RGA Tests and Measures the Accuracy of Mortality Risk Scores

Case Studies with Milliman Intelliscript® Risk Scores

Guizhou Hu
Vice President, Head of
Risk Analytics

Mark Ma
Vice President and
Managing Actuary

Taylor Pickett
Vice President and
Actuary

www.rgare.com

Introduction

With the increasing push to incorporate healthcare administrative data, such as pharmacy prescription fills (Rx) and medical billing or claims (Dx), into life insurance underwriting, carriers are taking an increasingly greater interest in the values of risk scores developed from those types of data. Milliman IntelliScript® offers a commercial product that offers risk scores on top of an existing data feed.

Milliman risk scores have evolved over the years, from earlier versions of Rx1.0 and Rx2.2, to the recently released RxDx 3.0. With each updated version, the scores were meant to be more accurate than the previous version as a result of better modeling and/or by integrating additional data inputs. However, a commonly agreed upon and widely accepted methodology and metrics for evaluating the accuracy of such mortality risk scores is yet to be established. In this paper, using Milliman risk scores as examples, we introduced two risk score accuracy concepts that have been widely documented and applied in statistical literatures, with special emphasis on one that was rarely discussed in most published validation studies of Milliman risk scores.

Two different score accuracy concepts

It is well documented in statistics literature[1-3] that there are two accuracy concepts for risk scores like Milliman's: (1) risk discrimination, or how well does the risk score differentiate risk; and (2) risk calibration, or how well does the risk score match actual risk.

Risk Discrimination

Since a risk score is primarily intended for underwriting, it makes sense that most risk score validation studies focus on the concept of risk discrimination. A common way of demonstrating such accuracy is through a lift curve, which shows risk elevation as the score goes up, assuming a higher score means higher risk. Also, a common metric of such lift is the risk differential by folds between top versus bottom score deciles (10 percentile). For example, the RxDx 3.0 risk score creates an approximate 20-fold mortality difference between top and bottom score deciles in a retrospective study data provided by Milliman, as shown in Fig 1. Such a metric is intuitive and a straightforward way of comparing various scores or testing incremental value of updated score versus previous versions. However, a drawback of the metric is its dependency on score distribution in the study population. The score might appear to have better performance when it is tested in a population with wider score distribution. For example, while limiting the Milliman study population to the health insurance line of business (N=3.3 million), the RxDx 3.0 top versus bottom decile score difference is about 14-fold with a mortality difference of 10 fold. On the other hand, a subpopulation of life insurance (N=15.6 million), the score and mortality difference are 26 and 24 fold, respectively. Without considering the score distribution, the model may appear to perform better in one population than the other, but as a matter of fact, they perform very similarly. The difference is more of score distribution difference between populations

Fig 1. Relative AE by RxDx Score Decile Among Cases with Both Rx and Dx Hits (N=23.6 M)

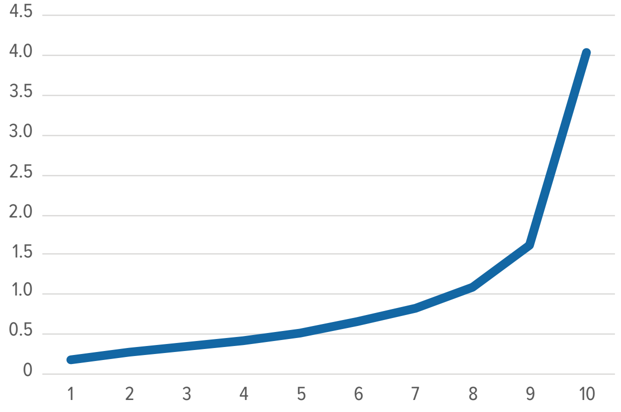
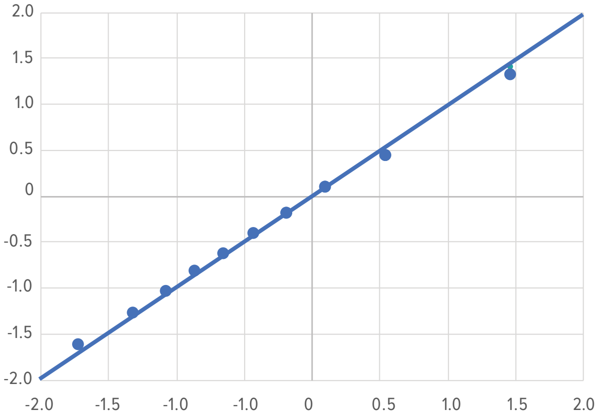


Fig 2. Log(AE) by Log(RxDx score) by Deciles Among Cases with Both Rx and Dx Hits (N=23.6 M)



than the model performance. The example indicates when we try to extrapolate lift information from one population to another, we need to be careful because the lift metric may not be transferable if the distribution of risk score varies.

Risk Calibration

The primary focus of this paper is the introduction of the other risk accuracy concept, risk calibration. This concept can be illustrated by Figure 2, which is a changed version of Figure 1. One of the changes is that the x-axis (axis for score) is changed from a group indicator to score average. This makes the plot a scatter plot. The other change is the log transformation on both axes. The Milliman risk score is a relative risk predictor, and relative risk changes are typically expressed as a mortality ratio or score ratio. For example, we are interested in whether the actual mortality risk is doubled if the risk score is doubled. The log transformation converts the risk difference between any two data points into a ratio. The line shows how close the predicted risk score changes match the actual risk changes, both as a ratio. How well the data align linearly and how close the fitted linear line is to having a 45-degree angle, or having slope of 1, represents the calibration accuracy of the score.

The interpretation of the slope on the plot is the relative mortality changes by a given relative score change. For example, if the slope is 0.95, then the estimated mortality changes would be exp(0.95*log(2))=1.93 fold, if score is doubled, as indicated by log(2).

It is a simple math that a slope can be calculated by any two data points, which is to describe the relation of mortality ratio to score ratio for those two data points. The slope on the Figure 2 plot can be viewed as an average slope across a wide score spectrum, which is defined by score decile in this example.

Why do we care about the score calibration accuracy, and what is the practical use of it?

First, unlike commonly used discrimination metrics (risk fold differential by top versus bottom score decile), the calibration accuracy tends to be independent of score distribution. This characteristic makes it more likely to be transferable across different populations because score distribution is often the main difference between populations.

Secondly, studying the slopes allows us to discover certain unique characteristics of the score behavior, which otherwise may be missed. Figure 3a and 3b show examples comparing slopes between Milliman Rx2.2 vs Rx3.0. It shows some difference, especially at the lower-risk-score range, between the two versions and how Rx3.0 was improved.

Lastly, studying the slopes of two specific groups can provide direct insights for setting mortality assumptions. For example, setting a cut-off point and excluding a subset that meets the cut from an application pool is a typical application of Milliman risk score for either triage or auto-decision. In this use case, the mortality of the excluded subset is a key interest but rarely available due to the difficulty of having a mortality experience study on-hand. By studying the slopes at such cut points in different populations would help to better estimate the mortality when the cut-off point is applied to the target population. For example, we studied the impact of having RxDx score >2 as a cutoff point among three very different sub-populations within the Milliman data. They are all life insurance applicants but have three different Rx drug colors (green, yellow, and red) in the Rx data. As it was showed in Table 1, the averages RxDx score and percentages of cases with scores >2 are quite different among the three populations. First, we found the mortality of scores >2 from the green

Figure 3a Log(AE) by Log(Rx2.2 score)

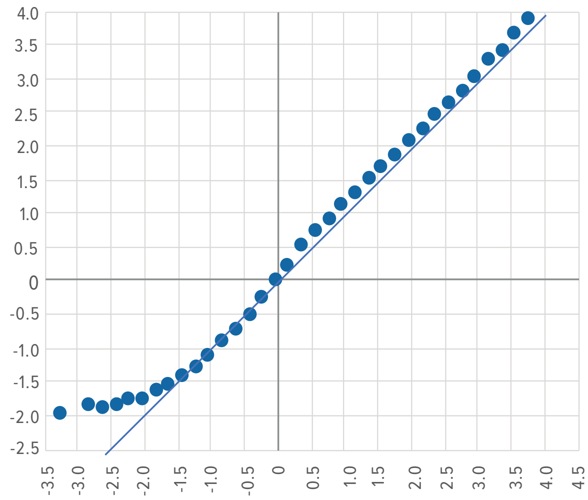


Fig 3b Log(AE) by Log(Rx3.0 Score)

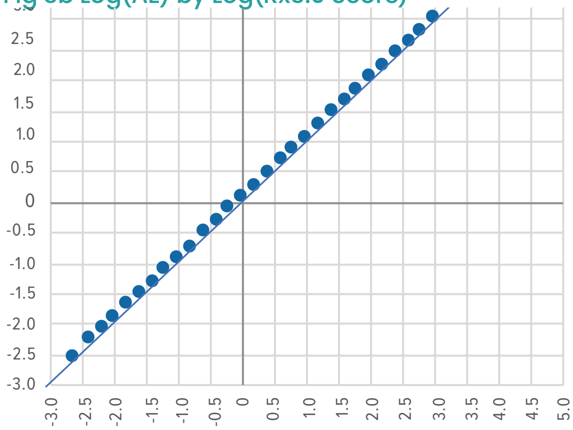


Table 1 Milliman RxDx 3.0 Score Comparisons Among Populations with Different Rx Drug Color Codes

			Cases with Score>2	
Drug Color	Average RxDx 3.0 Score	% with RxDx Score>2	Score Ratio to Overall Average	Mortality Ratio to Overall Average
Green	0.63	2%	375%	351%
Yellow	0.74	5%	347%	348%
Red	2.65	39%	445%	356%

and yellow populations are quite consistent with the score, suggesting a similar slope between the two populations even though the percentage of cases with scores >2 in the yellow population (5%) is more than double the green population (2%). Secondly, although the mortality of scores >2 among the red population does not appear to be as high as the score suggested, this is not a total surprise because many score outliers reside in this group (one case has a score of 934) and how dramatically different this group is from the other two groups. It is somewhat surprising, however, that the slope for this group is not shown to be more different from the other two groups.

Conclusions

To our knowledge, this paper is the first to apply the two concepts of risk score accuracy, which has been widely documented in the field of statistics, into score validation studies for the life insurance industry. The two concepts should help carriers better understand the value of the scores and factors that could impact them while applying the score to a target population. RGA has studied the calibration accuracy of Milliman's scores and other similar scores, such as ExamOne LabPiQture$^{\text{TM}}$ scores, in detail, including factors that could impact them. We are happy to discuss them in detail with those interested. ■

Reference

1. D'Agostino RB, Nam B-H. Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: Balakrishnan N, Rao CR, editors. Handbook of Statistics: Advances in Survival Analysis. San Diego, CA: Elsevier, Inc.; 2004

2. M. Shafiqur Rahman, Gareth Ambler, Babak Choodari-Oskooei and Rumana Z. Omar. Review and evaluation of performance measures for survival prediction models in external validation settings. BMC Medical Research Methodology (2017) 17:60

3. Guizhou Hu, Martin Root and Ashlee W Duncan. Adding multiple risk factors improves Framingham coronary heart disease risk scores. Vascular Health and Risk Management 2014:10 557–562