographic, product or other factors play a role in who became an emergency room frequent flyer. Logistic regression (without interactions), decision tree, and neural network models were developed to analyze the data. Because the emergency room frequent flyer event is quite small, it was necessary to oversample so that a frequent flyer event made up 10% of the data set. Validation of the models was performed by using separate training (40%) and validation (60%) data sets. The results from the logistic regression are presented in this brief article

Models were created for each line of business using a base year to predict the likelihood of a member becoming a frequent flyer in the following year. Members must have been continuously enrolled in both the base year and the prediction year to be included in the model. For Medicare and Medicaid, a second model was developed that incorporated data from the care management system[1]. Members must have been continuously enrolled and in one or more care/disease management programs in the base year to be included in the model. As with any model using health data, it is necessary to bucket or group medical and pharmacy claims into higher level diagnostic classifications. In our study, medical claims are bucketed into 25 major diagnostic categories (MDC) and pharmacy claims into 64 national drug code (NDC) groupings. *Exhibit 1* shows the variables that consistently had the most impact in order of relative importance. [2]

Results

The goal of the frequent flyer model is to generate chase lists targeting members for intervention. An understanding of model performance statistics provides insight into the likely business impact of using the model's results. As such, the precision statistic is important because it estimates the percentage of targeted members who are true frequent flyers. Sensitivity relates to the model's ability to identify actual high utilizers. A model with high sensitivity is associated with low Type II errors.

Based on the precision statistics for the study, the model's prediction is correct for roughly 70% of the targeted Commercial frequent flyers, 64% for Medicaid, and 68% for the Medicare population. Sensitivity ranges from 27% to 36%. The availability of care management data improves the precision and sensitivity in the model, though the number of members studied drops by approximately 50% each year since many members are not in a care management program in the base year.

Business impact

The average annual emergency room claim expense for a frequent flyer in 2012 was $1,531 ($2,030 Commercial; $1,549 Medicaid; $1,382 Medicare). The potential reduction in claims costs from successfully intervening on a percentage of the predictive model's 'true positives' can be substantial. With a 10% intervention rate (i.e. reducing the number of frequent flyers by 10%), estimated gross savings are $5.12 million.

---

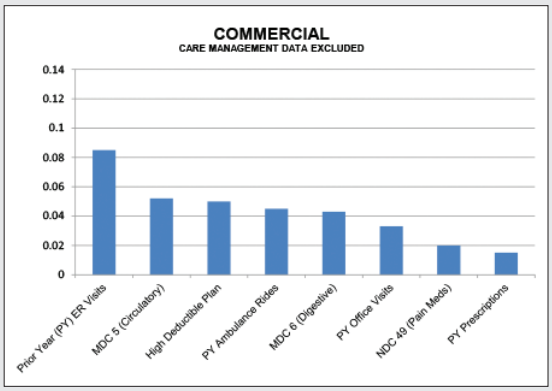
1 The rate of care management in the commercial population was too low to make any modeling effort credible.

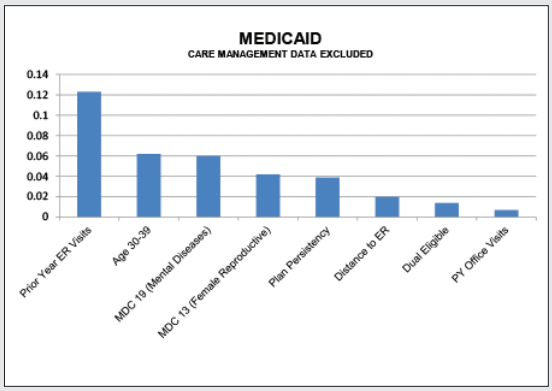
2 For simplicity, because many models were run for different time periods, only the variables that showed up as significant in most or all years are shown. On each graph, 'Variable Importance' is the average of the sequential r-square contribution across models for differing time periods.

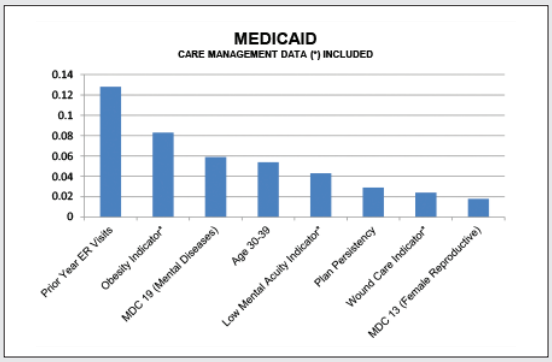
Net savings would be lower, after subtracting costs to develop and put into production a predictive model, clinical staff costs to contact and counsel frequent flyers, and costs of appropriate alternative care settings.

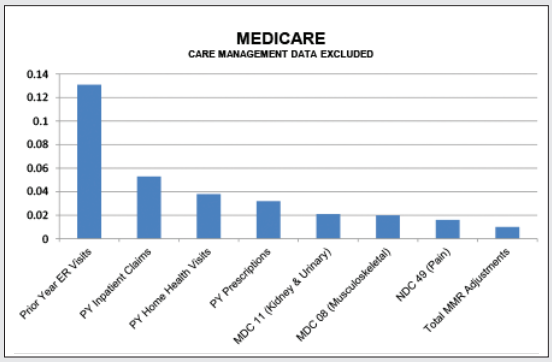
Exhibit 1

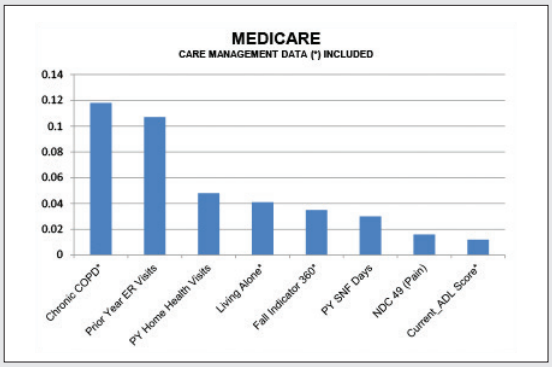
SIGNIFICANT VARIABLES (AVERAGE SEQUENTIAL R^2)









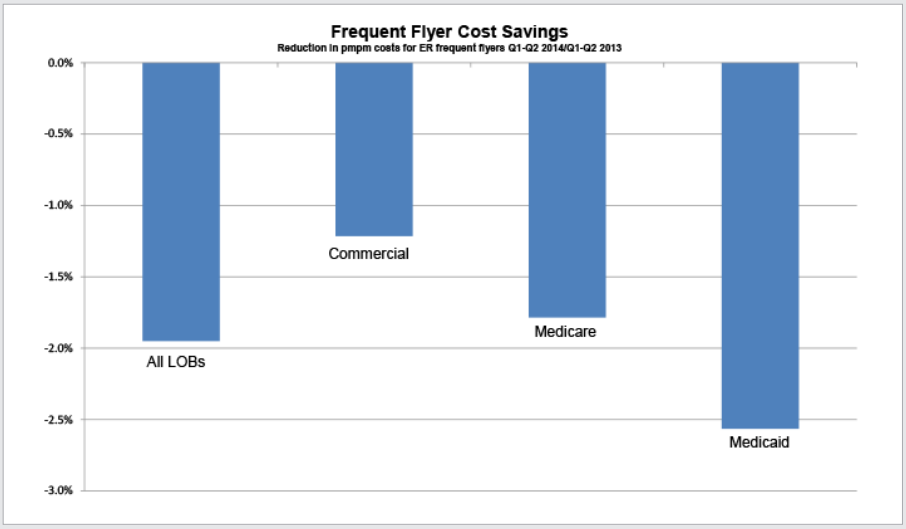


Conclusion

Using information from these models, two care management pilot programs were developed, one for the Commercial and Medicare LOBs and a separate one for Medicaid. Initial results for the first two quarters are promising, with a reduction in pmpm ER frequent flyer costs relative to Q1/Q2 2013 of 1.95%[3]. *Exhibit 2* shows this cost reduction broken out by LOB. The annualized gross savings is estimated to be in the neighborhood of $1 million. This is the first year-over-year reduction in frequent flyer costs observed during the five-year time period of this study.

Exhibit 2

MODEL COST SAVINGS



Overall, the results presented in this paper constitute a first step in developing new intervention programs or enhancing current care management programs. The goal of the intervention is to identify members likely to become ER frequent flyers and communicate with these members to inform them on routine physician visits, urgent care alternatives, and/or provide incentives for them to join a care management program.

Analytics is not a substitute for clinical best practices. It is meant to augment them. It is important when implementing such models for care management purposes to not only track whether or not the model produced a higher ROI than current intervention methods, but to also understand how the combination of model output with existing methods can enhance the valuable work already being performed by nurses and care management staff.

As in any predictive modeling process, there are limitations to our data and/or analysis, as well as identified opportunities for further improvement of the model. Space prevents us from discussing these issues in this article. Readers may contact the first author (Joe Randazzo) for additional information.

3 Frequent flyers are still defined as having three or more ER visits for the first six months in each year.

SOCIETY OF ACTUARIES

Producing Actionable Insights from Predictive Models Built Upon Condensed Electronic Medical Records

Sheamus Kee Parkes
FSA, MAAA
Actuary
Milliman, Inc.

Predictive modeling often has two competing goals: accuracy and inference. In healthcare, risk scoring is used to make different groups more comparable, and to explore drivers of costs. With care coordination specifically, patients need be prioritized for intervention while also understanding why a given patient was prioritized. Care coordination can benefit from custom trained models that adapt to service patterns and include any novel sources of available information. These custom models can include industry-leading risk scores as inputs to retain their strengths and insights. One important novel input could be electronic medical records (EMR) data.

Predictive modeling with EMR is commonly associated with mining physicians' notes for nuanced opinions not found in the coarse diagnosis coding of medical claims. Although valuable, physician notes are not the only information in EMRs; other novel pieces of information include vitals measurements and lab results. Vitals information includes items such as height, weight, and blood pressure. Labs information includes results of panels such as lipid, metabolic, and blood counts. These too can provide a more nuanced view of a patient's health than demographics and claims alone. This article will recount the process of including labs and vitals information into a set of custom models built for care coordination efforts and then understanding the added value in accuracy and insights.

Obtaining and standardizing

The first hurdle in utilizing EMR information is obtaining it; it is often stored separately from claims data and under control of different staff or even a different organization. EMR table structure is commonly even less standardized than claims tables. Limiting to just vitals and labs makes the acquisition process easier. Once acquired, the labs and vitals information need similar, but not identical, processes to make them useful in predictive modeling.

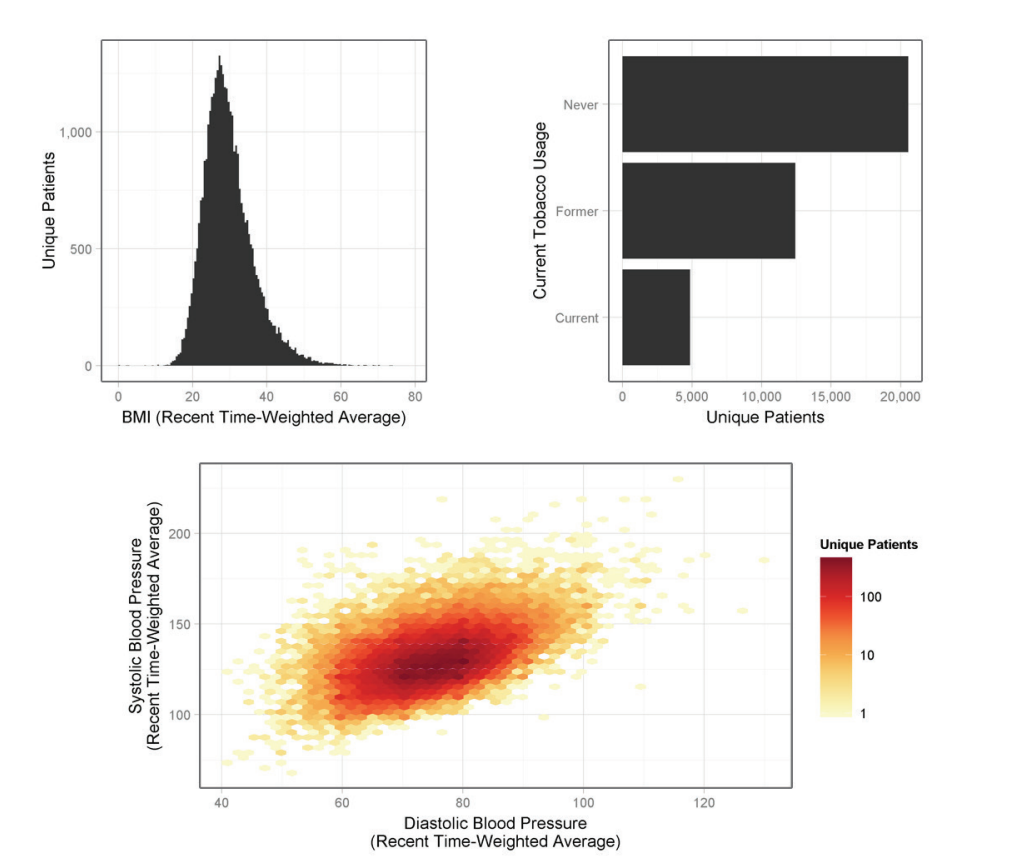
Labs and vitals both are needed on a timeline basis. Just having the most recent results for each patient would not be helpful unless pre-trained models were available that expected them as inputs. When training custom prospective models, a strong history of measurements is needed.

Labs and vitals are both subject to measurement and transcription errors. Although there is some clinical guidance available, concepts from robust statistics are invaluable in estimating useful bounds for outliers. Most items have generally symmetrical distributions of results.

While vitals data is collected more frequently than lab data, there are fewer types of information captured. Figure 1 shows the distribution of some key vitals information.

Figure 1

DISTRIBUTION OF EMR VITALS INFORMATION



Lab tests present additional hurdles. Results are collected from a variety of brick and mortar labs, and typically these entities do not report on a consistent basis. Most grievous is the lack of consistent naming of the item tested. For example, the following terms—BA%, BASOPHILS, Basophils %, and BASO%—all mean the same thing, which is separate from BA#, ABSOLUTE BASOFILS, and BASO (ABSOLUTE). A parsing library must be developed to standardize and categorize the labs data into consistent panel groups and individual items.
After being parsed and categorized, the measured outcome of each lab result needs to be analyzed. Up to 5% of the results are not numeric quantities and might be excluded for convenience. A combination of robust statistics and the central limit theorem can then be used to check the categorization logic for consistent concepts and units of measurement. This feedback can be used to iterate the parsing rules until a reasonable proportion of the lab results are tidied up for use.

Building feature vectors

In healthcare, many analyses use patients as the units of observation. To perform analysis at a patient level, a useful feature vector needs to be built for each patient for each pertinent time period. When training custom models at least two time periods are needed: a historical training feature period for which future outcomes are known, as well as a current prediction feature period for which future outcomes are not yet known (but are of interest).

Within each feature period a given patient may have many measures of a given vital or lab, or none at all. There are many useful ways to collapse these sporadic time series. Simple possibilities would include taking the most recent value or a straight average of all recorded values. A slightly more refined approach would be a weighted average that gave more credit to recent values; this can strike a nice balance between freshness of information and measurement error reduction. There are seldom enough measurements per member to estimate a trend, but differences between first/last and minimum/maximum can be interesting, as can the simple count of the number of measurements of each item. Missing values are coded for those items a patient did not have measured at all.

Choosing among all these encoding possibilities can be somewhat of an art. However, it should be influenced by what learning algorithms will be applied. A reasonable choice of algorithm could be ensembled decision trees, primarily because they gracefully handle missing values, nonlinearities, and interactions while maintaining excellent performance. They can also utilize random feature sampling similar to that championed by Random Forests, so having modestly redundant features can be tolerated, as long as the included EMR features are not so plentiful that the more standard claims and eligibility features become diluted.

Training models and estimating effects

Once the feature vectors are created, reasonable outcomes need to be chosen. Care coordination is often focused on avoiding the worst near-term outcomes, so useful outcomes can include the median and tail risk of total costs for the next six months.

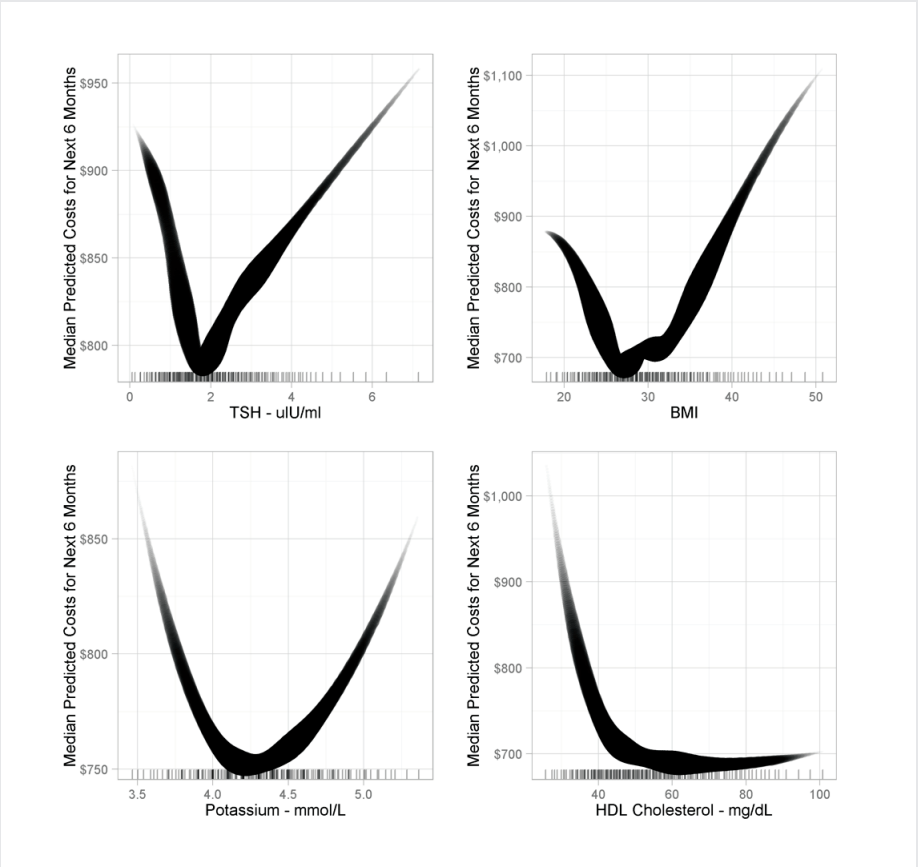
Ensembled decision trees provide useful insights into what features are important. In this example, the claims-based features were still the most important, but the EMR features provided a small lift in model performance when judged on a handful of different metrics. The EMR features did cause large shuffling in the ranking of

predictions, so similar performance was reached with a noticeably different cohort. Possibly more important, the EMR features provided new and potentially more actionable reasons for a given patient's predictions. Marginal effect estimates should likely be avoided when calculating and communicating the effects of individual features in this scenario; marginal effect estimates depend upon holding all other features constant. Given the highly overlapping and collinear nature of many of the features explored here, it is improper to even hypothetically hold all other features constant. Instead, reestimated univariate/single feature effects can communicate more useful information.

The reestimated relation between the median cost predictions and a few EMR features are shown in Figure 2. The rug plots and width of the lines emphasize the area of support that contains most of the example patients' results. The recurring horseshoe shape is very common in EMR effects and reflects a natural optimal equilibrium. These shapes also tend to align with general clinical guidance.

Figure 2

EXAMPLE EFFECTS OF EMR INFORMATION



Presenting results

Care coordination can use these results for both their accuracy and their insights. The predictions themselves can help prioritize what patients are selected for care coordination. The insights can be presented to care coordinators in the form of individual patient profiles. Each patient profile presents many of the features for that patient and ranks them by their importance to the patient's overall prediction. Individual feature importance is derived from the reestimated effects presented earlier in Figure 4 using a given patient's actual feature values. Labs and vitals that appear higher in the feature importance list can be especially valuable for care coordinators because they can represent more actionable information than just warnings of high historical utilization. Care coordinators could still go directly to an EMR for this information, but this feature importance reporting puts the information in a useful context. Adding EMR information provided value, but more to inferential insights than predictive accuracy. However, the value of EMR information depends upon the process used to extract it and this only recounts one useful approach.

SOCIETY OF ACTUARIES

Risk Segmentation: Application of Predictive Modeling in Life Underwriting

Richard Xu, PhD, FSA
Senior Data Scientist and Actuary
RGA Reinsurance Company

Minyu Cao, ASA
Assistant Actuary
RGA Reinsurance Company

Scott Rushing, FSA
VP & Actuary
RGA Reinsurance Company

As the data assets available to companies continue to expand, predictive modeling has become a powerful tool for assessing risks. The impact can sometimes be profound, and the utilization of these methods may forever change how insurance decisions are made. A useful application of predictive modeling in insurance is risk segmentation in underwriting.

Historically, the life insurance industry has developed underwriting tools to differentiate and select risks. These tools can be very effective in risk assessment, but the process can be time consuming and can create large acquisition costs to the insurer. Given adequate data, statistical modeling can also differentiate risks based on the risk profile of the applicant, but can result in a much more efficient and inexpensive process.

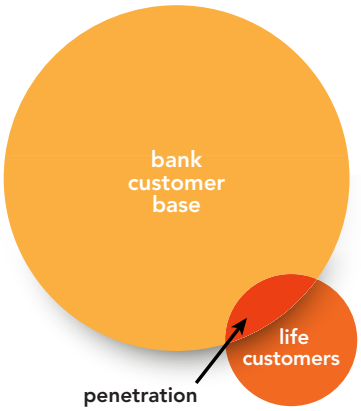
Background

A bank in Asia with a large customer base expressed a strong desire to increase the sales penetration of their life product, while streamlining the underwriting process. Their research attributed the relatively small proportion of life sales to their millions of banking customers to a burdensome underwriting process.

The bank was interested in understanding how their existing customer data could be used to simplify the current underwriting process. By knowing in advance which customers were most likely to qualify, the bank hoped to simplify the underwriting process for the lowest-risk customers. As a result, both sales volume and market penetration should increase.

Predictive modeling techniques were utilized to identify the most desirable risks using demographic and financial information on existing banking customers. The insurance company then tailored the resulting underwriting process to best align with the risks not properly identified by the model.

Figure 1



Data

Supporting data for this project comes from two sources: the underwriting data at the time of issue for the life policies and the bank's financial database. The two datasets were linked for the overlapping banking and insurance customers. To comply with regulatory requirements, all personally identifiable information was removed.

The data fields provided for modeling purposes were fairly comprehensive. The customer profile contained more than 80 predictor variables including demographic data, financial data, banking transaction summary data, product information and some behavioral data. Additionally, the underwriting decision (standard / sub-standard / decline) was provided for each case.

A key data concern was the limited information provided on the declined cases. Likewise, many policies had data fields that were unavailable, as we were attempting to collect the data values at the time of underwriting. Overall, the final dataset was relatively small with less than 10,000 underwritten cases available for modeling.

Modeling approach

For this project, a Generalized Linear Model (GLM) framework was chosen. GLM provides transparent model results and intuitive business insights. Although it is not the most advanced modeling technique available, the ease of interpreting results from a GLM improves the understanding of a model and leads to more effective decision making.

The primary goal of this model was to differentiate between standard and non-standard risks. There were no preferred risks available for this product. Non-standard risks are defined to include both rated policies and declined applications. The ultimate target variable is usually the mortality rate, but more data was required to build a credible model. For this project, the underwriting decision was chosen for the target variable.

Statistical procedures were used to identify those variables with the most predictive power. Balance between complexity and accuracy of the model were carefully monitored to ensure the final model was effective without over-fitting it to the data.

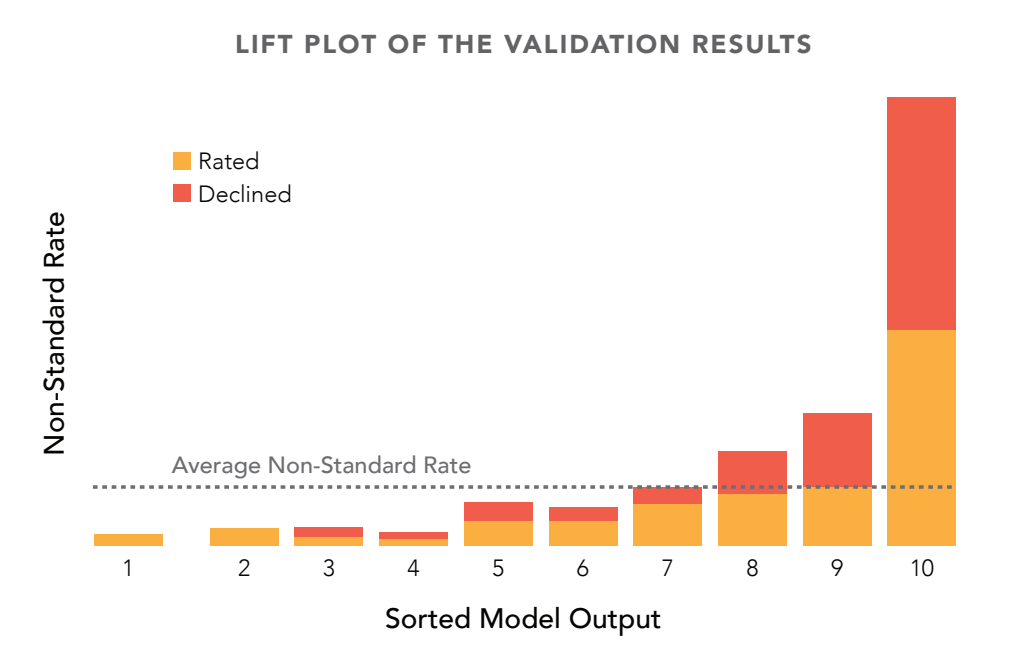
The final model contained 11 predictor variables – all statistically significant for prediction purposes and readily available in the customer database. The final variables include demographic information such as issue age, gender, marital status and socio-economic data such as total bank assets under management.

Results

To understand the predictive power, results from the model validation process were investigated. The model output was sorted into ten equally sized groups (deciles), with those customers "most likely to be standard" in the first decile and those customers "least likely to be standard" in the tenth decile. Using these results, a lift curve was generated to visually display the distribution of non-standard risks across the deciles.

For a perfect model, all non-standard policies would end up in the highest deciles (to the right). For a useless model, the distribution of non-standard policies would be nearly uniform (average rate) across the deciles. In reality, we expect to see a lift curve that falls somewhere between these two extremes. The lift plot displayed shows approximate model results.

Table 1



From the above graph, we can see that:

- The worst five deciles capture about 90% of all non-standard risks.
- The proportion of policies that would be declined is significantly lower in the best deciles (deciles 1 through 4).
- The average non-standard rate for the first 4 deciles is about 20% that of the overall average non-standard rate.

By leveraging existing data assets and building a predictive model to replicate the underwriting decision, the number of policies requiring traditional underwriting can be significantly reduced. The utilization of predictive

modeling for underwriting purposes can identify the best risks, simplify the underwriting process, speed up the underwriting decision, increase sales, and greatly reduce acquisition costs.

Implementation considerations

To simplify the implementation process, the results of the GLM formula are converted into a simplified underwriting index. The Index is calculated for each banking customer based on the characteristics for that individual. The index is further calibrated so that nearly all customers will get a value between 0 and 100.

Based on the score, the bank can offer reduced underwriting for those banking customers with a sufficiently high model score. Other restrictions, such as age and face amount limitations, are put in place to manage risk and reduce anti-selective behavior.

Once the model is deployed, procedures should be set up to monitor performance. A good quality assurance and feedback loop will help ensure the model is working properly and accomplishes the original objectives.

SOCIETY OF ACTUARIES

Predictive Modeling Techniques Applied to Quantifying Mortality Risk

Vincent J. Granieri,
FSA, MAAA, EA
Chief Executive Officer
Predictive Resources LLC

Actuaries are familiar with the interaction of art and science in their work. Some view underwriting in the same way, perhaps concluding that underwriting leans more toward art than science. With the advent of powerful computers and predictive modeling tools, it is possible to analyze survival data and produce statistically credible underwriting models that predict relative mortality risk among individuals based on demographic information and relevant conditions. In this paper, we will discuss the use of the Cox Proportional Hazards Model in developing a predictive underwriting model that produces a mortality multiplier for each individual.

Further, we wished to quantify the impact on survival, if any, of certain subpopulations. We were looking to validate the time–accepted concepts of the wealth effect (beyond the scope of this paper) and anti-selection in our population.

Cox Proportional Hazards Model

The Cox Proportional Hazards Model was introduced in 1972 as a method to examine the relationship between survival (mortality) and one or more independent variables, called explanatory variables. Some advantages of the Cox model are that it can handle many underwritings on the same life and can utilize data that is right censored; i.e. subjects can leave the study at any time or the study can end before all subjects have died. The Cox model does not require knowledge of the underlying (base) survival curve, but we will see that this advantage is also a challenge when analyzing mortality.

Cox Model results are expressed as the logarithm of the hazard so technically, the relative risk factor for each variable is obtained by raising e to the power of the log(hazard) ; e.g. consistent with Gompertz. The relative

risk factor is interpreted just as it sounds: it describes the force of mortality relative to the reference. A relative risk factor of two for a condition means the subject is twice as likely to die as another subject who does not have that condition.

As an aside, we utilized the R statistical package to produce our survival models. It is particularly well-suited for this type of analysis. Other popular statistical packages, such as SAS, also contain survival models using the Cox algorithms.

The issues

The most important issue was that of the underlying mortality distribution. We already had produced mortality tables that varied by age/gender/tobacco use. What then should we do with the results that also calculated the impact of these variables? We decided to use our existing base tables after reviewing the model results for consistency with them.

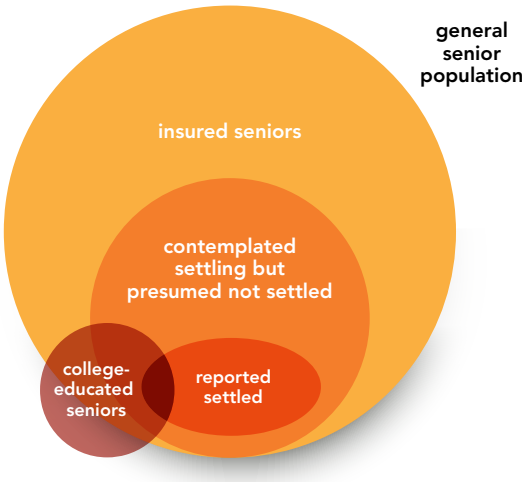
It was also very important to ensure that the explanatory variables were truly independent. If not, spurious results would ensue. We also had to redefine certain variables, such as BMI, where the risk was actually related to straying from the ideal BMI measurement, rather than the measurement itself. There were many other issues, too numerous to mention in a paper of this length.

Input data

For this exercise, we had available to us over 200,000 underwriting events on 80,000+ unique senior lives, which took place over a 15 year period, primarily in the life settlement market. Figure 1 is a graphic description of the major subpopulations of the universe of senior lives and the populations we studied. At the highest level, there is the general senior population. Some of these seniors have purchased insurance, creating a subpopulation, which can be further broken into two subpopulations; those who actually sold their policies on the secondary market and those who contemplated such a sale, but for some reason, did not conclude the sale. These latter two subpopulations were the basis for our study of antiselection. There is also a small population of college-educated seniors, some of whom can also be associated with the other populations above, which formed the basis for our study of the wealth effect. This data included demographic information such as age, gender, dates of birth and dates of death. It also included various underwriting conditions such as BMI, smoking status and indicators for various diseases. Included were favorable conditions, such as family history of longevity and good exercise tolerance.

Figure 1

SENIOR POPULATIONS



Creating Cox Proportional Hazards Models

There was significant data preparation involved. We set up the reference population, which we chose to be males who were age-appropriately active, who did not sell their policies and did not use tobacco. Variables were determined to be either continuous (age, BMI), where the condition has infinite possible values, or binary (CAD, osteoporosis), where the condition either exists or does not. This required considerable judgment and depended on the availability and form of the data.

Once the data were prepared, we began the process of determining which conditions were statistically significant in predicting mortality. We underwent an iterative process. The Cox models were run with every variable included at first. We then we reran the models, first eliminating most of those variables with a p-value

Figure 2

	All (<=0.05)				
	Log (hazard)	Hazard	Lower CI	Upper CI	P-Value
Age	0.077	1.080	1.075	1.085	-
Actual BMI less ideal BMI	0.002	1.002	1.001	1.002	0.000
Recurrent Cancer	0.458	1.581	1.365	1.832	0.000
Female	(0.365)	0.694	0.649	0.742	-
Active for their age	(0.141)	0.869	0.802	0.942	0.001
Sedentary	0.200	1.221	1.054	1.415	0.008
Unknown activity level	0.102	1.107	1.031	1.189	0.005
Family history of longevity	(0.087)	0.917	0.857	0.981	0.012
Family history of super longevity	(0.240)	0.787	0.722	0.857	0.000
College-educated population member	0.267	1.306	1.117	1.526	0.001
Settled population member	(0.370)	0.691	0.650	0.734	-
Current smoker	0.635	1.887	1.693	2.103	-
Discontinued smoking	0.178	1.195	1.128	1.267	0.000
Rare smoker	(0.339)	0.713	0.266	1.911	0.501
Tobacco replacement	0.576	1.780	1.187	2.668	0.005
Unknown tobacco use	0.119	1.127	1.018	1.247	0.021

Reference: Male, nonsmoker, normal activity level

greater than 0.2. This means we were excluding those conditions where the probability that the relative risk shown was due to random fluctuation was over 20%. These models were again rerun, this time eliminating those conditions with a p-value greater than 0.1. Finally, we reran the models, including only those conditions where the p-value was at most 0.05.

Figure 2 represents partial output from our models, consisting of conditions that were included in all runs even if they did not meet the criteria for continued inclusion above. As we advanced through the process, we felt strongly that these were fundamental variables that clearly impacted survival and should be included in the analysis regardless of their p-values. In reality, only one variable would have been eliminated, presumably due to data scarcity. Orange/yellow shading indicates that a condition is hazardous/protective, with the 95% confidence limits and p-values also shown. For example, the female hazard is 0.694 of that of males (1.0 as males are the reference) and the smoker hazard is 1.887 times that of nonsmokers. For the other explanatory variables, many were eliminated as the p-value criteria became more stringent.

Conclusions

The most important conclusion that we drew from this exercise was that despite our best efforts to quantify every aspect of underwriting, there is still considerable judgment brought to bear in that process. However, there is also much useful information that predictive models can provide us because of their ability to process large amounts of data quickly and efficiently. We did validate the anti-selection that occurs between those who actually sell their policy versus those who do not. Some results confirmed our clinical judgment; for example, an active lifestyle or family history of longevity are indicators of higher survival rates. Other things went against our clinical judgment; for example, cardiac related conditions, while still hazardous, were no longer as significant as we thought.

Then there were the confounding results. Hyperlipidemia was shown to be protective. We attributed this to the ubiquity of statins. There were a number of other conditions that were shown to be mildly protective, things such as BPH, sleep apnea, use of blood thinners and benign colon polyps. We concluded that these were indicators of frequent/better quality of healthcare, which would allow for early detection and mitigation of more serious risks.

Business outcomes

This analysis was the basis for changes in our debit/credit underwriting model. We replaced an additive model based only on clinical judgment with one that was more consistent with mortality research and provided us the flexibility to continue to factor in clinical judgment where appropriate.

SOCIETY OF ACTUARIES

Utilizing Frequent Itemsets to Model Highly Sparse Underwriting Data

Jeff Heaton
Data Scientist
RGA Reinsurance Company

Dave Snell, ASA, MAAA
Technology Evangelist
RGA Reinsurance Company

Predictive modeling holds great potential for insurance underwriting. Human underwriters spend hours distilling many pages of medical information into the few pertinent facts needed to classify a risk. Models such as linear regression, generalized linear models (GLM), random forests and others require input data to be a fixed-length numeric vector. This makes it difficult to model underwriting data, such as an attending physician statement (APS), because the APS is both highly dimensional (lots of potential body, lab, and other metrics) and sparse (various lab results, prescriptions, etc. may be missing).

This paper discusses our use of frequent itemsets[1] to preprocess drug information from APS data. Frequent itemsets are an example of a recommender algorithm used in data mining that deals effectively with highly sparse data. Companies, such as Amazon and Netflix[2] utilize them to suggest products for customers based on selections by other customers. Consider customers who frequently buy movies A, B and C. If another customer buys movies A and C, suggesting movie B is beneficial.

We are exploring two different frequent itemset applications. In both applications drugs are the items the algorithm recommends. Using frequent itemset mining we locate drug frequent itemsets. A partial set might alert the underwriter to seek the missing drug. Because data often comes from several different providers, such a warning is helpful. To reduce dimensions, each frequent itemset becomes a drug profile. We model with a single feature indicating the insured's nearest profile.

1 Mannila, H. T. (1994). Efficient algorithms for discovering association rules. (pp. 181-192). AAAI Press.

2 Bennet, James; Lanning, Stan; and Netflix, Netflix (2007). The Netflix Prize

Building a frequent itemset

We built a frequent itemset for APS documents with each transaction being an applicant's drug history. This is analogous to a Netflix customer's rentals, or an Amazon shopper's purchases.

allopurinol,atenolol,colchicine,dimenhydrinate,omeprazole
cephalexin,clonazepam,sertraline hydrochloride
escitalopram oxalate,lisinopril and hydrochlorothiazide,simvastatin
bupropion hydrochloride,hydrochlorothiazide,nicotine,sertraline hydrochloride
sertraline hydrochloride,trazodone hydrochloride
…

Each data line above represents a basket. For our application each basket represents drugs taken by a single applicant. Normally, the above data would specify drug codes, such as GPI, NDC or RxNorm; however, for clarity, we listed drug names. The baskets are expanded to a large, sparse matrix similar to Table 1.

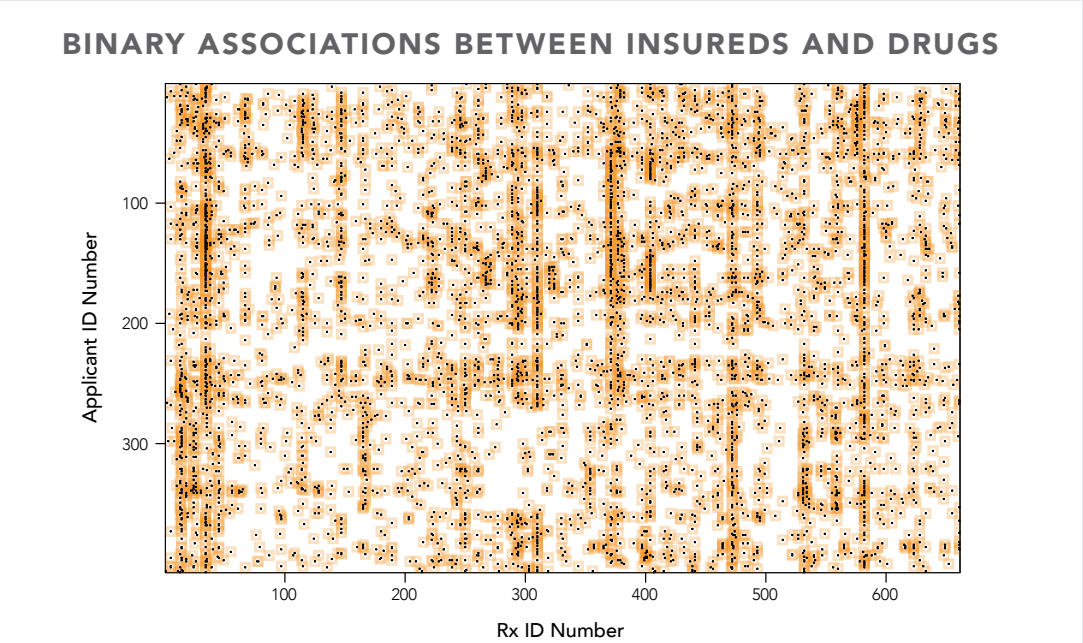
Table 1

BINARY ATTRIBUTE MATRIX

	Insulin	Simvastatin	Glipizide	Ibuprofen	…
Applicant #1	1	1	0	0	…
Applicant #2	0	0	1	0	…
Applicant #3	0	0	0	1	…
Applicant #4	1	1	1	0	…
Applicant #5	0	1	0	0	…
Applicant #n	…	…	…	…	…

Table 1 shows a simplified version of our actual binary association table. In reality, one column exists for every possible drug that the applicant might have taken. This results in a large number of columns. Most modeling packages internally store this data more efficiently as a sparse array. Figure 1 shows a visualization of an actual binary matrix of APS drug information.

Figure 1



The x-axis represents the individual drug types. The y-axis represents the insureds analyzed in this study. Figure 1 allows you to see relationships in the data. Vertical clusters indicate drugs common across a large number of insureds. Absence of any large white regions is an indicator the drugs span sufficient transactions to build relationships.

Modeling packages analyze the transactions to find frequent itemsets; specifically, combinations of drugs that frequently occur together. These relationships are expressed as rules.

$$\{Insulin, Simvastatin\} \Rightarrow \{Glipizide\}$$

If insulin and simvastatin are present, then glipizide might also be present. The data is analyzed to find these rules. Applicant #4 in Table 1 supports this rule. However, there might not be enough support in the entire data set for this rule to make it into the final ruleset. Constraints determine which rules are chosen for the final set. The following three calculations help determine if a rule should be included.

Support: $supp\ (X)$ is the percentage of transactions that contain the feature set X.

Confidence: $conf\ (X \Rightarrow Y) = \dfrac{supp\ (X \text{ and } Y)}{supp\ (X)}$ [confidence ranges from 0 to 1]

Lift: $lift\ (X \Rightarrow Y) = \dfrac{supp\ (X \text{ and } Y)}{supp\ (X) \times supp\ (Y)}$ [lift ranges from zero to a large but finite number]

Taken together, the set {insulin,simvastatin,glipizide} has a support of 1/5=0.2, being present with one of five applicants. The confidence for {*insulin, simvastatin*} ⇒ {*glipizide*} is 0.2/0.2=1.0, so for 100% of the cases containing insulin and simvastatin, glipizide is also present. Actuaries can think of confidence as the Bayesian conditional probability P(Y |X) times P(X). Lift for the above rule is calculated as:

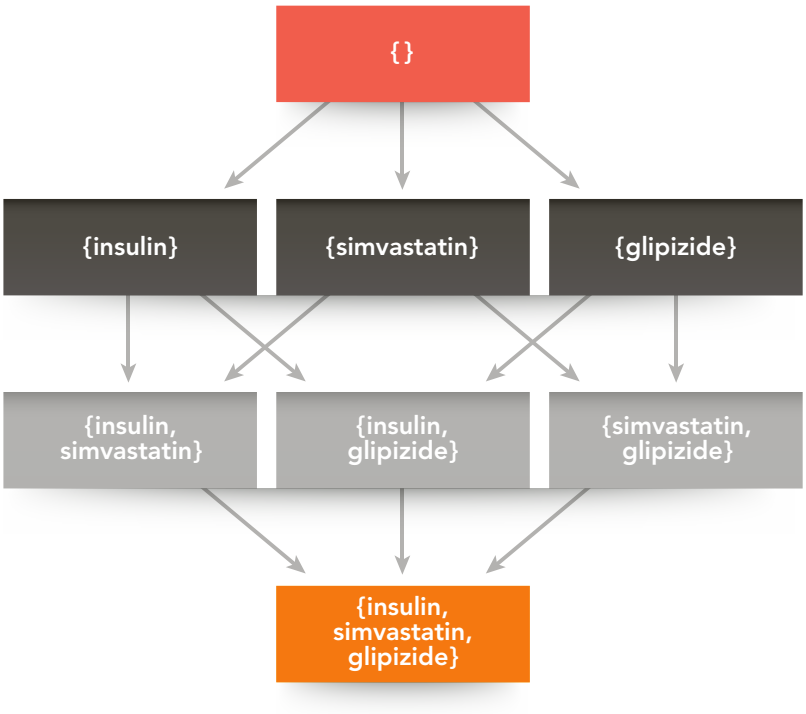
$$\frac{0.2}{0.4 \times 0.4} = 1.25$$

Lift is the ratio of support for the rule to the support for {insulin,simvastatin} and {glipizide} evaluated independently. Higher lift indicates that the probability of X and Y occurring together is greater than the probabilities of both X and Y occurring separately. For example, many families own cats or dogs and some families own both. A higher lift would indicate the prevalence of families that own both species of pet over their single pet species counterparts.

These calculations are used by a variety of algorithms to build the feature sets. One common algorithm is the apriori algorithm[3]. This algorithm builds a frequent itemset lattice, as shown in Figure 2.

Figure 2

FREQUENT ITEMSET LATTICE



3 Srikant, R. A. (1994). Fast algorithms for mining association rules in large databases. (pp. 487-499). Santiago, Chile: Proceedings of the 20th International Conference on Very Large Data Bases,.

In Figure 2's feature lattice, the algorithm starts with the empty set {}. Working downward, the first row adds all single item frequency sets with sufficient support. The next level adds supported two-element sets. The algorithm continues downward, always building new sets by combining supported feature sets from the previous level. Most algorithms limit the number of levels visited.

Building and using frequent itemset for drug codes

We analyzed digitized APS reports from previous cases using R with the ARULES package[4]. Table 2 shows frequent itemset rules with larger lift values.

Table 2

FREQUENT ITEMSETS AND RULES MINED FROM APS DATA

Rule	Support	Confidence	Lift
{sibutramine hydrochloride} → {orlistat}	0.01719902	1.0000000	40.70000
{orlistat} → {sibutramine hydrochloride}	0.01719902	0.7000000	40.70000
{zolmitriptan} → {sumatriptan}	0.01228501	0.7142857	26.42857
{albuterol, guaifenesin} → {ipratropium bromide}	0.01228501	1.0000000	22.61111
{formoterol fumarate} → {ipratropium bromide}	0.01228501	0.8333333	18.84259
…	…	…	…

This small set of rules provides interesting insights. Sibutramine Hydrochloride and Orlistat are both weight control drugs and often occur together. Zolmitriptan and Sumatriptan both treat migraine headaches and often occur together. Subject matter expertise can help identify patterns.

Conclusions and next steps

Medical information presents unique challenges for modeling applications. Medical data is very sparse, with many dimensions. We are beginning to investigate the use of frequent itemsets to assist models that use non-sparse, tabular data about insureds, such as build and disclosure information.

We plan to use frequent itemsets to investigate applicant medication use. We will look at how to balance rare, yet highly significant, drugs that might lack sufficient support for inclusion in an item set. We are researching how to ensemble models using feature sets with models better suited to detect rare, yet significant, drugs.

4 Hornik, M. H. (2005). arules -- {A} Computational Environment for Mining Association Rules and Frequent Item Sets. (pp. 1-25). Journal of Statistical Software.