Accuracy of Prediction Models in the Context of Disease Management

Authors
GUIZHOU HU and **MARTIN ROOT**

Abstract
There has been a significantly increased interest in the adoption of prediction modeling by many disease and case management programs to risk stratify members in order to optimize the utilization of available clinical resources. Before adopting any prediction model, it is crit- ical to understand how to evaluate the model's accuracy. This paper explains the basic con- cepts of prediction accuracy, the relevant parameters, their drawbacks, and their interpreta- tions. It also introduces a new accuracy parameter termed "cost concentration," which indicates the model accuracy more explicitly in the context of disease management.

GUIZHOU HU & MARTIN ROOT(2005) Accuracy of Prediction Models in the Context of Disease Management. *Disease Management*, (8:42-47) (ISSN1093-507X) copy available @ (http://www.liebertonline.com/doi/pdf/10.1089/dis.2005.8.42)

Accuracy of Prediction Models in the Context of Disease Management

GUIZHOU HU, Ph.D., and MARTIN ROOT, Ph.D.

ABSTRACT

There has been a significantly increased interest in the adoption of prediction modeling by many disease and case management programs to risk stratify members in order to optimize the utilization of available clinical resources. Before adopting any prediction model, it is critical to understand how to evaluate the model's accuracy. This paper explains the basic concepts of prediction accuracy, the relevant parameters, their drawbacks, and their interpretations. It also introduces a new accuracy parameter termed "cost concentration," which indicates the model accuracy more explicitly in the context of disease management.

INTRODUCTION

It is a well-known fact that a relatively small segment of the population is responsible for the majority of medical costs. For example, 20% of the population may be responsible for 80% of medical costs.[1] Based on this fact, disease management enterprises could manage a large portion of costs by directing resources toward a small segment of the population. However, the current high-cost members are not necessarily the high-cost members of the future. Therefore, the object of using prediction models in disease management is to try to identify the small segment of the population that will be responsible for a bigger percentage of the cost before it happens, so that preventive action can be taken.

In this context, the prediction model with higher accuracy should be the model which identifies the smaller segment of the population that is responsible for a bigger percentage of the overall cost before costs are incurred. However, there is no established statistical term that explicitly evaluates prediction model accuracy in this way. Therefore, we define a new term we are calling "cost concentration." Before introducing this new model-accuracy term, it is necessary to explain the traditional parameters used to evaluate prediction accuracy and to explain their inherent shortcomings in the disease management context.

TRADITIONAL PARAMETERS

Individual R-Squared (R^2) for continuous outcomes

When the outcome of the prediction model is a continuous variable, such as money spent, the prediction accuracy is determined by measuring how close each individual predicted cost is to the actual observed cost.[2] The statistical term for the accuracy measure is

R-squared (R^2). R^2 indicates the percentage of the total variation among individual observed costs that can be explained by the model. Its value ranges between 0 (0%) and 1 (100%). A value of 0 indicates that the model explains none of the variation in the costs, while a value of 1 indicates that the model explains all of the variation. This perfect fit, in which each predicted cost equals each actual cost is, in reality, unattainable due to many random and unpredictable factors.

R^2 is a single summary measure of predictive accuracy. Its usage has been widely accepted and understood in the statistical field. In the context of disease management, however, R^2 has several drawbacks. First, it reflects the model accuracy across the whole range of cost levels, while in disease management the interest is on the high-cost end only.[3] In other words, we are more interested in how accurate the model is in terms of identifying high-cost users. We are less concerned about how closely the model rank orders members in the low-cost range. R^2 cannot differentiate the model accuracy at the high- or the low-cost end of the range. Second, R^2 tends to be overly sensitive to the prediction error for individuals with very high costs because it squares the errors of prediction. This is a very important concern because health expenditures typically have a very skewed distribution, where a small number of members have relatively large expenditures. To overcome this problem, it has been suggested that the R^2 should be calculated after truncating large medical expenditures.[4] For example, we might set the ceiling value at $25,000 or $50,000. The Society of Actuaries also suggests a new parameter called the "mean absolute prediction error."[3] Instead of squaring the error, mean absolute prediction error calculates the absolute value of the error, making it less sensitive to outlier values.

Last, because of the unpredictable nature of medical expenditures, the R^2 of even the best prediction model has an appearance of poor performance. For example, the R^2 of a prediction model using administrative claims data is typically around 10%–20%. As a result, it can easily be interpreted that the model is only about 15% correct, or that it is about 85% wrong in predicting cost. This causes the healthcare decision makers to question the value of the prediction model, that is, "Why invest in an expensive and complicated process that explains, at most, only 15% of the variation in costs," or "Why use the prediction model to identify disease management candidates when the model is only 15% right?" This misconception is caused by the flaw of using R^2 to evaluate the accuracy of the prediction model. In the context of disease management, the objective of using a prediction model is to try to identify the right group of members with whom to intervene. For example, if there are 100 patients who indeed become high-cost users and a prediction model predicts all of them, and only them, as predicted high-cost members, then this model should be regarded as perfect or 100% accurate. However, it does not necessarily have an R^2 of 100%. In order to get $R^2 = 100\%$, the prediction model not only needs to predict members' future costs in the right order but also to predict the exact amount correctly. The model with only 10%–20% R^2 may still have fairly high accuracy in the context of stratifying risk, and the level of accuracy cannot be directly reflected by the R^2 value. Implying that the model is only 10%–20% right, or 80%–90% wrong, clearly is a wrong conclusion in this context.

Sensitivity, specificity, positive predictive value, and ROC curve for dichotomous outcomes

In the context of disease and case management, the target population of intervention is often predetermined. For example, the top 1% or 5% of the population or anyone with costs over $5000 may be chosen. As a result, the outcome can be defined as a dichotomous variable with two values, one for high cost (or positive) and one for low cost (or negative). In a similar manner, a cut-off point can be arbitrarily set in order to make any predicted cost into a dichotomous variable as the predicted high-cost member and predicted low-cost member.[5,6] Sensitivity, specificity, positive predictive value, negative predictive value and Receiver Operating Characteristic (ROC) curves are a set of accuracy parameters dealing with dichotomous outcome. Because it is so easy to confuse the definitions of those parameters, it is help ful to explain them using mathematical formulas first.

Let us define four table cells as a, b, c, and d, where a is high cost and predicted high cost; b is low cost and predicted high cost; c is high cost and predicted low cost; and d is low cost and predicted low cost (Table 1).

The following formulas then define the four terms noted above:

$$\text{Sensitivity} = a/(a + c)$$

$$\text{Specificity} = d/(b + d)$$

$$\text{Positive predictive value (PPV)} = a/(a + b)$$

$$\text{Negative predictive value} = d/(c + d)$$

Sensitivity can be interpreted as the fraction of all true high-cost members that were predicted as high cost. Specificity is the fraction of all true low-cost members that were predicted as low cost. The opposites of sensitivity and specificity are also termed false negative rate (1 − sensitivity) and false positive rate (1 − specificity). It is important to keep in mind that the sensitivity and specificity both depend on the cut-off point used to define the predicted high cost and low cost. If the prediction cut-off threshold is set lower, then more members are predicted to have high cost, and sensitivity will go up and specificity go down. Therefore it is difficult to compare the sensitivity or the specificity of two prediction models when the cut-off points are different. When the cut-off point is set, sensitivity and specificity are model specific.

The ROC curve is a measurement of model accuracy that is independent of the prediction cut-off point. The ROC curve is the curve of sensitivity versus 1 minus specificity across all prediction cut-off points. The area under the ROC curve indicates the predictive power of the model. Models can now be compared independent of cut-off points. However, it is hard to interpret the meaning of the area under the ROC curve.

Table 1. Defining Table Cells for a Model that Predicts Dichotomous Cost Outcomes

	Actual cost groups	
Predicted cost groups	*High cost*	*Low cost*
High cost	a	b
Low cost	c	d

Positive predictive value (PPV) is the fraction of all predicted high-cost members that are true high-cost members. The most common mistake in using PPV is not realizing that PPV is not only dependent on the model's predictive power but is also dependent on the prevalence of true positives in the population. For example, we may apply a prediction model to two different populations. In one population, the top 1% of costs is defined as high cost. In the other population, the top 5% of costs is defined as high cost. Even when the models' sensitivity and specificity are constant, the PPV will be higher in the population in which the prevalence of true positives is higher (the population where the top 5% is set as high cost). Negative predictive value works similarly with PPV but is seldom used in the risk stratification context because the positives are what we are interested in.

Sensitivity, specificity, and PPV may be more appealing than R^2 as measures of prediction accuracy of a model because they focus on the accuracy of identifying potential high-cost users, which is directly relevant in the disease and case management setting. However, they have many drawbacks as well. First, as mentioned above, sensitivity, specificity, and PPV are all dependent on other conditions that have nothing to do with the prediction accuracy, such as high-cost prevalence and choice of prediction cut-offs. Because of this, it is very difficult to directly interpret the sensitivity and PPV and to compare different models. ROC is independent of the choice of prediction cut-offs, but the area under the ROC curve does not have a direct interpretation in terms of risk stratification. Like R^2, it also measures accuracy over the whole range of costs, while only the high-cost subjects are of interest. Second, similar to R^2, sensitivity and PPV also give the general impression of poor performance to a model. For example, most administrative claim-based prediction models have a PPV of around 30% when both the true high-cost and the prediction high-cost cut-offs are defined as the top 5%. In other words, only about one third of the predicted high-cost members are true high cost (true positives), or two thirds of them are false positives. In the context of disease management, false positives may easily be interpreted as a waste of clinical resources on those who do not need the intervention. This certainly will make any healthcare decision maker question the real value of a prediction model.

This apparent poor performance is caused by the intrinsic flaw of treating the outcome as dichotomous. For example, if we predefine high risk as costs greater than or equal to $2000 a year, then two prediction models will not be differentiated if one predicts a $1999 member as positive and the other predicts a $0 member as positive. This is because they are both regarded as false positives even though there are different magnitudes of misclassification between the two. Simply regarding all false positives as a waste of clinical resources is thus overly strict. In a similar manner, the correct classification also has different magnitude. For example, correctly identifying as high cost a member whose true cost is $2001 or another member whose true cost is $100,000 has the same contribution to the sensitivity and PPV. But they have different disease management implications. The sensitivity, specificity, and PPV are not able to differentiate the magnitude of misclassification (or correct classification). A prediction model could still have value even if it has a PPV of 0 (100% misclassification rate), as long as the misclassified members only miss the cut-off by a small amount. For example, as long as the model can identify a group of members with true costs higher than average, then the prediction model has value (or has some level of accuracy), even when the PPV or sensitivity may turn out to be 0% for a higher cut-off.

NEW PARAMETER

Cost concentration

Because of the drawbacks of the traditional parameters in evaluating the accuracy of prediction models in the disease and case management setting, we introduce a new parameter we are calling "cost concentration." It is defined as the percentage of true cost of the total population that is concentrated among the subset of the population that was predicted to be high cost. For example, a given model may predict 5% of the population as high cost. If a model had no predictive value at all, then we would expect that about 5% of the cost is concentrated in that group. However, if the model was reasonably accurate, then this subset of the population may have a cost concentration of 25%. The difference between the two percentages (from 5% to 25%) indicates the value of prediction model.

Similar to sensitivity and PPV, in order to use cost concentration the prediction needs to be expressed as a dichotomous outcome, namely, predicted high cost versus predicted low cost. This makes it directly relevant in the disease management context. On the other hand, different from the sensitivity and PPV which also expresses the outcome as dichotomous (true positive and true negative), the outcome of cost concentration is expressed as a continuous value of total cost among predicted high cost. This enables it to differentiate the magnitude of misclassification and correct classification.

Cost concentration is expressed in the same way as Pareto's principle of the "20% who account for 80% of the cost." Pareto's principle also defines the highest possible value of cost concentration in a population. For example, a perfect prediction model might identify that 20% of high-cost members who will account for 80% of true future cost. The lowest cost concentration value would be that 20% of the members who account for 20% of the cost, which represents an indifferent model (or random selection model), similar to the indifferent line in an ROC curve.

By analysis of the cost distribution of typical health claim data, we defined the approximate highest cost concentration value at different cut-off points in Table 2, along with the cost concentration value of a typical prediction model based on claim data.

The main advantage of using cost concentration is that it explicitly indicates the value of the prediction model in the context of disease management. For example, if a model has a cost

Table 2. Cost Concentration at Different Prediction Cut-Off Points

Prediction cut-off point (%)	1	5	10	15	20
Highest cost concentration (%)	21	44	58	67	73
Indifferent (lowest) cost concentration (%)	1	5	10	15	20
Cost concentration of a prediction model[a]	8	22	35	44	52

[a] This prediction model has an R^2 of 25% while truncating the cost at $50,000, and both sensitivity and PPV of 31% at the 5% cut-off point and an area under the ROC of 0.82.

concentration of 25% among the 5% predicted high-cost members, then it means a disease management program that uses this model will be able to manage 25% of the total cost by focusing on the 5% of the population. The value of the prediction model becomes intuitively simple. But with R^2 or PPV of 0.3, it can be easily and mistakenly interpreted as 30% correct and 70% wrong, as explained earlier.

Another advantage of the cost concentration concept is its ability to take the magnitude of misclassification and correct classification into account. For example, model A and model B may have the same number of misclassifications and correct classifications and therefore have the same values for sensitivity and PPV. However, model A may have a lower magnitude of misclassification or a higher magnitude of correct classification, and therefore it has a higher cost concentration. In that case, model A should be regarded as a better model.

Cost concentration has two main drawbacks. First, similar to sensitivity and PPV, its value depends on the prediction high-cost cut-off point. Therefore, the cost concentration is only comparable between models in which the cut-off point of predicted high cost is the same. Second, similar to R^2 it is sensitive to outlier value. For example, correctly predicting a $1,000,000 member as a high-cost member may have a strong impact on the cost concentration value, even though that member is just one true positive.

There could be a situation in which model A has higher cost concentration than model B but has significantly lower sensitivity and PPV than model B. There can be two reasons for this. One is that model A has more misclassification but with less magnitude (the false positives have high true cost value but barely make the cut-off point as true positive), and model B has less misclassification but high magnitude (including some $0 cost members as predicted positive). Under this situation, the cost concentration is a better parameter to balance the overall effect of the model, and we can conclude that model A still is better than model B. The other reason that causes this situation is that some outliers of true positive of model A may skew the cost concentration, meaning model A has a smaller number of true positives, but many of them have extremely high-cost value. Judging which model is a better prediction model would be difficult in this situation. The combined information, including R^2, area under ROC, sensitivity, PPV, cost concentration, and clinical consideration, need to be evaluated together to make the judgment.

CONCLUSION

There has been an increase of interest in using prediction models in disease and case management. However, a proper term for evaluating the accuracy of prediction models in this context is lacking. Traditional model accuracy terms such as R-squared, sensitivity, PPV, and ROC all have significant limitations. The new term, "cost concentration," appears to be a more appropriate term in evaluating model accuracy in this context. It is directly relevant to risk stratification, and it is intuitively easy to understand. Due to its own drawbacks, one still needs to be cautious in judging model accuracy solely by cost concentration, especially when there are significant discrepancies between cost concentration and the more traditionally used evaluation parameters.

REFERENCES

1. Buchanan M. Wealth happens. Harv Bus Rev 2002; 80:49–54.
2. Ridinger MH, Rice JJ. Predictive modeling points way to future risk status. Health Manage Technol 2000;21: 10–12.
3. Cousins MS, Shickle LM, Bander JA. An introduction to predictive modeling for disease management risk stratification. Dis Manage 2002;5:157–167.
4. Cumming R, Knutson D, Cameron B., et al. A comparative analysis of claims-based methods of health risk assessment for commercial populations. Research Study Sponsored by the Society of Actuaries, 2002 [On-line]. www.soa.org:80/sections/riskadjfinalreport1. pdf.
5. Ash AS, Zhao Y, Ellis RP, et al. Finding future high-cost cases: comparing prior cost versus diagnosis-based methods. Health Serv Res 2001;36:194–206.
6. Lieu TA, Capra AM, Wuesenberrry CP, et al. Computer-based models to identify high-risk adults with asthma: is the glass half empty or half full? J Asthma 1999;36:359–370.