SOCIETY OF ACTUARIES

Article from:

Health Watch

January 2013 – Issue 71

Uncertainty in Risk Adjustment

By Syed M. Mehmud, Rong Yi and Danielle Bergmeier

Syed M. Mehmud, ASA, FCA, MAAA, is a director and senior consulting actuary with Wakely Consulting Group, Inc. He can be reached at *syedm@wakely.com*.

Rong Yi, Ph.D., is the director, Risk Adjustment and Predictive Modeling Practice, and senior consultant with Milliman Inc. She can be reached at *rong.yi@milliman.com*.

Note: This article is intended to present a high-level review of a recently concluded research project (Mehmud & Yi, 2012) with the same title. The research was funded by the Health Section of the Society of Actuaries. The researchers used Medicare fee-for-service (FFS) 5 percent data sample, the CMS-HCC risk adjustment model and bootstrapping to construct empirical confidence intervals. While the actual results may not be applicable to a non-Medicare population, the methods and procedures described herein are population- and model-neutral. Practitioners may be able to use any risk assessment model and the methodology described in the report to calculate uncertainty-related metrics on their own data and model.

The report, in its entirety, is available at: *http://www.soa.org/research/research-projects/health/uncertainty-risk-adjustment.aspx*

The practice of risk adjustment has long carried significance for Medicare Advantage, Medicaid managed care and commercial health insurance plans. With national health care reform, especially in the areas of health insurance exchanges and provider payment reforms, many more stakeholders are hoping to better understand implications of risk adjustment. No longer the realm of specialists; risk adjustment now concerns most practicing health care actuaries.

The accuracy of risk assessment as measured by statistics such as R-squared, mean absolute prediction error (MAPE) and predictive ratio have been well studied (Winkelman & Mehmud, 2007). Uncertainty in the context of risk assessment is due to the fact that predictions are not perfect. In practice, risk adjustment typically concerns group-level relativities in risk scores, such as a health plan, a provider organization or a physician's panel. Because risk adjustment models are statistically based, the risk scores and predicted expenses should be viewed as point estimates within confidence intervals.

It is worth spending some ink explaining why we should think about a risk score as a distributed variable rather than a point estimate. We are used to a separation of payment from risk in risk adjustment—in other words, risk scores need not necessarily and/or exactly track cost for a particular organization, since that cost will depend upon many things (contracts, efficiency, benefits, etc.). Development of coefficients is, however, different from application and in the former case we *do* want risk to track cost (since that is the objective or dependent variable of the modeling). It is here that we find that predictions do not equal cost, and that the properties of the error term distribution provide a picture of *how confident we can be in the point estimate determined by the software*.

Table 1 shows a simplified example of a risk adjustment calculation. Cells that are affected by uncertainty in risk score estimates are in bold.

We show two hypothetical health plans, A and B, with identical member months. We assume that the projected per member per month (PMPM) health care expense for both plans is $450. The two plans attract different members and therefore have different risk scores and risk-adjusted expenses.

The concepts and methodologies developed in the aforementioned research report allow a practitioner to calculate confidence intervals for risk scores. The last two columns in Table 1 show the range of the 90 percent confidence intervals for the estimated risk and the risk-adjusted expense.

Risk adjustment moves money around. As such, among the many functions that actuaries perform,

Table 1: Illustrative Example of a Risk Adjustment Calculation for Two Health Plans

	Member Months	Projected Expense	Average Risk	Risk Adjusted	Average Risk (90% CI)	Expected Expense (90% CI)
Plan A	5,000		1.03	$463.50*	{0.988 - 1.076}	{$445 - $484}
Plan B	5,000		0.97	$436.50	{0.928 - 1.016}	{$418 - $457}
Total	10,000	$450.00	1.00	$450.00		

* $450 x 1.03 = $463.50

risk adjustment is expected to be highly scrutinized. The questions from stakeholders concerning risk adjustment are expected to grow in their complexity. We studied the following questions as part of this research:

1. When are differences in risk scores statistically significant?
2. How confident can we be that the predictions from a risk adjustment model will be close to the actual values?
3. How does the predictive accuracy of a risk adjustment model affect uncertainty around the prediction?
4. What are the sources of uncertainty and bias in risk scores?

At this point it is important to emphasize that quantifying uncertainty should not and does not undermine the value in a sound application of risk adjustment. In fact, the research aims to strengthen the foundations of the concept, providing new tools for greater rigor in its application, and therefore enabling more success in meeting the policy goals of risk adjustment.

Question 1: When is a difference in risk scores statistically significant?

When calculating group average risk scores from a sample, we need to understand whether differences in risk scores are statistically significant enough to justify budget movements. We have seen questions like this coming up in the context of provider global risk payments and expect that it will be relevant in the risk adjustment program in health insurance exchanges as well.

Table 2 shows the minimum difference in risk scores required to be statistically significant. Please note that the minimum difference in risk scores will differ by the risk adjustment model and the dataset used to construct confidence intervals. We do not guarantee that the same results can be found beyond our study sample and the CMS-HCC model.

We can see that the minimum difference required decreases by group size. For instance, for groups of 5,000 lives, the minimum difference required is at 3 to 4 percent. In other words, if we have two groups of 5,000 members each, an observed difference in risk scores of 0.0273 or greater would be considered statistically significant at the 0.1 significance level. For groups of 50 lives, the difference in risk scores needs to be greater than 0.2811 in order to be significant at the 0.1 significance level. (This can be a relatively high threshold to cross in practice.)

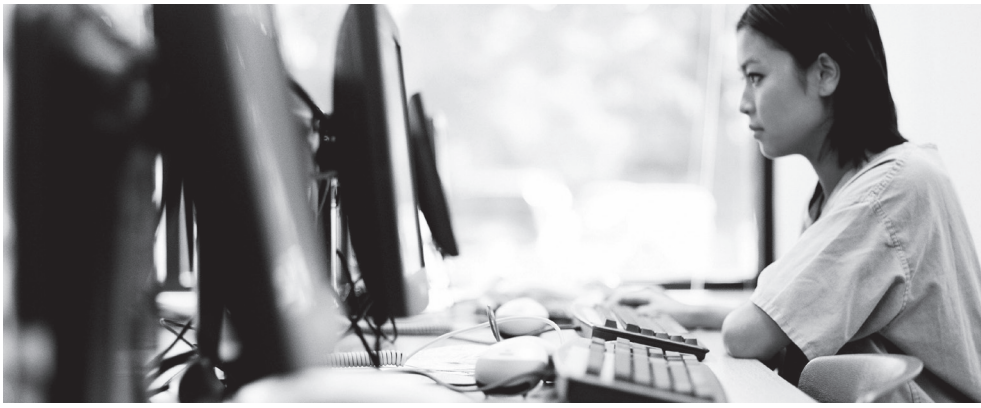
Table 2: Minimum Difference in Risk Scores Required for Statistical Significance

# of Members		Min. Diff. for Sig.	
Group 1	Group 2	90%	95%
50	50	0.2811	0.3370
250	250	0.1229	0.1467
1,000	1,000	0.0611	0.0728
5,000	5,000	0.0273	0.0326

Question 2: What is the size of prediction error at various group sizes?

In this study we provide a methodology to calculate confidence intervals empirically. This methodology applies to individuals and groups, and is independent of the risk assessment model that is used. Table 3 shows the empirical 90 percent confidence interval (90% CI) for groups of size one (i.e., individual) to 5,000 lives. The confidence intervals shown through this report are provided as *adjustments* to the risk scores (i.e., +/- adjustments). For example, if the individual risk score is 2.00, the 90% CI from Figure 2 is {2.00 – 1.6 = **0.4**, 2.00 + 2.81 = **4.81**}; or {0.4, 4.81}.

Danielle Bergmeier is an actuarial analyst with Wakely Consulting Group. She can be reached at *danielleb@wakely.com*.



CONTINUED ON **PAGE 12**

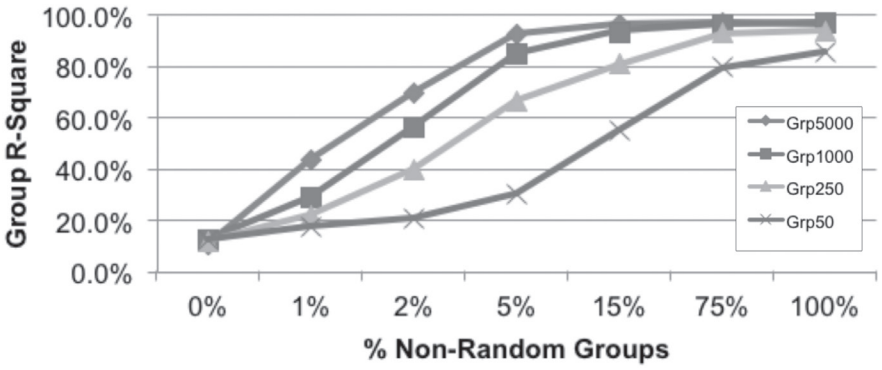
As shown in Table 3, the confidence intervals are asymmetric, owing to the asymmetric distributions of health care costs, and become narrower as group size increases.

Table 3: 90% Confidence Intervals by Group Size Medicare 5% Sample and CMS-HCC Model

Group Size	Confidence Interval
1	Score + {-1.6,2.81}*
2	Score + {-1.25,2.21}
5	Score + {-0.89,1.51}
25	Score + {-0.49,0.69}
50	Score + {-0.37,0.49}
250	Score + {-0.19,0.21}
1,000	Score + {-0.097,0.1}
5,000	Score + {-0.042,0.046}
10,000**	Score + {-0.033,0.038}
25,000**	Score + {-0.022,0.023}
50,000**	Score + {-0.016,0.016}
100,000**	Score + {-0.012,0.011}

* Refer to the report for derivation of these ranges.

Chart 1: R-Squared vs. Percentage of Groups Created Non-Randomly



Question 3: How do accuracy and uncertainty change by group size?

Group R-squared is a commonly used metric to evaluate group-level accuracy of risk adjustment models. A limitation of the data used in this study is that it is de-identified and does not contain actual grouping information of the individual. One approach would be to form groups of various sizes randomly and calculate the R-squared and confidence intervals by random groups, such as in earlier research by Ellis and Yi (Ellis & Yi, 2003). In reality, groups—either employer groups or provider groups—are not randomly formed. To draw more meaningful inferences, we took a hybrid approach by blending in a certain percentage of non-random groups in the bootstrapping simulations, where the non-random groups are determined by risk score bands. The results are illustrated in Chart 1. Overall, we find that only a small amount of non-randomness is needed in order for the group R-squared to increase significantly, and larger group sizes have a steeper increase than smaller group sizes. We expect that when real grouping is used, the group R-squared can be quite high. There is a lot of anecdotal evidence among practitioners for high R-squared for actual groups (vs. the low individual-level R-squared results that are most commonly cited). It is satisfying to experimentally confirm the relationship of the R-squared statistic to non-random grouping.

We find that calculated confidence intervals are quite resilient to changes in group R-squared, and in fact slightly expand when group R-squared is high. In summary, what is responsible for the increase in accuracy, i.e., the variation in actual cost, also leads to an increase in the spread of the error term, thus increasing the width of calculated confidence intervals.

The fact that a higher R-square does not decrease uncertainty in risk scores does not negate the need for higher accuracy in risk adjustment. Higher R-square results are widely desired; however, a disciplined methodology does not exist to run a cost-benefit-style analysis on an incremental increase in accuracy. The research report (Mehmud & Yi, 2012) makes the point that in order to value

accuracy, we need to look at risk scores through the lens of the rating actions or business decisions that can take place as a result of assessed risk.

Another key point discussed in the report is drawing inferences regarding accuracy of risk adjustment from application to a small population. Accuracy metrics such as the R-square *tend* to their "true" value as the sample size or population gets larger, and this value can bounce around significantly at smaller sizes. The report describes (section E.1) *convergence* of R-square as a concept, and determines that with the Medicare population and the CMS-HCC adjuster, convergence generally occurs for over 10,000 members. This concept is not at all new, and there is a good discussion on the perils of drawing inferences regarding loss ratios from data with low credibility in a recent *Health Watch* article (Wrobel, 2012).

Question 4: **What are the other sources of uncertainty in risk scores?**
In practice, uncertainty in risk scores does not accrue solely from the quality of the risk adjustment model being used. It may also come from the data fed into a model. For instance, the CMS-HCC model, as well as many other risk assessment models, was developed using a calendar year of fully run-out data to predict cost for the next year. In practice there are several constraints that do not allow an application of the model that is consistent with its development, and potentially this could lower accuracy and introduce bias in the predictions. Such practical constraints include:

- **Claim and administrative lag:** Health care claim lag is usually at least three months. Administrative lag includes time needed to aggregate, edit and validate the data and to run analytics.

- **Partial eligibility:** Members may either be newly eligible or partially eligible for a benefit year, and their diagnostic information is either missing or incomplete. There are techniques that commercial risk adjustment models have used to address prediction bias associated with partial eligibility. In Medicare Advantage, members with less than 12 months of eligibility would be scored by an age/gender model. In many state Medicaid managed care programs, we have seen the eligibility threshold to be six or seven months. However, the report concludes that the risk score at nine months is only 93 percent of its true value (i.e., value if full year of data were available), and the score at six months is 85 percent of its value. This is a pretty important result, indicating that the magnitude of bias due to differences in average eligibility may overwhelm any underlying differences in average morbidity.
- **Data quality:** Claim data varies in quality. Diagnoses codes may not be complete or reliable—leading to large differences in risk scores.

Risk adjustment is important not only from the perspective of issuers, but also has critical public and health policy implications. Any differences in risk scores driven by factors other than underlying morbidity undermine the policy goals of risk adjustment—and can potentially disrupt the market. We need to carefully understand the impact of such factors and make adjustments accordingly. The report focuses on quantifying the impact from several methodological and data constraints.

Risk adjustment is important not only from the perspective of issuers, but also has critical public and health policy implications.

The accuracy and confidence interval metrics for these issues are summarized in Table 4. While individual R-squared results vary significantly by input data quality, the confidence intervals are quite stable within a specified group size. The reason for the stability of confidence intervals with respect to data quality is the same as for accuracy, in that the interval width is driven by the variance in actual cost and not predicted cost.

Confidence intervals are related to the variance of the error term, which can be expressed as:

CONTINUED ON **PAGE 14**

Table 4: 90% Confidence Interval by Various Models and Group Size

Input Data	Individual R-Sq	Grp: 1	Grp: 5	Grp: 50	Grp: 5,000
1: Demo Only**	0.5%	{-1.09,3.23}	{-0.82,1.65}	{-0.39,0.53}	{-0.047,0.048}
2: Standard***	12.3%	{-1.6,2.81}	{-0.89,1.51}	{-0.37,0.49}	{-0.042,0.046}
3: Lag-1Q†	10.2%	{-1.62,2.86}	{-0.91,1.53}	{-0.38,0.51}	{-0.044,0.046}
4: Lag-2Q	8.9%	{-1.63,2.88}	{-0.92,1.53}	{-0.38,0.51}	{-0.046,0.046}
5: Elig-9Mo††	12.3%	{-1.56,2.82}	{-0.88,1.49}	{-0.37,0.51}	{-0.044,0.046}
6: Elig-6Mo	12.1%	{-1.52,2.83}	{-0.87,1.49}	{-0.38,0.49}	{-0.043,0.045}
7: Elig-3Mo	11.6%	{-1.46,2.87}	{-0.86,1.52}	{-0.37,0.51}	{-0.044,0.045}
8: Turnover-10%	11.0%	{-1.53,2.84}	{-0.88,1.51}	{-0.37,0.5}	{-0.044,0.046}
9: Turnover-30%	9.0%	{-1.37,2.95}	{-0.87,1.56}	{-0.38,0.51}	{-0.045,0.047}
10: Quality-Inp†††	12.3%	{-1.6,2.8}	{-0.89,1.5}	{-0.38,0.49}	{-0.043,0.046}
11: Quality-Out	12.3%	{-1.6,2.81}	{-0.89,1.49}	{-0.38,0.5}	{-0.044,0.045}
12: Quality-Prof	9.0%	{-1.43,2.92}	{-0.86,1.55}	{-0.38,0.5}	{-0.045,0.046}

** Demographic only prediction. *** Using a standard application of the CMS-HCC model.

† Incorporating one-quarter (1Q) or two-quarters (2Q) lag between the experience and prediction periods.

†† Consideration of partially eligible members (e.g., members eligible for nine months—Elig-9Mo).

††† Quantifying impact of data quality issues by systematically ignoring diagnoses (e.g., Quality-Inp implies that all diagnoses codes from an inpatient setting are ignored when calculating risk scores). Similarly, outpatient (-out) and professional (-prof) codes are ignored and accuracy/uncertainty results are recalculated).

We calculated the variance of the predicted and actual costs for the study population and found that the variance in actual cost is over six times the variance in predicted risk, and therefore is by far the dominant contributor to the total variance of the error term. This is the reason why even when the R-squared changes (i.e., prediction quality changes) the confidence intervals remain relatively unaffected. This can also be observed in Table 4, wherein the width of confidence intervals are dependent upon group size, and not as impacted by changes in the accuracy of various applications of a risk assessment model as measured by the R-squared statistic.

We have described thus far some of the main results of the study for the four questions above. There is however, one more key question:

How can a practitioner utilize the results generated from methods such as those used in this study?

A focus of the research was to produce a methodology that is of practical use and of interest to actuaries working in risk adjustment.

The detailed results presented in Appendix F of the report can be condensed into a simple lookup table that an actuary can use to determine the appropriate confidence interval to apply to a risk score. Confidence intervals vary by group size, lag, turnover, partial eligibility, risk score percentile and expected accuracy of the risk-score predictions (at individual or group level). Confidence intervals also will vary by the type of population/data and model used. The methods presented in the study, however, may be used to develop the appropriate set of results given other data or models.

Once these results have been developed, we can look up the corresponding confidence interval. For example, say we have a group of 995 Medicare FFS members with an average risk score of 1.02. The average eligibility of the group is 12 months, and there is a lag of three months between the experience period used to assess the risk and the period

during which the score is effective. From Appendix F, the nearest corresponding 90 percent confidence interval (i.e., group size of 1,000) for this situation is {1.02-0.099, 1.02+0.1} = {0.921, 1.12}. This means that while the best estimate for risk is 1.02, the actual risk for a plan may lie between 0.92 and 1.12 with a 90 percent confidence.

While the best estimate is still the one used in pricing and calculating adjustments, the quantification of uncertainty provides the practitioner key information regarding the expected variation of actuals from estimates. At least one risk assessment model (WRA model; for details, please see wramodel.com) has incorporated the calculation of confidence intervals into risk score estimates.

Conclusion

The authors would greatly appreciate feedback from and discussion among risk adjustment practitioners regarding recognition of uncertainty in risk assessment and risk adjustment. We hope that, through discussion and exchange of ideas, best practices would emerge that would enrich this area of actuarial expertise.

Works Cited

Ellis, R., R. Yi, et al. August 2003. Applying Diagnosis-Based Predictive Models to Group Underwriting. *Health Section News* No. 46. Retrieved May 15, 2012, from *http://cms-staging.soa.org/library/newsletters/health-section-news/2003/august/hsn-2003-iss46-elliskramer.aspx.*

Mehmud, M. and Rong Yi. 2012. *Uncertainty in Risk Adjustment*. Schaumburg, Ill.: Society of Actuaries. Web link: *http://www.soa.org/research/research-projects/health/uncertainty-risk-adjustment.aspx*

Winkelman, R. and S. Mehmud. 2007. *A Comparative Analysis of Claims-Based Tools for Health Risk Assessment.* Schaumburg, Ill.: Society of Actuaries. Web link: *http://www.soa.org/research/research-projects/health/hlth-risk-assement.aspx*

Wrobel, K. 2012. The Actuarial Profession and Complex Models: Knowing the Limits of Our Knowledge. *Health Watch,* Issue 68, 5–8.

We hope that, through discussion and exchange of ideas, best practices would emerge that would enrich this area of actuarial expertise.