Using Asymmetric Cost Matrices to Optimize Wellness Intervention

Zoe Gibbs[a], Brian Hartman[a,*]

[a] *Department of Statistics, Brigham Young University, Provo, UT, USA*

Abstract

The majority of healthcare expenditures are incurred by a small portion of the population. Care management or intervention programs may help reduce medical costs, especially those of extremely high-cost members. For these programs to be effective, however, the insurer must identify and select potential high-cost members to be assigned to an intervention before they incur those costs. Because high medical costs are often connected to an accident or traumatic event that cannot be anticipated, it can be difficult to predict who will be high-cost in the future. In this paper, we explore the use of machine learning in predicting high-cost members. Specifically, we use the extreme gradient boosting algorithm to develop risk scores for members based on demographic, medical, and financial histories. To select members for intervention, we develop asymmetric cost matrices that account for potentially unequal savings or losses for assigning interventions to members. We show how these matrices can be reduced to a function of the expected savings per dollar of intervention, which is easily used to optimize the risk score threshold at which members are assigned an intervention. These techniques, which can be tailored to the specific needs of an insurer, may help insurers select the optimal members for intervention programs, reduce overall costs, and improve member health outcomes.

1. Introduction

Because a small portion of the population accounts for the vast majority of health care expenditures, insurers and policymakers are interested in developing methodology to select potential high-cost members for cost-saving intervention programs. Ideally, insurers and medical personnel will employ methodology to aid in the selection of members for interventions that will not only reduce costs, but also improve quality of life. Many insurers are already employing such practices, but the literature on member selection for intervention is sparse. This paper seeks to add to the available methodologies of member selction and offer insight into potential areas for future research.

With the increasing popularity of wellness and care-management programs, it is important to understand the methods used to select patients for interventions, as well as the costs and benefits associated with the interventions. Depending on availability, many types of data can assist in patient classification in addition to traditional claims and enrollment data. Electronic medical records may have great potential in optimizing patient selection for intervention by identifying characteristics of high-cost patients and patterns of triage or other costly medical events (Bates et al., 2014). Self-reported health assessments can also provide valuable information in selecting members for care-management programs (Kim and Rosenberg, 2018).

Several algorithms have been proposed for scoring members in terms of risk or cost. A Dirichlet process mixture of log-normals has been shown to accurately predict future costs while appropriately accounting for uncertainty in future claims (Hong and Martin, 2017). To account for the variation in frequency of health care events (i.e. hopsital stays), Frees et al. (2011) outlined a two-part model that first predicts the number of events and second, models expenditure per event. In models that classify members as either low or high risk, the number of chronic conditions has been shown to be a significant predictor (Fleishman and

[*] Corresponding author

Email addresses: zgibbs8@gmail.com (Zoe Gibbs), hartman@stat.byu.edu (Brian Hartman)

Cohen, 2010). Other authors have found that it can be helpful to focus an analysis on a specific source of high costs. One study focused on predicting costs for some of the most expensive members–those in need of long-term care–who might benefit from further interventions (Lally and Hartman, 2016). Prediction of hospital readmissions for those with and without chronic illness could also provide information in patient selection for care-management. (Billings et al., 2006; Rosenberg and Farrell, 2008).

In recent years, many researchers have cited boosting as a promising tool in fine-tuning insurance claims prediction models. Duncan et al. (2016) concluded that alternative models to regression, including boosted trees, resulted in more accurate predictions. Lee and Lin (2018) furthered this research by finding that Delta Boosting outperforms Gradient Boosting in prediction. These methods build upon earlier publications that cite cost-sensitive boosting algorithms as ideal for classification of imbalanced data sets, including medical claims data (Sun et al., 2007).

Each of these models is only as effective as the thresholds used to inform resource allocation. Thus, cost-effectiveness should drive care-management decisions to maintain transparency (Eichler et al., 2004). To calibrate decision thresholds, translating various thresholds chosen with different performance metrics into expected losses may be helpful (Hernández-Orallo et al., 2012). Although traditional cost functions often model symmetric losses, these models are not suitable for many economic and management situations where an asymmetric loss function more accurately describes reality (Granger, 1969). Moreover, asymmetric surrogate loss functions display potential in binary classification for situations with unbalanced training data sets or unequal misclassification costs (Scott et al., 2012), such as medical claims data.

The development of asymmetric loss functions to optimize threshold selection for member intervention centers around the cost-effectiveness and quality of the intervention programs. Duncan et al. (2011) analyzed savings for members with chronic conditions enrolled in a health management program, finding that savings were highest for those with more costly diseases and for whom the duration since selection for intervention was longer. Another study showed that patients with mental illness may benefit from participating in a wellness program that helps them combat the negative and costly side-effects of psychopharmacologic treatment, including weight gain (Hoffmann et al., 2005). Billings and Mijanovich (2007) researched the net savings of mandatory Medicaid managed care programs in New York City, finding that although some members with lower risk scores enrolled in the program experienced savings that were merely equal to the cost of the intervention, total savings across the entire program were significant. As mobile and virtual technologies continue to improve, the effectiveness of telehealth management programs have been analyzed. One program that selected members with high predicted costs to take part in a telephone-based care management program saw a 3.6% reduction in costs and a 10.1% reduction in hospitalizations compared to the control group (Wennberg et al., 2010). A systematic review of mobile health technology intervention programs, found evidence that text-messaging interventions could improve medication adherence and smoking cessation (Free et al., 2013). This is significant because for some chronic conditions, increased medication adherence is associated with decreased disease related medical costs such that the savings offset the increased cost of medicine (Sokol et al., 2005).

2. Methods

Extreme gradient boosting (XGBoost) with a binary logistic objective function was used to create risk scores for each member in the data set. The algorithm creates a series of decision trees, such that each new tree improves upon the error of the previous tree. Within each tree are several branches, or splits in the tree. At each split, a random sample of the covariates is taken. From the random sample, the covariate that provides the greatest reduction in error is chosen to create a split in the tree. The splitting continues until a maximum tree depth is reached, after which the tree is pruned back to remove all splits beyond which there is no gain in accuracy. With each iteration, a different weighted sample of the training set is chosen to focus on the errors of the previous trees. A weighted average of all the trees is computed to form the final model. The model output is a score for each member, indicating the probability that the member will be high-cost next year.

To implement the XGBoost algorithm, we optimize several hyperparameters. The following hyperparameters were optimized over the ranges indicated:

- Column Sample by Tree (.5,1) : The proportion of columns (covariates) that are randomly chosen at each split for consideration for use.
- Eta (0,1) : The learning rate. A lower value implies a slower learning rate that may improve prediction precision but also slows convergence.
- Gamma (1,100) : The minimum reduction in loss acceptable to justify the creation of a new node. Lower values imply greater tree depth.
- Max Depth (3, 10) : Maximum number of splits in each tree. The risk of overfitting increases with increasing depth.
- Minimum Child Weight (1, 10) : Indicates the threshold at which if splitting a node results in a child weight (number of instances in a node) less than the parameter provided, the current iteration of tree-building stops. A higher value implies simpler trees.
- Subsample (.5,1) : The proportion of the training set that is selected at each boosting iteration. A smaller subsample helps prevent overfitting.

Training and test sets were formed such that the training set contained approximately two-thirds of the data and the test set contained the other one-third of the data. The training and test sets were stratified so that the proportion of high-cost members in each of the training and test sets were approximately equal to the proportion of high-cost members in the original data set, an especially important feature for unbalanced datasets. The parameters were tuned using five-fold cross-validation over the training data set with random searches over the parameter spaces indicated above. The measure for optimization was accuracy with a default threshold of 0.5. The tuned values of the hyperparameters can be seen in Table 1.

Hyperparameter	Tuned Value
Column Sample by Tree	0.967
Eta	0.016
Gamma	4.485
Max Depth	7.000
Minimum Child Weight	5.978
Subsample	0.993

Table 1: Tuned Parameters

Because each tree is built using a different weighted sample of the training set designed to reduce the error of previous trees, XGBoost is a promising algorithm for unbalanced data sets, including medical claims data. Additionally, the random selection of covariates for examination at each split in the tree helps reduce collinearity among variables.

To assist in our goal of selecting members for intervention, the output of the XGBoost may be interpreted as a risk score, where a higher score represents higher risk. Once the risk scores are calculated, a threshold for intervention must be chosen such that those with risk scores above the threshold are assigned a care management program, and those below the threshold are not. This threshold should be calculated to maximize total savings. We assume that members selected for intervention who are truly high-cost will experience savings. Conversely, if a low-cost member is selected for the same intervention the insurance company must pay for the entirety of the intervention without experiencing savings. By this logic, if the cost of intervention is near zero but the savings are positive, we would want to select all members for intervention and would apply a selection threshold close to zero. Because selecting low-cost members costs money without yielding savings, as the cost of intervention increases, we want to select fewer members in order to reduce the number of selected low-cost members. Thus, the optimal threshold should approach one as the cost of intervention increases (holding the savings constant).

A loss function that considers the expected costs and savings associated with care-management programs must be developed such that interventions with high savings per dollar of intervention have a low threshold, and interventions with low savings per dollar of interventions have a high threshold. Because cost and savings of interventions may vary by demographics, disease, etc., it is important that the loss function for optimization is both intuitive and computationally simple. To build the intuition behind the proposed loss function, we have organized our predictions in the form of the confusion matrix shown below:

$$\begin{array}{r|cc|} & \text{Negative Prediction} & \text{Positive Prediction} \\ \text{Actual Negative} & \text{True Negative} & \text{False Positive} \\ \text{Actual Positive} & \text{False Negative} & \text{True Positive} \end{array}$$

It is important to consider that losses are unlikely to follow a symmetric pattern, meaning that the cost of a false prediction might not equal the savings associated with a true prediction. Thus, to accurately model the financial impact of prediction, it is necessary to develop asymmetric cost matrices. Below, we describe the simplest form of an asymmetric matrix that gives way to a loss function for threshold optimization. These matrices can be adjusted to include demographic or intervention specific rewards or penalties.

Because those who are predicted to be low-cost are not assigned an intervention program, zeroes can be inputted on the first column of the matrix since they have neither intervention costs nor savings. Those who are predicted to be high-cost but are, in fact, low-cost are assigned an intervention program, but may not experience the expected savings from the program. Thus, false positives can be assigned a penalty of the cost of intervention (I). Those who are predicted to be high-cost and are actually high-cost experience a net reward of the amount of savings (S) less I.

$$\begin{array}{r|cc|} & \text{Negative Prediction} & \text{Positive Prediction} \\ \text{Actual Negative} & 0 & -I \\ \text{Actual Positive} & 0 & S-I \end{array}$$

This matrix can be further simplified by dividing all elements by I to represent a function of S/I or the savings per dollar of intervention:

$$\begin{array}{r|cc|} & \text{Negative Prediction} & \text{Positive Prediction} \\ \text{Actual Negative} & 0 & -1 \\ \text{Actual Positive} & 0 & \frac{S}{I}-1 \end{array}$$

This matrix gives rise to a function (1) that is maximized at the optimal threshold value.

$$\frac{S}{I}(\#\text{ of True Positives}) - (\#\text{ of False Positives} + \#\text{ of True Positives}) \tag{1}$$

3. Data

The data used in this analysis contains member information from a large insurance company in 2012 and 2013. The data was cleaned to include only those who had 12 member months of coverage in both 2012 and 2013, resulting in a total of 967,031 members. The explanatory variables are listed in Table 2.

ETGs refer to episode treatment groups. ETGs, which were devleloped and patented by OPTUM, group related medical and pharmacy claims into medically relevant units that describe a complete episode of care. Our data contain 320 ETG indicators. A value of one means that a member had a claim related to a particular ETG while a value of zero means that a member did not have a claim related to that ETG.

Table 3 contains summary statistics on a few covariates. The sample population contains a nearly equal proportion of male and female members, but appears to be somewhat young with a mean age of 34 years. The mean diagnosis count is much greater than the mean ETG count, which is to be expected since several

Variable Name	Description
Funding Arrangement Type	To maintain confidentiality for the data provider, the following codes will be used: 0, 1, 2, 8
Gender	Member Gender (M for male and F for female)
Member Age	Age in years
Total Cost 2012	Total allowed dollars in 2012
Out of Network	1 indicates out of network claims, 0 otherwise
Diagnosis Count	Count of distinct ICD-9 codes in member's claim history
Specialty Drug Count	Number of prescribed specialty pharmaceuticals
Standard Drug Count	Number of prescribed standard pharmaceuticals
Injectable Drug Indicator	1 if the member has been prescribed injectable drugs, 0 otherwise
Psychotropic Drug Indicator	1 if the member has been prescribed psychotropic drugs, 0 otherwise
Maintenance Drug Indicator	1 if the member has been prescribed maintenance drugs, 0 otherwise
Prospective Risk Score	Prospective risk score
Prospective Risk Category	Binning of prospective risk numbers
ETG Count	Number of indicated episode treatment groups
Group 2012	Binning of total allowed dollars in 2012 (<100K, 100K-250K, 250K-500K, >500K)
320 ETG Indicators	1 if the member has claims associated with the specified episode treatment group, 0 otherwise

Table 2: Variable names and descriptions

Members	967,031
Mean Age	34.63
Percent Female	50.33
Percent Male	49.67
Mean Diagnosis Count	4.88
Mean ETG Count	1.81

Table 3: Summary Statistics

diagnoses may fit into one ETG and there may be diagnoses that fit into an ETG not indicated in the dataset.

Table 4 displays comparative statistics for allowed costs in 2012 and 2013. Allowed costs were significantly higher in 2012 than 2013. Additionally, the number of members with allowed claims totaling over $100,000 was much greater in 2012. The stark difference in claims between the two years may pose a challenge to prediction if a suitably robust method is not used.

Our goal is to identify members who are most likely to incur large claim costs and to recommend them for a health intervention program that could improve quality of care while yielding medical cost savings. Therefore, we are focusing on the binary classification of members as high or low-cost. For simplicity, we define high-cost members as those who incur at least $100,000 in allowed claims in a single year. In 2012 the number of high claimants was 4,351 while in 2013 the number of high claimants was 1,783. The rareness of extremely high-cost members necessitates extra consideration in the modeling process.

4. Results

Although inference is not the goal of this paper, it can provide background and context to our predictions. Thus, we will briefly explore the effect of each covariate on prediction. To infer the relative importance of the covariates in predicting high-cost members, a feature importance plot may be used. The plot is a visual

	2012	2013
Mean	4,912	3,049
Median	1,102	662
Maximum	2,631,172	1,800,314
>100,000	4,351	1,783

Table 4: Comparison of Allowed Dollars in 2012 and 2013

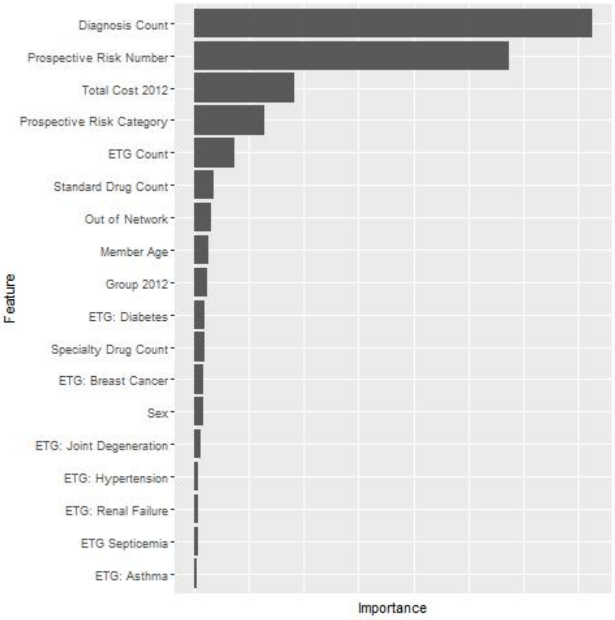

Figure 1: Feature importance plot

representation of the improvement in accuracy that results from including each covariate in the model. The feature importance plot for our fitted model is shown in Figure 1.

As can be seen in Figure 1, diagnosis count was the most important covariate in predicting high-cost members. While a member's total cost in 2012 is significant in predicting 2013 costs, the diagnosis count and risk numbers provide important information that potentially help differentiate between those who are consistently high-cost and those who had an isolated but expensive medical event in 2012. The most significant ETG code was for diabetes, followed by hypertension.

Table 5 includes the quantiles associated with the risk scores calculated for each member using XGBoost.

Figure 2 displays the confusion matrix calculated at a standard threshold level of 0.5 over the test set. Most members are correctly identified as low-cost. Although there are many high-cost members who are predicted to be low-cost, the majority of members who are predicted to be high-cost are actually high-cost.

While a threshold value of 0.5 produces adequate results (a positive predictive value of 0.6305 and a sensitivity of 0.1833), the arbitrary threshold offers no insight into the true financial impact of these predictions. Thus, it is important to employ the asymmetric cost matrix (Equation 1).

If the estimated savings per dollar of intervention is 5, then the optimal threshold calculated over the training set risk scores is 0.1281523. The confusion matrix in Figure 3 is formed over the test set risk scores using this threshold.

The difference between the 0.5 threshold confusion matrix and the 0.128 threshold confusion matrix is

Quantile	Risk Score
0.0	0.0000512
0.2	0.0000761
0.4	0.0000845
0.6	0.0001332
0.8	0.0004146
1.0	0.9597558

Table 5: Risk Score Quantiles

shown in Figure 4.

When compared to the confusion matrix at a threshold of 0.5, it is clear that with the threshold of 0.128, we were able to capture more true high-cost members in our predictions. However, we also predicted many low-cost members to be high-cost. Thus, when performing these matrix optimization techniques, it is important that the value for the expected savings per dollar of intervention is as accurate as possible so that the savings from identifying an increased number of truly high-cost members more than offset the cost of extra interventions.

$$\begin{array}{rc} & \begin{array}{cc}\text{Negative Prediction} & \text{Positive Prediction}\end{array} \\ \begin{array}{r}\text{Actual Negative}\\ \text{Actual Positive}\end{array} & \begin{pmatrix} 643,388 & 126 \\ 958 & 215 \end{pmatrix}\end{array}$$

Figure 2: Test set predictions at a standard threshold of 0.5

$$\begin{array}{rc} & \begin{array}{cc}\text{Negative Prediction} & \text{Positive Prediction}\end{array} \\ \begin{array}{r}\text{Actual Negative}\\ \text{Actual Positive}\end{array} & \begin{pmatrix} 642,479 & 1035 \\ 623 & 550 \end{pmatrix}\end{array}$$

Figure 3: Test set predictions at a threshold of 0.128 (corresponding to \$5 savings per \$1 of intervention)

$$\begin{array}{rc} & \begin{array}{cc}\text{Negative Prediction} & \text{Positive Prediction}\end{array} \\ \begin{array}{r}\text{Actual Negative}\\ \text{Actual Positive}\end{array} & \begin{pmatrix} 909 & -909 \\ 335 & -335 \end{pmatrix}\end{array}$$

Figure 4: Difference between test set predictions at thresholds of 0.5 and 0.128

Because the majority of risk score values are small and close together, a small change in threshold can result in drastic changes in net savings. Figure 5 displays the thresholds optimized according to Equation 1 for different values of savings per dollar of intervention.

Clearly, when there are no savings per dollar of intervention, the optimal threshold approaches one, meaning that no one should be assigned the intervention. Conversely, as savings per dollar of intervention

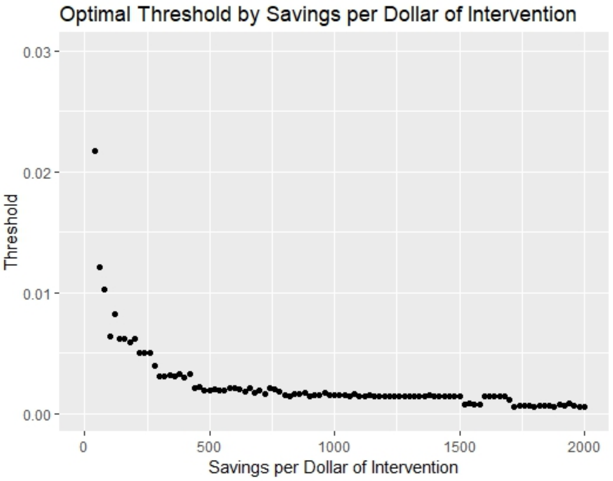

Figure 5: Selected threshold at different values of savings per dollar of intervention

increase, the optimal threshold approaches zero. One may notice a few points on the plot in which the threshold appears to increase with increasing savings per dollar of intervention. In these locations, there are no risk scores between the lower and upper points, so the optimization algorithm selects the lower and upper threshold value with equal probability.

The financial and action-based trade offs at each selected threshold can be seen in Figure 6.

As the savings per dollar of intervention increase, the number of true high claimants who receive cost-saving interventions increases. However, the number of members who participate in unnecessary interventions also increases. Conversely, as the savings per dollar of intervention decreases, fewer high-cost members receive potentially beneficial interventions, but less money is also wasted in interventions on low-cost members.

5. Conclusion

When insurance companies take action to both maximize profits and increase member wellness, careful member selection for intervention programs is critical. XGBoost is one possible algorithm whose iterative nature and flexible attributes result in relatively accurate risk scores. Savings, however, are a direct result of the risk score threshold chosen for intervention. Asymmetric cost matrices offer a solution to threshold optimization. The matrices' inherent simplicity facilitates intuitive and frequent updates to the model, an important feature as wellness programs continue to evolve. Additionally, the asymmetric cost matrix is easily adaptable for optimizing disease-specific intervention outcomes.

More research should be done to evaluate the effectiveness of this model for intervention selection in practice. The data used in this analysis was cleaned to contain only members with 12 member months in 2012 and 2013. Thus, it excludes perhaps some of the costliest members: those who died during the year. Additional research should be done to determine how the inclusion of members who passed away in the time of study affects the methodology.

One could possibly also improve the results of the algorithm by tuning the XGBoost hyperparameters using a cost matrix as the function for optimization rather than accuracy at a level of 0.5. This was explored, but found to be computationally intensive with only marginal improvements to member selection. Furthermore, it does not facilitate the flexible adjustments to the threshold, an important feature to the proposed model. However, additional research could be done to improve processing speed, making the incorporation of the cost matrix into the boosting algorithm more feasible.

Further research should also be performed to evaluate the effectiveness of modifying cost matrices for different circumstances. For example, the threshold for intervention can be optimized over different blocks of

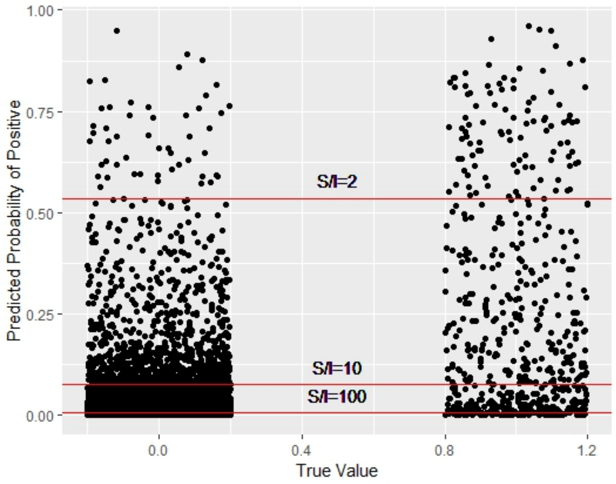

Figure 6: A visualization between the trade-offs associated with selected thresholds

members (such as populations with a given diagnosis or in a certain age group) according to the estimated savings of potential interventions. The matrices can also be fine-tuned to include some savings for those who were actually low-cost members but who received an intervention. Although low-cost members enrolled in an intervention program may not capture all the possible savings that a true high-cost member would, there may be some benefits from the intervention that could slightly alleviate the initial cost of intervention. Additionally, researchers could explore the use of hierarchical matrices to select members for different, non-mutually exclusive interventions. With proper selection methodology, high-cost members will receive the help they need while reducing costs for the member and insurer.

6. Acknowledgments

The authors are grateful to Rebecca Owen of HealthCare Analytical Solutions, Inc. for her suggestions and support.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., Escobar, G., 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Affairs 33 (7), 1123–1131.

Billings, J., Dixon, J., Mijanovich, T., Wennberg, D., 2006. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. Bmj 333 (7563), 327.

Billings, J., Mijanovich, T., 2007. Improving the management of care for high-cost medicaid patients. Health Affairs 26 (6), 1643–1654.

Duncan, I., Beatty, B., Day, B., 2011. A risk-based evaluation methodology for cost effectiveness of chronic condition health management programs. North American Actuarial Journal 15 (1), 1–12.

Duncan, I., Loginov, M., Ludkovski, M., 2016. Testing alternative regression frameworks for predictive modeling of health care costs. North American Actuarial Journal 20 (1), 65–87.

Eichler, H.-G., Kong, S. X., Gerth, W. C., Mavros, P., Jönsson, B., 2004. Use of cost-effectiveness analysis in health-care resource allocation decision-making: how are cost-effectiveness thresholds expected to emerge? Value in health 7 (5), 518–528.

Fleishman, J. A., Cohen, J. W., 2010. Using information on clinical conditions to predict high-cost patients. Health services research 45 (2), 532–552.

Free, C., Phillips, G., Galli, L., Watson, L., Felix, L., Edwards, P., Patel, V., Haines, A., 2013. The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review. PLoS medicine 10 (1), e1001362.

Frees, E. W., Gao, J., Rosenberg, M. A., 2011. Predicting the frequency and amount of health care expenditures. North American Actuarial Journal 15 (3), 377–392.

Granger, C. W., 1969. Prediction with a generalized cost of error function. Journal of the Operational Research Society 20 (2), 199–207.

Hernández-Orallo, J., Flach, P., Ferri, C., 2012. A unified view of performance metrics: translating threshold choice into expected classification loss. Journal of Machine Learning Research 13 (Oct), 2813–2869.

Hoffmann, V. P., Ahl, J., Meyers, A., Schuh, L., Shults, K. S., Collins, D. M., Jensen, L., 2005. Wellness intervention for patients with serious and persistent mental illness. The Journal of clinical psychiatry 66 (12), 1576–1579.

Hong, L., Martin, R., 2017. A flexible bayesian nonparametric model for predicting future insurance claims. North American Actuarial Journal 21 (2), 228–241.

Kim, K., Rosenberg, M. A., 2018. The role of unhealthy behaviors on an individual's self-reported perceived health status. North American Actuarial Journal, 1–18.

Lally, N. R., Hartman, B. M., 2016. Predictive modeling in long-term care insurance. North American Actuarial Journal 20 (2), 160–183.

Lee, S. C., Lin, S., 2018. Delta boosting machine with application to general insurance. North American Actuarial Journal, 1–21.

Rosenberg, M. A., Farrell, P. M., 2008. Predictive modeling of costs for a chronic disease with acute high-cost episodes. North American Actuarial Journal 12 (1), 1–19.

Scott, C., et al., 2012. Calibrated asymmetric surrogate losses. Electronic Journal of Statistics 6, 958–992.

Sokol, M. C., McGuigan, K. A., Verbrugge, R. R., Epstein, R. S., 2005. Impact of medication adherence on hospitalization risk and healthcare cost. Medical care, 521–530.

Sun, Y., Kamel, M. S., Wong, A. K., Wang, Y., 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition 40 (12), 3358–3378.

Wennberg, D. E., Marr, A., Lang, L., O'Malley, S., Bennett, G., 2010. A randomized trial of a telephone care-management strategy. New England Journal of Medicine 363 (13), 1245–1255.