An Introduction to Predictive Modeling for Disease Management Risk Stratification

MICHAEL S. COUSINS, Ph.D., LISA M. SHICKLE, M.S., and JOHN A. BANDER, M.S.

ABSTRACT

Predictive modeling tools are used by disease management programs to risk-stratify members in order to optimize the utilization of available clinical resources. This paper provides an introduction to predictive modeling within the context of disease management by describing how predictive modeling tools can be used, how they work, and how modeling results should be evaluated. These topics will be described at a relatively general level to provide a basic understanding for non-statisticians.

INTRODUCTION

Chronic disease is the leading cause of illness, disability, and death in the United States. In the year 2000, approximately 125 million Americans had one or more chronic conditions. It is projected that in 2020, 157 million, or approximately 50% of Americans, will have one or more chronic illnesses. Healthcare costs for people with these conditions are staggering, exceeding $500 billion per year. Fortunately, clinical interventions by disease management programs are effective at controlling these costs and improving health outcomes. One way that disease management programs can improve their effectiveness is to focus resources (i.e., care managers, pharmacists, physicians) on the care of patients who are at risk for their condition(s) to worsen. Ideally, patients should be risk-stratified to identify opportunities for intervention before the occurrence of adverse outcomes that result in increased medical costs. The use of predictive modeling tools is one way that some disease management programs are efficiently allocating resources to make their programs more successful and cost-effective.

Predictive modeling tools incorporate mathematical formulas that allow users to interpret historical data and make predictions about the future. More specifically, these tools are used to create a predictive model by mapping associations and their statistical relationships among data elements to a specific target. The empirically derived model is then used to forecast future events based upon the identified relationships. These models are adaptable, can effectively utilize most any data source, and have proven useful in a variety of disciplines and industries, including marketing and merchandising, engineering, finance, psychology, medicine, and healthcare.

In disease management, predictive models can help to improve staff efficiency, reduce administrative costs, focus patient counseling and education efforts, and improve financial and health outcomes. The model identifies rela-
Although Pareto's principle provides a basis to explain the disproportionate distribution of healthcare costs in a given year, the problem of identifying members who will accrue the most costly claims in future years remains. The members who are in the small segment of interest (e.g., highest costs or highest utilization) in any given year may not be among the group that will account for the majority of costs the following year. Because of this, methodologies that risk-stratify members based on current events or medical costs in order to select members for high levels of intervention in subsequent years may prove to be inaccurate.

It is common for some disease management programs to select a high-intervention group by including any member who meets a specified criterion or threshold for a parameter of interest. This is referred to as a rules-based, criterion-based, or threshold-based model. A program may provide intensive management to individuals who exceed a threshold based on frequency or type of adverse medical events or conditions, risky medical procedures, or hospital readmissions. Thresholds are also often set on overall healthcare costs or costs related to a specific condition. For example, a program interested in reducing readmission rates among survivors of heart failure may focus clinical resources on patients who have high-risk factors for readmission such as prior heart failure, recent hospital admissions, diabetes, and/or creatine levels that are above threshold. Another program may allocate re-
Selection bias occurs when members are selected because they are outliers who represent some extreme or exceed a given threshold. For example, selection bias occurs if the average medical claims costs for a health plan are $500 per year and members are selected for intervention if they exceed a threshold of $5,000 per year. This is a simple example using one variable, costs per year, but selection bias is inherent to the threshold-based approach even if the model uses hundreds of variables.

A significant problem with models that suffer from selection bias is a concept known as regression to the mean, or average. Regression to the mean refers to the fact that, over time, the measurement of a given event moves toward the average. Therefore, the problem with selecting members for intervention who represent some extreme is that over time there is a high likelihood that the members will show improved health, reduced medical costs, or better outcomes, even without intervention. This reduces opportunities to effectively manage the appropriate patients and makes it difficult to truly evaluate the effects of interventions.

Examples of how models that are subject to selection bias can be associated with regression to the mean are relatively common whether the threshold is based on hospital readmission rates, HbA1c values, or medical claims costs. For example, this year’s top 1% of members account for 30% of medical claims costs this year, but next year they will account for just 15% of medical claims costs. This year’s bottom 80% of members account for only about 15% of claims costs this year, but next year they will account for about 50% of claims costs. The exact percentages vary across health plans, but the general point remains, that members with measurements at the extreme one year rarely have extreme measurements the following year. These extreme values regress over time toward the group’s mean.

There are two important implications of selection bias and regression to the mean that are associated with threshold-based models. The first deals with inappropriate allocation of resources. Assume that a disease management program is interested in reducing claims costs for diabetic members with the highest medical claims costs. The program’s administrators calculate that the average claims costs for diabetics is $5,000 per year and may target members for high intervention who exceed a threshold of $20,000 in a year. Although this is a reasonable approach, it highlights two fundamental problems with the threshold-based method. First, it is likely that members chosen for the highest level of intervention based on this year’s claims costs will have reduced claims costs next year (i.e., will regress to the average) even without intervention. Indeed, by choosing members in this manner, it is likely that the highest level of intervention is provided to the wrong members. Therefore, application of this method would likely result in misallocation of resources, missed opportunities to manage members who had high costs, and generally unfavorable net financial and clinical outcomes. The second implication centers on program evaluation. Unless there is an unmanaged control group to use as a reference for the high-cost members, program evaluators could mistakenly attribute reduced claims costs for these members to the high-intervention program and incorrectly believe that opportunities for management were maximized. In fact, the threshold-based approach almost ensures that the managers of a disease management program would mistakenly believe they are succeeding at improving outcomes, or worse, unfairly take credit for improved outcomes, when in reality the wrong members were targeted for intervention.

Selection bias and regression to the mean are problematic regardless of the type or number of thresholds. For example, the flaws described above with a single measure such as medical costs wholly apply to other single measures such as hospital readmission rates or HbA1c. Regression to the mean will be a deleterious factor if congestive heart failure members are...
selected for intervention based on having above average hospital readmission rates or diabetic members are selected for intervention based on having HbA1c levels greater than 8%, which is well above the average HbA1c levels for diabetics. Combining various thresholds, such as risk-stratifying diabetics who have greater than $20,000 in medical claims costs and HbA1c levels above 8%, into high-intervention treatment helps to reduce the number of incorrect members receiving high intervention but does not mitigate the limitations that arise from regression to the mean. In summary, because the threshold-based method incorporates selection bias when selecting members for intervention—whether based on a single criterion or dozens of criteria—there is a high probability of regression to the mean.

PREDICTIVE MODELS VERSUS THRESHOLD-BASED MODELS

Predictive models have advantages over threshold-based models because they avoid selection bias and subsequently regression to the mean. They are also generally considered to be more accurate than threshold-based models. Selection bias and regression to the mean do not occur with predictive models because predictive models do not risk-stratify members by identifying members who are outliers. This lack of selection bias obviates the possibility of regression to the mean. The accuracy of predictive models is supported by numerous peer-reviewed publications, including those applicable to disease management that describe risk stratification for patients with asthma, diabetes, or coronary artery disease. However, we are not aware of any published, peer-reviewed articles that explicitly compare the accuracy of threshold-based models with that of predictive models. Therefore, the increased accuracy of predictive modeling tools is either anecdotal or inferred from the fact that these tools do not suffer from selection bias. This inference is reasonable but should be validated in published peer-reviewed studies.

One unpublished study compared the accuracy of a predictive model with a threshold-based model (Cousins and Bander, unpublished observations). To ensure that the results from the different methods could be meaningfully compared, the models had the same objective: to identify, using medical claims data, the 15% of diabetics who would be most expensive in the subsequent year. Both methods were applied to the same population of approximately 11,000 diabetic members of a health management organization. The threshold-based model was a multicriteria model that reviewed total medical costs and emergency room, inpatient, and outpatient utilization patterns. The predictive model was based on a simple regression tool and used the same cost and utilization information, as well as other claims-based data such as diagnosis and prescription codes. This comparison showed that the predictive model’s accuracy, as measured by the true positive rate (i.e., sensitivity), was approximately twice that of the threshold-based model (21% vs. 43%). These results must be interpreted cautiously for the following reasons. First, other measures of model performance such as rates of true negative, false negative, and false positive, which will be described below, are unavailable. Thus, although the true positive rate is substantially higher with the predictive model, it is important to look at these additional parameters to be certain that the model’s performance is optimal. Second, as mentioned above, these are not published data that have benefited from the scrutiny of peer review. Therefore, although this study substantiates the increased rigor of predictive modeling tools, caution should be used when interpreting or generalizing these results.

In contrast to the paucity of published data that support the greater accuracy of predictive models over threshold-based models, there is some evidence that predictive models are more accurate than the predictions of clinical experts. For example, compared with clinical experts, predictive models have demonstrated higher accuracy in predicting newborns at risk for sudden infant death syndrome and in diagnosing progressive brain dysfunction. These findings are of interest because many threshold-based models are built using clinical judgment.

In summary, although peer-reviewed pub-
Predictive modeling tools are the logic or software used to create predictive models. Once a predictive model is constructed and tested, it can be applied and used to predict the future event of interest. Essentially, a predictive modeling tool builds a model by examining the historical relationships between known data and a variable of interest. Typically, in disease management programs, these data are derived from a plan’s healthcare claims of members identified with managed conditions (i.e., diabetes, end-stage renal disease, etc.). Once the historical relationships are known and the model is built, it is applied and used to forecast the targeted variable. The primary steps involved in model construction are: (1) the identification of a target variable, (2) selection of time periods to use for development, (3) review of candidate (i.e., potential) drivers or predictors, and (4) selection of the significant drivers from the candidate drivers and determination of weights. This construction and application process is outlined graphically in Figure 1 and explained conceptually in the following sections. Examples of the construction and application of predictive models are described in more detail in several published articles.19-22

Selecting a target variable is the first step in the construction process. The target variable is an element or event that the model will be developed to predict. For disease management programs wishing to focus resources on enrollees with the greatest risk for adverse outcomes, target variables could be most anything, including overall or disease-related costs (where cost is a proxy for health status), comorbidity, or a specific health event. In this description, the target variable for the model will be defined as next year’s predicted medical costs (Predicted Target Year +1 in Fig. 1C).

The second step in building a predictive model, selection of the time periods, is the “when” for the target variable. The users must determine whether the goal is to forecast medical costs six months, one year, or several years into the future. In the construction of the model the actual, not predicted, historical medical costs are totaled for each member for Year −1,
Before a model is adopted as final, it is critical that it is tested. One way to test the model is to assess its performance with data that are different from those used during the construction process. This different dataset is known as a "holdout" or "validation" dataset that contains data not used by the modeling tool to build the model (e.g., Lieu et al.27).

It is typically created before the model-building process is started and usually consists of a randomly selected sample of members who are withheld from the modeling dataset. It is important that the model is evaluated on the holdout dataset, because extreme values or outliers present in

which is the 1-year period prior to the forecast period (Fig. 1B).

For step 3, candidate drivers are assembled for two years prior (Year -2) to the year for which the target is to be predicted. Of the historical data available for review, most any information can be a potential, candidate driver. This could include demographic information such as gender, marital status, and age, or information from medical, pharmacy, and laboratory claims such as medical and drug costs, diagnoses, medical procedures, major surgeries, laboratory results, co-morbidities, and utilization patterns. Other sources of data include information from electronic medical records or health surveys. It is prudent to consider all available sources of data, since the ability of the model to accurately predict the target variable increases with the availability of accurate and complete candidate drivers.

For step 4, the modeling tool statistically evaluates all the available data in Year -2, which at this stage are "candidate" drivers, in relationship to the actual target in Year -1 through an iterative process of "testing" and "learning" (Fig. 1A and B). The candidate drivers are evaluated alone and in combination, and weights are assigned that indicate the influence of the candidate drivers. The magnitude of the weight is proportional to the relationship of the driver with the target. The modeling tool identifies the combination of Year -2 drivers and weights to optimize the prediction of the Year -1 actual target (Fig. 1A and B). Since the actual target in Year -1 is already known, the iterative testing and learning process continues until the best drivers and optimal weights are selected that best "predict" the actual target. When the process is complete, the product is a predictive model. An example of a hypothetical model with the selected drivers and weights is shown in Figure 2.

**EVALUATING A PREDICTIVE MODEL**

Before a model is adopted as final, it is critical that it is tested. One way to test the model is to assess its performance with data that are different from those used during the construction process. This different dataset is known as a "holdout" or "validation" dataset that contains data not used by the modeling tool to build the model (e.g., Lieu et al.27). It is typically created before the model-building process is started and usually consists of a randomly selected sample of members who are withheld from the modeling dataset. It is important that the model is evaluated on the holdout dataset, because extreme values or outliers present in
the data used to build the model may have overinfluenced (or underinfluenced) variable selection and weighting. The holdout dataset will have different members with different extreme values. Evaluating the model on the holdout dataset will reveal whether the model will perform equally well using the new data as on the data used to build the model. If the model performs well with both datasets, this suggests that it will successfully generalize over time to new members, including members added to a plan because of churn. Models that do not generalize are said to “overfit” the original dataset used in the construction of the model. These models must be rebuilt.19

To this point, a model’s performance has been described using general terms such as “accuracy” and “performance.” But a description of a model’s performance that is general can be confusing at best and misleading at worst. Therefore, when comparing the accuracy of models between or within disease management programs, it is critical to use definitions of accuracy or performance that are precise. There are several ways to precisely describe the performance of a predictive model.13 For example, models can be evaluated in terms of $R^2$ or on what is termed a receiver operating characteristic (ROC) curve. An $R^2$ value and ROC curve describes the performance of a model over the whole range of risk levels for the population identified with the condition. Although these are important measures, a better assessment can be achieved by examining the model’s accuracy at the desired intervention cutoff level. Since the goal of predictive modeling in a disease management setting is to identify highest risk members so that those members can receive the program’s highest level of intervention, it is critical to evaluate the model’s success at predicting the members for that group.22

The model’s performance at accurately predicting the health risk of low-risk members who (by definition) are in the low-intervention group is irrelevant.

Four metrics that precisely describe the accuracy of a model for a given intervention cutoff point are the true positive rate (i.e., sensitivity), true negative rate (i.e., specificity), false positive rate, and false negative rate (e.g., Ash et al.10 and Lieu et al.27). There are two important points to keep in mind when examining these rates. First, the rates are interrelated, and the most complete picture of a model is described by considering all four measures. Second, the rates need to be defined together with the cutoff point for the intervention level (Fig. 3). For example, the target could be the 5% of diabetics predicted to be next year’s most expensive, and these members would receive the program’s highest level of intervention. But if the cutoff point for intervention changes, the measures described above change.

A model’s true positive rate indicates how well the model has accurately identified the target. A higher true positive rate is desired—ideally, a rate of 100%—so that members predicted to be highest risk or cost actually were highest.
as targets (which decreases the false positive rate). A model that performs well has low false negative and false positive rates.

As mentioned above, the true positive, true negative, false positive, and false negative rates change as the cutoff point for high intervention changes. This is shown graphically in Figure 3. The cutoff point for the highest level of intervention increases from Figure 3A to D. As this cutoff point increases from 5% (Fig. 3A) to 10% or 15% (Fig. 3B and C, respectively), the true positive rate increases and the false negative rate decreases, even though the performance of the model is unchanged. But the increase in the true positive rate is at the expense of an increase in the false positive rate. If all else is constant, as the cutoff point for intervention increases, the true negative rate decreases and the false positive rate increases. Finally, as shown in Figure 3D, if the cutoff point for intervention is moved to include all members with a condition, the true positive rate is 100% and the false negative rate is 0% (with the true negative and false positive rates in between). These rates may seem ideal, but the cutoff point for the highest level of intervention includes the entire population.

These measures described above should be determined during testing before a model is implemented. This can be accomplished by using the holdout dataset, as described above, on a randomly selected subset of members where the actual target is known.\textsuperscript{19,20} In addition, these rates can be determined after a model is implemented.

### FIG. 3. Effect of moving intervention cutoff points on rates of true positive, true negative, false positive, and false negative. Cutoffs at: (A) top 5%, (B) top 10%, (C) top 15%, (D) none.
and the actual target is known, in order to evaluate its performance on an ongoing basis. This latter determination is useful as an assessment of the model’s real-world performance.

These rates are crucial to aid in understanding how well the predictive model assists in the allocation of resources. Since disease managers wish to focus limited resources towards members in greatest need of assistance, the accuracy of the model as defined by these measures affects whether the members at highest risk for future adverse outcomes are targets for the highest level of intervention. Likewise, an accurate model excludes members with low risk for adverse outcomes from being targets for the highest level of intervention. This helps to reduce the improper use of limited resources. Ideally, returning to Pareto’s principle, disease management efforts aim to improve the outcomes of the “20%” who account for “80%” of the costs and adverse events. The model’s true positive rate, for example, explains whether the members included in the target group and identified for high intervention are the “correct 20%.”

Since the success of disease management programs is dependent upon changing the behaviors and improving the health status of the members, it is critical that members predicted to be high-risk by the model are “intervenable.” Ideally, the model identifies high-risk patients who have conditions, utilization patterns, or events that are preventable or manageable. For example, does a member with end-stage congestive heart failure who is already compliant with optimal medical therapy, but who has been predicted to be high-risk or high-cost because of the severity of his or her condition, represent a viable opportunity for the disease management program? Questions such as this must be asked when reviewing the output from the predictive model. Otherwise, the disease management organization’s resources may be misappropriated to individuals who, although predicted to be high-risk, have conditions that are not likely to be improved by care managers.

**TYPES OF PREDICTIVE MODELS**

There are several predictive modeling tools that are useful for disease management applications, including time series models, classification tree models, linear and non-linear regression models, and neural network models.6,19-22,25,28 Each has strengths and weaknesses depending upon the specific application and particulars of the data. Time series modeling tools are useful to map changes over time. Non-linear regression and neural network modeling tools are useful to map non-linear relationships. Neural network modeling tools are useful because interactions between predictors do not need to be explicitly identified. Classification tree models are useful when data are incomplete and contain a lot of missing values.

Of these, there has recently been particular interest in neural network tools. A few published healthcare studies have evaluated neural network models against others, notably regression models. In general, the neural network models perform similarly or slightly better than regression models. One study attempted to predict the likelihood of 5-year survival after first-course treatment for colon cancer.29 The other attempted to determine whether pre-angioplasty patient risk factors, demographic characteristics, and procedural information can be used to predict major complications after coronary angioplasty in patients with end-stage renal disease.30 Both studies reported that the neural network model had a slightly higher true positive rate and similar true negative rate to the regression model. Although neural networks performed slightly better than the regression models, it must be noted that neural networks present some difficulties in implementation because they are more complicated. In addition, the neural network does not yield clinically useful information about the predictors (for example, see Fig. 2). Indeed, the investigators in the latter study, as well as others, report this as a specific shortcoming of the neural network model.5,30 Therefore, although the neural network model performed as well or slightly better than the regression model, the regression model identified specific drivers that were significantly associated with the target and provided estimates of the risk via the weights assigned to the drivers. This information may be useful to clinicians in designing interventions. Whether this information is useful relative to differences in performance illus-
trates the tradeoffs that one must make when choosing a predictive modeling tool.

CONCLUSION

The appropriate allocation of healthcare resources to members in highest need of intervention is essential for a disease management organization to be successful. Predictive modeling tools can be extremely useful in this regard and offer clear advantages over traditional threshold-based models. Since predictive models are empirically derived from the data, the model's performance increases as more comprehensive data are incorporated into building the model. Thus, with the steadily increasing availability of healthcare information—such as from electronic medical records and laboratory results from more accurate, diagnostic, and prognostic tests—the potential for predictive modeling tools to contribute to the success of disease management organizations increases.

REFERENCES

25. Leli D, Filskov S. Using SPRs to diagnose brain dys-

Address reprint requests to:
Michael S. Cousins, Ph.D.
Health Management Corporation
6800 Paragon Place
Richmond, VA 23230

E-mail: mcousins@choosehmc.com