IBM **Watson Health**™

Cost of care models performance evaluation summary

IBM **Watson Health**™

Contents

In early 2015, Truven Health Analytics® (aquired by IBM in 2016 and now part of the IBM® Watson Health™ business) released the Cost of Care Models (CCM), a suite of claims-based risk assessment models designed to predict healthcare spending on a per-capita basis. These predictive models estimate both retrospective and future expected healthcare payments for a commercially insured population. As part of the release, Truven Health contracted with an independent external party to measure the statistical performance of these models. This paper presents a summary of those results. The full report is available on request.

Model creation

The models were created using the Truven Health MarketScan™ Research Databases (now known as the IBM® MarketScan® Research Databases), and apply both linear and non-linear modeling methods to predict cost of care. Both concurrent and prospective models were developed for total cost of care (medical and prescription drugs) and medical costs alone: none, $100,000, and $250,000. For brevity, only results for the "none" and $250,000 truncation criteria are discussed in detail or presented in the tables herein.

Evaluation description

Truven Health contracted with an independent third party, David Knutson, who has an extensive background in this type of evaluation. Mr. Knutson was an investigator in the risk adjustment tools comparative study sponsored by the Society of Actuaries (SOA) and conducted in 2002 (Cumming and Knutson, 2002),[1] which was then updated in 2007 (Winkelman and Mehmud, 2007).[2] Those studies measured and compared the predictive performance of the prominent risk assessment tools within the marketplace at the time. Truven Health requested that Mr. Knutson conduct an evaluation of the overall predictive performance of the CCM and, where reasonable, compare results to the 2007 SOA study.

Data

The MarketScan Research Databases were used for the evaluation. Importantly, this data represented claims from a later time period than that used to estimate the models. The evaluation data represented about 25 million covered lives enrolled during 2011 and 2012. Model performance on several subpopulations was also evaluated. These groups were defined by age and gender, ranges of incurred costs and specific disease categories.

Methodology

Each individual was assigned risk scores by the models based on 2011 experience (age, sex, diagnosis history) for each of the 12 model variations (concurrent and prospective, three truncation options and total cost with and without pharmacy costs). Actual incurred cost for each individual was provided for both 2011 and 2012 for comparative purposes. These costs represented total allowed payments; that is, they included both plan liability and member payments.

The overall evaluations of each model included two commonly applied tests: R2 correlation analysis and Mean Absolute Prediction Error (MAPE). In addition, analyses were conducted for subpopulations based on age and sex groupings, incurred cost ranges and groups with specific diagnoses.

Results

Summary

Table 1 provides performance results for both R2 and MAPE for each of 8 models—concurrent and prospective, two truncation options and total cost with and without pharmacy costs, both for the total population and for a population limited to full-year enrollees. Here and throughout this document, full-year enrollees are defined as members that were born in or enrolled during all 12 months in 2011 for the concurrent models, and those who were also enrolled for all 12 months in 2012 for the prospective models.

Table 1. Overall predictive performance of all models

		Full-year enrollees	All enrollees	Full-year enrollees	All enrollees
Concurrent models					
Cost	Truncation	R2		MAPE	
Medical	No Truncation	0.529	0.423	68%	67%
	$250K Truncation	0.635	0.614	65%	66%
Total	No Truncation	0.538	0.412	63%	65%
	$250K Truncation	0.638	0.618	60%	61%
Prospective models					
Cost	Truncation	R2		MAPE	
Medical	No Truncation	0.182	0.172	107%	108%
	$250K Truncation	0.226	0.228	104%	105%
Total	No Truncation	0.206	0.189	94%	94%
	$250K Truncation	0.261	0.272	92%	92%

The differences in both R2 and MAPE results between all enrollees and full-year enrollees were relatively small. Where there was a difference, the difference was as expected, with the R2 slightly higher for full-year enrollees and MAPE slightly lower. The results between the two groups are highly correlated for all the models.

A high correlation across truncation-level results was found for both concurrent and prospective models. It is notable that the gain in predictive accuracy measured by correlation analysis between the $250,000 and the $100,000 truncation levels is small. It may suggest that the Truven Health CCM perform better at higher actual cost levels than is typical of such models. MAPE results follow a more typical pattern across truncation levels, with MAPEs decreasing at fairly similar intervals across levels.

The models that predict total cost of care were compared with models estimated to predict medical cost only. Analysis showed that the results are very highly correlated for both concurrent and prospective models.

The 2007 SOA evaluation used total annual cost and applied the same truncation levels as did this evaluation. It is important to note that while the methodologies are generally comparable and credible comparisons can be made, the studies were conducted separately and methodological specifications may differ. The most directly comparable results to the 2007 SOA study are those calculated for full-year enrollees. A comparison of these results to the comparable models in the SOA evaluation is provided in Table 2 below.

Table 2. Comparison to 2007 SOA study results

		SOA study	CCM	SOA study	CCM
Concurrent models					
Cost	Truncation	R2		MAPE	
Total	No Truncation	0.274 to 0.498	0.538	81% to 65%	63%
	$250K Truncation	0.297 to 0.518	0.638	81% to 65%	60%
Prospective models					
Cost	Truncation	R2		MAPE	
Total	No Truncation	0.162 to 0.174	0.206	96% to 88%	94%
	$250K Truncation	0.192 to 0.206	0.261	95% to 88%	92%

In summary, the Truven Health model results compare favorably with findings from the 2007 SOA evaluation. In nearly all comparisons, the Truven Health model R2 performance is better than or as good as the comparable models evaluated by the

SOA. MAPE results follow a consistent pattern across all models, with consistently lower MAPEs for full-year versus all enrollees and lower MAPEs as truncation thresholds are lowered. MAPEs are lower for concurrent compared with prospective models, as would be expected. The pattern is similar for the total cost and medical cost models. Total cost model MAPEs are consistently but only modestly lower than for medical cost models. MAPE results also follow a typical pattern across truncation levels, with MAPEs decreasing at fairly similar intervals across levels.

Subpopulation results

The most common and arguably most important statistic for insurance applications is evaluating predictive ratios by subgroups. Therefore, predictive ratios were calculated to determine how well the models perform for specific groups. The following tables present predictive ratios for total cost and medical cost models for age and sex, cost and diagnosis subgroups for selected models.

Because the difference between full-year and all enrollees was consistently very small and correlated at the 0.99 level, results are presented for only full-year enrollees to reduce information burden. For the same reason, only the total cost models are presented rather than both the total cost and medical cost models. Full results on all model permutations are available on request.

Age and sex groups

Table 3. Predictive ratios for age/sex groups –total cost models, full-year enrollees

Age/ sex groups	Concurrent – No truncation	Concurrent – Truncated at $250K	Prospective – No truncation	Prospective – Truncated at $250K
0-18 M	0.998	1.01	1.08	1.05
0-18 F	0.967	1.03	1.02	0.999
19-30 M	0.972	0.961	0.997	1.00
19-30 F	1.06	1.06	1.06	1.05
31-50 M	1.03	1.02	1.03	1.03
31-50 F	1.22	1.03	1.03	1.02
51-64 M	0.978	0.989	0.968	0.979
51-64 F	0.986	0.997	0.990	0.988

As would be expected, the range of predictive ratios is relatively small and all values are close to 1.00. These results show no major over-prediction or under-prediction by age group. The consistent under-prediction for the 51-64 age groups is common with similar models and, therefore, somewhat expected.

Cost groups

Table 4. Predictive ratios for groups by percentile of actual cost–total cost models, full-year enrollees

Cost	Concurrent – No truncation	Concurrent – Truncated at $250K	Prospective – No truncation	Prospective – Truncated at $250K
0-20	4.25	4.48	37.79	36.7
21-40	2.21	2.43	5.77	5.65
41-60	1.82	1.85	3.02	2.95
61-80	1.40	1.40	1.72	1.69
81-90	1.13	1.12	1.10	1.07
91-95	0.938	0.948	0.709	0.712
96-99	0.783	0.756	0.444	0.437
>99	0.606	0.582	0.247	0.249

Predictive ratios progress from low-cost to high-cost groups as expected and are generally closer to 1.00 for concurrent models than for prospective. The very high predictive ratios in the lowest percentile are typical, as are the low predictive ratios in the highest percentiles. Nearly all risk adjustment models under-predict high-cost individuals and over-predict low-cost individuals, particularly on a prospective basis.

The results from this evaluation are similar to results from the SOA evaluation for comparable risk assessment tools. In general, it appears that the Truven Health CCM may be better at predicting the tails of cost groups than the other tools in the marketplace.

Diagnosis groups

Computing predictive ratios for groups with specific diagnoses is also a common and important evaluation to identify any bias for individuals with certain conditions. Table 5 reports the predictive ratios for patients diagnosed with the eight conditions shown.

Table 5: Predictive ratios for diagnosis groups–total cost models

Predictive ratios by selected diagnosis groups	Concurrent – No truncation of total annual cost	Concurrent – Truncated at $250K of total annual cost	Prospective – No truncation of total annual cost	Prospective – Truncated at $250K of total annual cost
Diabetes	0.983	0.962	0.837	0.826
CHF	1.05	0.996	0.708	0.551
Asthma	1.00	1.00	0.803	0.793
Hypertension	1.06	1.04	0.843	0.834
CAD	0.965	0.954	0.564	0.574
COPD	0.983	0.987	0.644	0.655
Breast Cancer	1.05	1.05	0.672	0.659
Mental Illness	0.996	0.988	0.724	0.704

The few under-predictions and over-predictions by diagnosis group are relatively small and compare favorably with the performance of other commercial models. The pattern of prediction error and bias for specific condition categories is generally quite consistent. The prospective models, which are unable to adequately predict acute events occurring in the future year, consistently under-predict, as expected.

One interesting, small departure from a commonly found pattern in risk assessment model evaluations across truncation levels is that for the Truven Health CCM, lower thresholds of truncation increases or reduces bias only very slightly. This may be another manifestation of the models performing well for high-cost cases, which is beneficial not only for users targeting to lower the risk of these patients, but also for applications involving specialty providers, medical homes and ACOs first seeking to improve quality and build trust with their high-cost complex patients.

Key terms used in this paper

R-squared (R2) is a measure of how well a regression model fits a collection of data. It is constrained by 0 and 1; the closer R2 is to 1, the better the model fits the data. R2 is an estimate of the amount of variation in the data that is captured by the model. For example, an R2 of 0.86 means that 86 percent of the variation observed in the data is captured in the model.

The MAPE compares the model-predicted values against the observed values in the data. It estimates how far off the predictions are from the observed values in terms of percentage. For example, a MAPE of 5 indicates that the predictions deviate from the observations by an average of 5 percent. Put another way, the model forecast is off by 5 percent. The MAPE is constrained at the bottom by 0 with no maximum limit.

Get connected
Email: watsonh@us.ibm.com

About IBM Watson Health
Each day, professionals throughout the health ecosystem make powerful progress toward a healthier future. At IBM Watson Health, we help them remove obstacles, optimize efforts and reveal new insights to support the people they serve. Working across the landscape, from payers and providers to governments and life sciences, we bring together deep health expertise; proven innovation; and the power of artificial intelligence to enable our customers to uncover, connect and act — as they work to solve health challenges for people everywhere.

For more information on IBM Watson Health, visit:
ibm.com/watsonhealth

Footnotes

1 A Comparative Analysis of Claims-based Methods of Health Risk Assessment for Commercial Populations; By: Robert B. Cumming, FSA, MAAA*; David Knutson+; Brian A. Cameron, FSA, MAAA*; Brian Derrick*; A research study sponsored by the Society of Actuaries; May 24, 2002; *Milliman USA, Inc., Minneapolis; +Park Nicollet Institute Health Research Center: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.550.8912&rep=rep1&type=pdf

2 A Comparative Analysis of Claims-Based Tools for Health Risk Assessment; By: Ross Winkelman, FSA, Principal & Consulting Actuary, Milliman; Syed Mehmud, Actuarial Assistant, Milliman; Peer reviewed by: Leigh Wachenheim, FSA, Principal & Consulting Actuary, Milliman, April 20, 2007; https://www.soa.org/Files/Research/Projects/risk-assessmentc.pdf

HPW03066-USEN-01

IBM **Watson Health**™