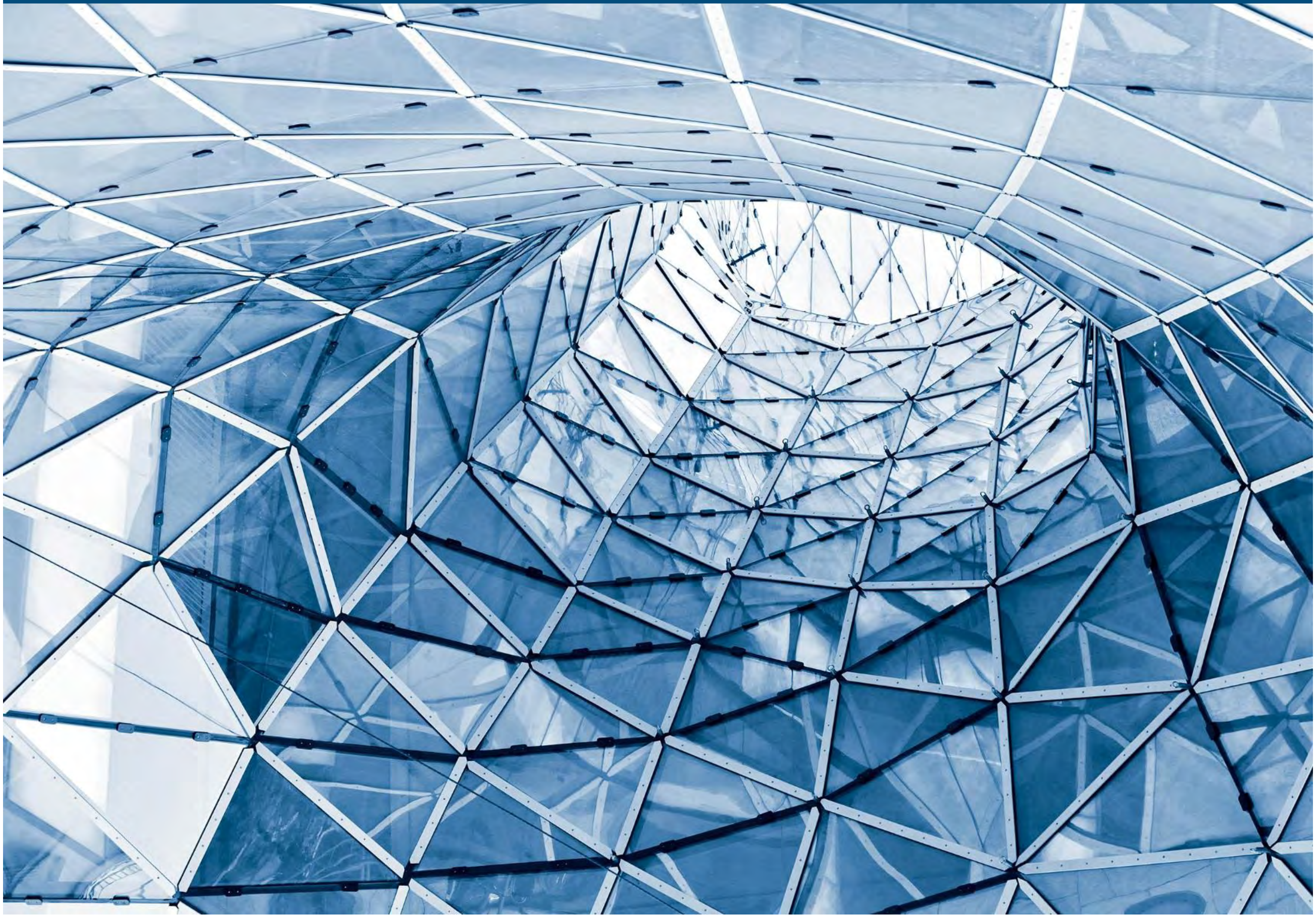
SOCIETY OF ACTUARIES

Accuracy of Claims-Based Risk Scoring Models



October 2016

Accuracy of Claims-Based Risk Scoring Models

SPONSOR	Health Section Research Committee	**AUTHORS**	Geof Hileman, FSA, MAAA Spenser Steele

Caveat and Disclaimer

The Society of Actuaries does not endorse, explicitly or implicitly for any purposes, software packages evaluated in the study. The opinions expressed and conclusions reached by the authors are their own and do not represent any official position or opinion of the Society of Actuaries or its members. The Society of Actuaries makes no representation or warranty to the accuracy of the information.

TABLE OF CONTENTS

Accuracy of Claims-Based Risk Scoring Models

This paper presents the results of a study comparing the accuracy of over 40 risk scoring models from 11 different vendors / sources. The study builds on prior studies with a similar objective published by the Society of Actuaries in 1996, 2002, and 2007. (Dunn, et al., 1996) (Cumming, Knutson, Cameron, & Derrick, 2002) (Winkelman & Mehmud, 2007)

Section 1: Acknowledgements

We are grateful to the Society of Actuaries and the Health Section Research Committee in particular for their funding of this research effort. We are also very appreciative of Steven Siegel and Barbara Scott of the Society of Actuaries for their administration of the project and the associated oversight group.

The Project Oversight Group provided valuable feedback and dialogue throughout the course of the project. The POG included the following individuals:

- (Chair) Ian Duncan, FSA, FIA, FCIA, FCA, MAAA
- Brandon Barber, MS
- Christopher Coulter, FSA, MAAA
- HsinTing Tina Liu, CPA, MBA
- Bill O'Brien, FSA, MAAA
- Rebecca Owen, FSA, MAAA
- Dan Pribe, FSA, MAAA

We also wish to recognize Casey Kangas of Kennell & Associates for his peer review of our SAS code and Randall Brown, Ph.D., of Mathematica Policy Research for consulting support and peer review throughout the project. We are also indebted to the previous research teams that laid a foundation for this study. The 1996 study was conducted by researchers from the Harvard University School of Public Health and Coopers & Lybrand, LLP, and included Daniel Dunn, Alice Rosenblatt, Deborah Taira, Eric Latimer, John Bertko, Thomas Stoiber, Peter Braun, and Susan Busch. The 2002 study was conducted by Robert Cumming and Brian Cameron of Milliman and David Knutson and Brian Derrick of the Park Nicollet Institute Health Research Center. The 2007 study was conducted by Ross Winkelman and Syed Mehmud of Milliman.

Finally, each of the vendors participating in the study worked extensively with us to ensure that their models were properly installed and functioning in our environment. Results of the study were shared with them prior to publication in order to provide an opportunity for comments. Each vendor also provided us with the use of their software on a no-cost basis, which was essential for the completion of this project. We are grateful for the collaborative and helpful spirit exhibited by all of the vendors throughout the research effort.

Section 2: Background and Scope

2.1 MOTIVATION FOR THE STUDY

Since the publication of the most recent SOA study comparing risk scoring models, the field of such models has become considerably more crowded. In the 2007 study, twelve models were compared from a total of six distinct vendors (Winkelman & Mehmud, 2007). Since then, commercially available models are now being marketed from additional vendors - we have included such newer offerings from Milliman, SCIO Health Analytics, Truven Health Analytics, and the Wakely Consulting Group. Also, some of the existing vendors have expanded their suite of available products. The Johns Hopkins ACG System is now available in a pharmacy-only, diagnosis-only, and combined version, where only two versions had been included in the 2007 study.

In addition to new commercially-available models, risk scoring has taken on a much more prominent role within the U.S. healthcare system through the implementation of the Affordable Care Act's reforms. Specifically, all individual and small group health insurance plans (both on- and off-exchange) are subject to financial transfers governed by a risk adjustment program. This program is designed to ensure that carriers attracting higher-risk individuals within each community-rated market are compensated by the plans attracting lower-risk individuals. Risk adjustment has essentially replaced medical underwriting as the primary mechanism for normalizing revenue for health risk in the individual and small group segments. While risk adjustment has been prominent in Medicare and Medicaid managed care financing for years, this new focus in the commercial marketplace supports a fresh look at the various models that can be used for this purpose. We have also included the model that governs most of these financial transfers (the HHS-HCC model) in our comparison study.

2.2 STUDY SCOPE AND COMPARISON TO PRIOR STUDIES

The primary objective of this research effort remains unchanged from prior studies – to evaluate the predictive accuracy of the current set of commercial risk scoring models available in the marketplace. Although some vendors have ventured into related predictive modeling efforts such as the risk of hospitalization, we have kept our focus on the traditional "risk score" value. The models that are included in this study are based on data elements found in claim encounter records, such as diagnoses, procedures and prescription drugs. These models do not include information from medical records such as clinical indicators of severity, measures of prior use, lifestyle or supplemental demographic information, or survey-based data.

While the number of risk scoring models and vendors has increased considerably from the prior study, we have reduced the number of variations in the comparison. For example, the prior study compared models with health care costs censored at two different levels; we have reduced this to one level. We have also not reproduced the prior study's evaluation of data with and without a lag in the data periods between the diagnosis and evaluation period. These topics were treated thoroughly in the previous study and their exclusion permitted time to be spent on new areas of focus.

We have introduced new comparisons among the models. First, we have developed a measure to indicate the likelihood that a model's predicted risk score is accurate at the individual level within a specified tolerance. We have also evaluated the binary prediction of the top one percent of healthcare spenders. Finally, we have also included a comparison of the accuracy of each model when predicting biased samples of groups of individuals. This is particularly relevant to the current commercial marketplace, where risk scoring methods are being called upon to serve as a substitute for underwriting.

We have attempted to provide a variety of methods for comparisons across models, so that model selection can be guided by metrics that match up more closely with the business problem at hand for a given application.

We have found that this variety of means of comparison is essential to understanding differences among models and the areas in which risk scoring models excel. One of the key points stressed throughout the paper is the observation that R-Squared values alone are not sufficient to explain the predictive abilities of a risk scoring model. We have found that R-Squared values are particularly susceptible to the influence of outlier observations and that other measures are needed to fully evaluate a set of models.

2.3 OVERVIEW OF TERMINOLOGY

Throughout this paper, we have referred to the category of models that we evaluated as *risk scoring models*. While perhaps more commonly referred to as "risk adjustment" models, we feel that the latter term connotes a specific purpose for which these models are typically used – that is, the normalization of plan premiums or claims experience across a covered population or market to account for differences in risk. Risk scoring is the first step of that process – the means by which each individual is assigned a specific risk score. The second step of risk adjustment, the payment transfer, is not within the scope of this study, as it is typically not coupled with a particular model.

In all of the model comparisons, we have categorized the models as prospective or concurrent models. This is a critical distinction for any approach to risk scoring. A *prospective* model uses information from one year to predict medical expenditures for the following year. A *concurrent* model, by contrast, uses information from one year to explain medical expenditures in that same year. For the purpose of this comparison of statistical power, the most important distinction between the two approaches is that concurrent models are far more accurate in their predictions. This is due to the fact that the claims experience being predicted is more closely associated with the data period from which the independent variables have been drawn. A full discussion of the differences between these two approaches and their application can be found in a recent report published by the Society of Actuaries (Hileman, Rosenberg, & Mehmud, 2016).

Another important difference between the tested models is the type of inputs, or independent variables, that are used in determining the predictions. We have classified models according to their use of diagnosis data, pharmacy data and, for prospective models, prior year cost data. While there may be other minor differences in the types of input data used (such as the inclusion of procedure codes as an input), we have categorized models into the following groups: diagnosis-only (DX), pharmacy-only (RX), diagnosis-and-pharmacy (DX+RX), and diagnosis-and-pharmacy with prior year costs (DX+RX+$).

Section 3: Study Design

In this section we provide a summary of the models that were selected for inclusion in the study, the data source used in the statistical evaluation, and the methods used for comparing the predictive power of the various models.

3.1 INCLUDED MODELS

Since the release of the 2007 study (Winkelman & Mehmud, 2007), there have been notable new entries into the risk scoring model market. Several existing vendors have also increased the variety of models offered. One new development over the past decade has been the introduction of predictive models that aim to predict more than simple relative risk. For example, some models now produce probabilities of hospitalization as an additional dependent variable. The evaluation of these predictive variables is beyond the scope of this study, which focused solely on the traditional risk score measure of relative risk.

We have included 23 prospective models and 19 concurrent models from a total of ten distinct sources / vendors in this study. For all comparisons, we have grouped the models by their prospective or concurrent design but also by the type of input data used in the risk score generation. These groupings included diagnosis-only (DX), pharmacy-only (RX), diagnosis-and-pharmacy (DX+RX), and diagnosis-and-pharmacy plus prior cost (DX+RX+$; prospective applications only). Table 3.1.1 summarizes the models that were included in the study.

Table 3.1.1: Included Models

Source	Model Name / Version	Model Types
Johns Hopkins University	Johns Hopkins ACG System, v11.0.1	Concurrent (DX; DX+RX) Prospective (DX; RX; DX+RX; DX+RX+$)
University of California at San Diego	Chronic Illness & Disability Payment System, v5.5 MedicaidRx, v5.5	Concurrent (DX; RX; DX+RX) Prospective (DX; RX; DX+RX)
3M Health Information Systems	3M Clinical Risk Groups (CRG), v2.0	Concurrent (DX+RX) Prospective (DX+RX)
Verisk Health	DxCG Intelligence, v4.3.1	Concurrent (DX; RX) Prospective (DX; RX; DX+RX+$)
Centers for Medicare and Medicaid Services	HHS-HCC Model, v3	Concurrent (DX)
Optum	ImpactPro	Prospective (DX; RX; DX+RX)
Milliman	Milliman Advanced Risk Adjusters (MARA), v3.6	Concurrent (DX; RX; DX+RX) Prospective (DX; RX; DX+RX; DX+RX+$)
SCIO Health Analytics	Prospective Cost of Care Model	Prospective (DX+RX+$)
Truven Health, an IBM Company	Cost of Care Model, v2.0	Concurrent (DX) Prospective (DX)
Wakely Consulting Group	Wakely Risk Assessment Model, v5.01	Concurrent (DX; RX; DX+RX) Prospective (DX; RX; DX+RX)

We excluded the Medicare Advantage CMS-HCC model, because it is specifically tailored for a Medicare population and would not be appropriate to compare alongside the models focused on commercial-aged populations, particularly given the inclusion of the HHS-HCC model now used in the commercial sector.

We have provided brief summaries of each of the models that were provided for this study. These summaries are drawn from the background materials and documentation provided by each of the vendors, where available.

3.1.1 ACG System (Vendor: Johns Hopkins University)

The ACG System concurrent and prospective cost models measure the morbidity burden of patient populations based on disease patterns derived from the diagnostic and/or pharmaceutical code information found in insurance claims or other electronic medical records. A distinguishing feature of the ACG System is its "person-focused" approach emphasizing the constellation of morbidities rather than individual disease categories or stages allowing the System to capture the multidimensional nature of an individual's health over time. The program offers a suite of risk models (e.g likelihood of hospitalization, unexpected high pharmacy use, continuous high utilizer) as well as a range of clinical markers (e.g. to coordination of care, active treatment for specific disease categories, frailty …) that provide additional context to the interpretation of generated risk scores.

3.1.2 Chronic Illness & Disability Payment System and MedicaidRx (Vendor: University of California at San Diego)

The Chronic Illness and Disability Payment System (CDPS) is a classification system for Medicaid programs to use to make health-based capitated payments for TANF and disabled Medicaid beneficiaries. There are both concurrent and prospective weights provided along with diagnosis and pharmacy weights for all flags the model creates. The provided weights are summed across the flags on a member level.

3.1.3 Clinical Risk Groups (Vendor: 3M Health Information Systems)

CRGs are a classification system for describing the health status and burden of illness of individuals in an identified population. CRG relates the historical clinical and demographic characteristics of the enrollee (claim based diagnosis, procedure, pharmaceutical, and functional health status) to the amount and type of healthcare resource that enrollee will consume in the future. In addition, CRGs can be linked to critical outcomes such as rates of potentially Preventable Readmissions and Emergency Department Visits.

The CRG system is a categorical clinical model that classifies each member of the population based on his or her burden of chronic medical conditions, assigning each individual into one of over 1,400 mutually exclusive risk categories. Individuals without a chronic condition are assigned to groups for healthy or significant acute illness. CRGs offer the user the choice of two models for both prospective and concurrent applications. The prospective model has 346 base categories and a total of 1,434 risk groups with severity level breakouts. These are also aggregated to three tiers with 618, 206, and 44 risk groups in each tier, respectively. The concurrent model is similar but with slightly more risk groups.

Although only the diagnosis-plus-pharmacy version of the CRG model weights were provided and tested, CRGs can also be run with diagnosis data only.

For this study, a pre-release working version of the V2.0 CRG software from August 2015 was used. The fully updated V2.0 CRG software released by 3M in May 2016 contained additional updates to its categories and logic, but was not available in time to be used for this study.

3.1.4 DxCG Intelligence (Vendor: Verisk Health)

Using predictive models, DxCG Intelligence turns healthcare data into risk scores for individual patients. Scores correlate with the cost of the underlying illness burden that individuals carry. Aggregating the scores of individuals with key attributes generates group-level predictive results that can be applied to answer questions fundamental to the ability to manage clinical and financial risks.

Consisting of more than 100 models, DxCG Intelligence includes both concurrent and prospective variants. Models are grouped into three primary functional bundles—budgeting and underwriting, medical management, and performance assessment—that can be tailored for commercial, Medicare, and Medicaid populations.

3.1.5 HHS-HCC (Centers for Medicare and Medicaid Services)

The HHS-HCC model was developed by CMS to fulfil the need for risk normalization in the post-ACA commercial marketplace. The HHS-HCC model uses diagnoses and demographics to assign a risk score to each individual. There are separate models provided for infants, children, and adults, each of which reflect the specific contribution of particular conditions to risk for these groups. One unique aspect of the HHS-HCC model is that the model does not predict allowed costs, but rather predicts plan liability at each of the five ACA metal levels: platinum, gold, silver, bronze, and catastrophic. Because in this study we are measuring accuracy in predicting total allowed costs, we have used the HHS-HCC platinum model as it represents the closest available proxy for allowed costs.

It is also very important to note in the context of this comparison study that the maximization of R-Squared and other measures of predictive accuracy was not a primary goal in the development of the HHS-HCC model. According to Kautter et al. (2004), “the HHS-HCC models are intended to balance high predictive ability with lower sensitivity to discretionary diagnostic coding.”

3.1.6 Impact Pro (Vendor: Optum)

Optum ImpactPro is a clinical, episode-based predictive model. It:

- Uses information readily available from medical and pharmacy claims, as well as member enrollment files.
- Uses a member's clinical episodes of care, prior use of health care services, prescription drugs, and lab results as markers of their future health care use.
- Creates markers of use that can be both predictive and provide clinical insights into why a patient is high risk.
- Predicts both future expenditures and calculates the probability of one or more hospitalizations.
- Produces outputs that can be used to design and implement effective care and case management strategies and to support actuaries and underwriters.
- Includes a reporting application that allows users to explore model results to better understand patients of highest risk and their most important diseases and conditions.

3.1.7 Milliman Advanced Risk Adjusters (Vendor: Milliman)

Milliman Advanced Risk Adjusters (MARA) release 3.6 was used for this study. MARA uses demographic and claim data in conjunction with its library of risk adjusters to estimate morbidity and healthcare resource use. One of the distinguishing features of the MARA model is the calculation of six service-specific risk scores in addition to the total risk score. Risk scores are calculated separately for pharmaceutical, inpatient facility, outpatient facility, emergency room, physician, and other medical services. MARA output includes clinical condition flags as well as risk scores. The MARA library of models includes prospective and concurrent models calibrated for commercial and Medicare populations, as well as an implementation of the HHS-HCC risk adjustment model specification published by the federal government.

3.1.8 Prospective Cost of Care Model (Vendor: SCIO Health Analytics®)

The SCIO® Prospective Cost of Care Model™ is a commercial risk assessment model developed by SCIO Health Analytics®. The model aims at predicting the total costs and financial risk per member using their health care utilization, prior year's total health expenditures, and demographics. The model was developed using two years of commercial claims data. In addition, the model leverages enrollment data, demographic details, medical claims, and pharmacy claims data. The intent is to better align the risk factors with more recent treatment patterns and heath care costs.

The Prospective Cost of Care Model assigns each member to one or more of the 75 SCIO proprietary condition categories and/or CCS diagnosis groupers based on medical and pharmacy claims. This helps generate a member risk profile that is based on age, gender, and condition categories. Member's prospective risk cost is then assigned based on age, gender, utilization, prior year costs, and condition categories.

3.1.9 Cost of Care Model (Vendor: Truven Health Analytics, an IBM Company)

Truven's Cost of Care Model estimates both retrospective and future expected healthcare payments for a commercially insured population. The models were created using the Truven MarketScan® research database, 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. The models predict relative costs under three alternative high-cost outlier truncation criteria: None, $100,000, and $250,000.

3.1.10 Wakely Risk Assessment Model (Vendor: Wakely Consulting Group)

The Wakely Risk Assessment (WRA) model was developed with the goal of keeping the model design simple and transparent. To this end, the WRA model includes fewer than 90 medical markers and less than 60 pharmacy markers in a simple linear-additive model. The required model inputs are also designed to use a minimum amount of information (to reduce administrative burden of running a model). Another motivation for the model was to anticipate what the HHS-HHS model may look like. Towards this, the model explicitly disallows a substantial number of diagnosis codes and pharmacy NDC codes that may be vague, discretionary, or otherwise susceptible to 'gaming'. The model includes an implementation of the HHS-HHS model as an option.

3.2 DATA SOURCE

The primary data source for this study was Truven Health Analytics' MarketScan Commercial Claims and Encounters database for calendar years 2012 and 2013. The MarketScan database contains experience covering nearly 50 million lives, including demographics, medical, and prescription drug encounter data. In order to keep the computational requirements manageable, we selected a sample of one million individuals for the calculation of measures of predictive accuracy. We sampled these individuals in two stages: first, we excluded individuals with inadequate data from the study (the exclusion criteria are identified below); second, we randomly sampled one million individuals from the remaining eligible pool of individuals.

We excluded from the study individuals who met any of the following three conditions:

- We excluded individuals for whom prescription drug data was not available in MarketScan. (Approximately 22 percent of the potential sample).
- We excluded any individual with at least one capitated service in either year of the historical data. The financial data associated with the capitated services are not necessarily consistent with the non-capitated services. (Approximately 6 percent of the potential sample).
- We excluded individuals with less than 12 months of enrollment in 2012. (Approximately 39 percent of the potential sample). There was no minimum enrollment period for the second year, 2013, provided the individual was represented in the database.

The three criteria combined, including any overlap, eliminated 53 percent of the MarketScan population. Once we applied these exclusion criteria, we selected a sample of one million individuals using a simple random sampling approach. We determined one million to be an adequate size for our analytic sample by calculating the mean risk score and R-Squared statistic for samples increasing in size from 100,000 to one million. These statistics were stable with a random sample of one million lives. Previous studies used considerably smaller samples: 620,000 in the 2007 study and 375,000 in the 2002 study.

The 2012 MarketScan data were used to generate the inputs for the prospective models, with target costs being drawn from the 2013 data. Testing of concurrent models was conducted using inputs and costs drawn from 2013.

It is important to stress the importance of viewing the results of this study within the context of the data source. MarketScan is a nationally representative data source with high quality data. However, the individuals covered by the MarketScan database may differ in demographics or health status from other populations to which risk scoring models may be applied. For instance, our exclusion of individuals with fewer than 12 months of enrollment in 2012 effectively removes newborns from the study. While this was consistent with prior research, it is not representative of a typical commercial population.

One disadvantage of the use of the MarketScan database is that it is used in the development and specification of several of the tested models. While we believe we have mitigated this risk as discussed in Section 3.4, a more ideal dataset would have been completely independent of all tested models. However, we were not able to identify a sufficiently rich dataset that met this criterion.

3.3 MEASURES OF FIT

As in each of the previous studies, we have computed three familiar series of statistics indicating measures of predictive accuracy: R-Squared, the Mean Absolute Error (MAE) statistics, and a series of predictive ratios. R-Squared and MAE are both indicators of individual goodness-of-fit, as they both describe the distribution of the error in predicting individual risk levels. Predictive ratios, by contrast, provide a snapshot of accuracy within a specific subgroup of individuals.

The coefficient of determination, denoted by R-Squared or R^2 for the remainder of this report, is a commonly used metric for describing the fit of a model that predicts a continuous variable. It is best described as the percentage of model variation in the dependent variable that is explained by the specified model. The R-Squared is mathematically equivalent to the square of Pearson's Correlation

Coefficient. For both measures, a value of 0 percent indicates a perfectly random distribution with no relation to the specified model while a value of 100 percent indicates a model that perfectly predicts the dependent variable for every observation in the sample. The general formula for R-Squared is given by:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where f_i is the prediction for observation i and y_i is the actual value for observation i. Note that for the fraction portion of the R-squared measure the numerator is the residual sum of squares and the denominator is the total sum of squares.

For each model, we have also calculated the Mean Absolute Error (MAE). Mean Absolute Error is used to measure how close a prediction is to the outcome. Established alternatives to the MAE are the mean absolute scaled error and the mean squared error. The MAE formula is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

where $|e_i| = |f_i - y_i|$ is the absolute error for observation i and n is the total number of observations. For MAE, lower values are preferable. An MAE of zero would indicate that the estimated risk score was always perfectly accurate (and thus the average error was zero). There is no theoretical upper limit on the MAE. Because all risk scores and actual costs in this study have been rescaled to a mean of 1.0, we have expressed the MAE as a percentage.

Although R-Squared has been featured in each of the prior SOA studies and continues to be included in this current study, it is a statistic that should be interpreted with great care. One well-known example, reproduced in (Leida & Siegel, 2014), shows four contrived datasets with identical correlation coefficients that exhibit quite different relationships. This illustration, known as Anscombe's Quartet, demonstrates that an R-Squared alone cannot accurately convey the relationship between two variables (Anscombe, 1973). In one prior SOA study (Cumming, Knutson, Cameron, & Derrick, 2002), the authors provide a demonstration of the effect that a single outlier observation can have on the squared prediction error, which is the basis for R-Squared. In Section 5.3, we provide a discussion of the development of an alternate model with a very low R-Squared but other metrics that indicate a potentially better fit than most of the included models.

By including a variety of alternative measures, our intention is to begin to draw attention away from the R-Squared measure as a definitive metric of risk scoring success. Some of these measures include analysis at the group level, rather than the individual.

Predictive ratios are defined as the mean risk score divided by the mean actual cost for a subgroup of individuals from the sample population, with both values scaled to 1.0 over the entire population. Thus, a predictive ratio of 100 percent indicates that a model is – on average – perfectly unbiased for that group of individuals. A predictive ratio above or below 100 percent suggests a possible bias. Predictive ratios are an important counter-balance to the individual-level metrics of R-Squared and MAE that we have previously discussed. We have calculated predictive ratios by the presence of specific medical conditions, by age/sex group, by cost range, by level of benefit richness, and by geographic region.

In addition to these three measures, we have measured models on other bases as well: the accuracy of each model in identifying the individuals in the top one percent of health care expenditures and a measure that quantifies the degree to which each model is accurate at an individual level within a specified allowable error limit.

3.4 MODEL RECALIBRATION

In each of the prior SOA-sponsored comparison studies, risk scoring models have been compared on two bases: using the coefficients provided by the software vendors, and using recalibrated coefficients. The vendor-supplied coefficients are tailored specifically to the population on which the model was built, while recalibrated coefficients may be used to more closely reflect the relationships between healthcare expenditures and the model's independent variables for a population similar to that which is the basis for the comparison study. In the initial design of our approach, we had not intended to perform any recalibration of the models. This was out of recognition of two observations. First, most end-users of risk scoring models are using them with weights as supplied by the vendors, thus this offered-weight comparison is the most relevant to typical actuarial practice. Second, we observed that with the notable exception of CDPS, there was very little movement in the relative accuracy of the models tested in the 2007 study between the offered-weight and the recalibrated comparison. The CDPS exception was expected due to its specific focus on the Medicaid population.

Ultimately, we did determine that some manner of recalibration of the tested models was appropriate. One of the main considerations was our understanding that some of the tested models are developed using MarketScan data and some are not. This can affect both the overall structure of a model and its supplied weights. While we have mitigated this risk by using only a sample of the 2012 and 2013 MarketScan files, there is at the very least the appearance of potential favorable bias for those models that are already aligned with MarketScan experience. Our decision to also test the models on a recalibrated basis was further supported by Actuarial Standard of Practice 45, which states that "recalibration is often used to make the risk adjustment (or risk scoring) model more specific to the population, data, and other characteristics of the project for which it is being used." (Actuarial Standards Board, 2012)

We considered several approaches to the recalibration of the various risk scoring models. First, we considered conducting full recalibrations of each model. In this approach, we would attempt to recreate the regression model that was used in the specification of each vendor's model. In order to successfully use this approach, we would need full visibility into the independent variables that comprise each model, including any hierarchies or combinations. This information was not available for all included models. The second approach we considered was the technique used in the previous study, which was a credibility-weighted recalibration. This approach relies on a regression of any known independent variables against the model residuals, in order to create an adjustment to each individual's risk score. (Winkelman & Mehmud, 2007) Finally, we evaluated the use of a ridge regression approach. (Parkes, 2015)

The results of our evaluation of these three approaches was discussed in detail in a recent article in the SOA *Predictive Analytics and Futurism* newsletter (Hileman G. , 2015). In short, we determined that with an adequate sample size, the three methods converge to the same risk scores. We chose to use the same method as was employed in the 2007 study, the credibility-weighted recalibration, because it represented the most straightforward implementation and required the least amount of transparency into the inner workings of the various models.

In order to recalibrate each model, we specified a linear regression equation of the form:

$$Y_{Actual} - Y_{Predicted} = \sum_{i=1}^{A} \alpha_i \times \ Age/Gender\ Bin_i + \sum_{i=1}^{B} \beta_i \times \ Condition\ Bin_i$$

Where Y_{Actual} is the actual allowed healthcare expenditures relative to the average, $Y_{Predicted}$ is the risk score relative to the average, $\alpha_i$is the regression coefficient that specifies adjustments to the demographic component of the risk score, and β_i is the regression coefficient that specifies adjustments to the condition- and pharmacy-based components of the risk score.

To recalibrate the models, we used any independent variable (or condition category) that was specifically provided within the model's output. For some models, we were able to confirm that this was the complete set of independent variables used in the risk score calculation. For others, this set of predictor variables may not have been all-inclusive. While we used all of the input data required for the operation of each model, in some cases the models convert raw data directly into risk scores without any reporting of the intermediate condition categories. Thus, we were not able to observe all independent variables. Additionally, combinations of variables or logical hierarchies may have also been used in the model but were not apparent to the end user.

Our approach avoided the need for full visibility into each algorithm we tested. To recalibrate the models, we regressed the known independent variables against the model residuals. To the degree that some variables or combinations were not visible to us, these variables' contributions to the risk score remained in the score.

We multiplied each specified coefficient by a factor of $(1-p)^{5.95}$, where *p* represents the *p*-value associated with the specific coefficient[1]. Through this adjustment, we were able to minimize the influence of highly insignificant results (with very high *p*-values). This approach is also consistent with the 2007 study. We then calculated an adjustment to the risk score for each individual and each model by computing the cross-product of the adjusted coefficients with the binary indicator variables for each demographic, condition-based, and pharmacy-based indicator at the person-level. These recalibrated risk scores were then compared to the actual costs in the same manner as the original out-of-the-box results.

There was one exception to our recalibration process. 3M's CRG model is structurally different than the other models in that its categories are a series of mutually exclusive conditions. As such, the development of model weights for the CRG system is simpler than for the regression-based approaches. With 3M-specified adjustments for credibility, we calculated the average relative cost for each of the mutually exclusive categories and replaced the original weight with these recalibrated weights.

Because of time constraints due to the models being provided later in the process, we were not able to perform the recalibration for the ImpactPro models. Because the effects of the recalibration process were minimal and our focus has remained on the comparison of the models with offered weights, we do not believe that this omission will have a material effect on the utility of this study.

[1] The choice of 5.95 as the exponent for this weighting was informed solely by its use in the 2007 SOA study. In discussions with the authors of that study, this factor was selected after performing a series of statistical tests. In practical application, the specific value of this exponent made very little difference.

Section 4: Results

4.1 CHARACTERISTICS OF SAMPLED POPULATION

Table 4.1.1 presents a comparison of the demographic distribution of the study population to the sample population used in the 2007 study. It compares both samples to the age-gender distribution of a reference population derived from the *Milliman Health Cost Guidelines,* 2006 edition, as reported in the prior SOA study. (Winkelman & Mehmud, 2007) We observe that the current analytic sample, drawn from the 2012 and 2013 MarketScan databases, conforms more closely to the reference population than the sample used in the prior study, which was drawn from the 2003 and 2004 MarketScan databases.

Table 4.1.1: Demographic Characteristics of Study Population Compared to Reference Population

Demographic Category	% of Total		
	Current Study	2007 Study	Reference
Child, 0-1	1%	1%	3%
Child, 2-6	6%	4%	7%
Child, 7-18	18%	16%	21%
Child, 19-22	6%	5%	5%
Male, 23-25	1%	0%	2%
Male, 25-29	3%	1%	3%
Male, 30-34	3%	2%	4%
Male, 35-39	3%	2%	5%
Male, 40-44	4%	3%	5%
Male, 45-49	4%	5%	5%
Male, 50-54	5%	7%	4%
Male, 55-59	5%	8%	2%
Male, 60-64	4%	5%	1%
Female, 23-25	1%	0%	2%
Female, 25-29	3%	1%	3%
Female, 30-34	4%	2%	4%
Female, 35-39	4%	2%	5%
Female, 40-44	4%	4%	5%
Female, 45-49	5%	6%	5%
Female, 50-54	5%	8%	4%
Female, 55-59	5%	10%	3%
Female, 60-64	5%	7%	2%

Table 4.1.2 shows the number of individuals by disease cohort during 2013.

Table 4.1.2: Sampled Individuals by Disease Category

Condition Category	Unique Members (out of 1 million)
Heart Disease	49,644
Mental Illness	130,936
Diabetes	76,105
Low Back Pain	88,776
Asthma	13,920
Arthritis	6,660

4.2 R-SQUARED AND MEAN ABSOLUTE ERROR

In each of the previous SOA studies, models have been compared using R-Squared and MAE statistics. While these measures, like any, have some drawbacks, they are still useful to assess the degree to which each risk scoring model successfully explains individual-level health expenditure risk. In this section, we present the R-Squared and MAE for each tested model using both uncensored costs and also using allowed costs censored at $250,000 for an individual to limit the influence of extreme outliers. The models are grouped by the type of input data used in producing the risk scores.

Table 4.2.1: R-Squared and MAE, Concurrent Models

	R-Squared		MAE	
	Uncensored	Censored at $250k	Uncensored	Censored at $250k
Diagnosis-Only Models				
ACG System	44.1%	52.4%	75.3%	73.3%
CDPS	24.2%	30.0%	92.5%	90.6%
DxCG	52.6%	61.0%	67.6%	65.0%
HHS-HCC	41.3%	45.2%	86.8%	85.5%
MARA	52.7%	62.6%	64.0%	61.8%
Truven	52.6%	62.7%	64.9%	61.6%
Wakely	43.2%	51.0%	76.5%	74.3%
Pharmacy-Only Models				
DxCG	29.6%	38.4%	83.0%	80.8%
MARA	30.1%	40.1%	81.8%	79.6%
MedicaidRx	12.9%	18.0%	100.3%	98.3%
Wakely	19.9%	28.8%	91.4%	89.2%
Diagnosis-and-Pharmacy Models				
ACG System	45.9%	56.4%	70.0%	67.6%
CDPS-MRx	25.6%	32.4%	90.0%	88.1%
CRG	41.0%	49.3%	78.2%	76.2%
MARA	55.4%	66.7%	57.9%	55.6%
Wakely	44.3%	54.2%	73.8%	71.3%

This comparison highlights the extraordinary influence that extreme values can have on the results, with significant jumps in R-Squared resulting from censoring at $250,000, a level reached by less than a tenth of a percent of the individuals in our testing sample. This comparison also shows the much stronger predictive power resulting from diagnostic inputs rather than using only pharmacy inputs. There is only a modest gain in predictive power when comparing the diagnosis-and-pharmacy models to the diagnosis-only models, suggesting a considerable amount of overlap in explanatory power between the diagnosis and pharmacy inputs.

Table 4.2.2: R-Squared and MAE, Prospective Models

	R-Squared		MAE	
	Uncensored	Censored at $250k	Uncensored	Censored at $250k
Diagnosis-Only Models				
ACG System	16.2%	21.0%	100.7%	98.7%
CDPS	9.1%	11.9%	109.2%	107.5%
DxCG	18.6%	23.8%	98.9%	96.9%
Impact Pro	18.9%	22.8%	98.2%	96.2%
MARA	20.1%	24.9%	97.3%	95.3%
Truven	20.7%	26.4%	96.4%	94.0%
Wakely	17.0%	21.3%	100.5%	98.6%
Pharmacy-Only Models				
ACG System	11.6%	16.5%	102.7%	100.7%
DxCG	14.8%	19.9%	100.4%	98.4%
Impact Pro	13.7%	19.1%	101.6%	99.6%
MARA	15.1%	20.1%	99.8%	97.8%
MedicaidRx	8.6%	12.8%	107.6%	105.7%
Wakely	9.9%	14.9%	103.9%	101.9%
Diagnosis-and-Pharmacy Models				
ACG System	17.2%	23.0%	97.6%	95.5%
CDPS+MRX	10.0%	13.3%	107.0%	105.1%
CRG	17.0%	21.7%	99.6%	97.6%
Impact Pro	20.7%	25.8%	94.6%	92.5%
MARA	22.0%	27.7%	93.3%	91.3%
Wakely	18.5%	23.7%	97.1%	95.1%
Prior Cost Models				
ACG System	17.8%	23.7%	96.7%	94.6%
DxCG	23.8%	27.7%	91.2%	89.1%
MARA	24.8%	26.9%	91.8%	90.1%
SCIO	15.1%	22.4%	95.8%	93.5%

The results in Table 4.2.1 and Table 4.2.2 were calculated using the out-of-the-box models – that is, using the weights as provided by each of the software vendors. For most of the models, we recalibrated the offered weights to be consistent with the MarketScan population, as discussed in Section 3.5. Table 4.2.3 shows the R-Squared and MAE for both the offered weights and the recalibrated concurrent models, without censoring, while Table 4.2.4 shows the same information for the prospective models.

With infrequent exceptions, the recalibration process had very little impact on the goodness of fit statistics. Because the offered weights are based on a Medicaid population which differs considerably from a commercial population, the CDPS model did exhibit substantial improvement from the recalibration, but most models' changes were extremely minimal. Somewhat more improvement was demonstrated in the 2007 study, but the improvements in the representativeness of the underlying MarketScan data since the prior study (which used 2002 and 2003 MarketScan data) may have diminished the importance of the recalibration process.

Table 4.2.3: R-Squared and MAE, Recalibrated Concurrent Models, No Censoring

	R-Squared		MAE	
	Offered	Recalibrated	Offered	Recalibrated
Diagnosis-Only Models				
ACG System	44.1%	44.0%	75.3%	74.0%
CDPS	24.2%	34.1%	92.5%	84.3%
DxCG	52.6%	53.2%	67.6%	67.9%
HHS-HCC	41.3%	42.2%	86.8%	86.4%
MARA	52.7%	51.9%	64.0%	67.6%
Truven	52.6%	53.0%	64.9%	65.6%
Wakely	43.2%	43.2%	76.5%	77.4%
Pharmacy-Only Models				
DxCG	29.6%	30.5%	83.0%	82.6%
MARA	30.1%	30.3%	81.8%	83.2%
MedicaidRx	12.9%	17.4%	100.3%	96.2%
Wakely	19.9%	20.0%	91.4%	92.7%
Diagnosis-and-Pharmacy Models				
ACG System	45.9%	45.7%	70.0%	69.7%
CDPS-MRx	25.6%	35.8%	90.0%	82.0%
CRG	41.0%	42.2%	78.2%	74.4%
MARA	55.4%	54.6%	57.9%	62.2%
Wakely	44.3%	45.5%	73.8%	74.0%

Table 4.2.4: R-Squared and MAE, Recalibrated Prospective Models, No Censoring

	R-Squared		MAE	
	Offered	Recalibrated	Offered	Recalibrated
Diagnosis-Only Models				
ACG System	16.2%	16.0%	100.7%	100.7%
CDPS	9.1%	14.7%	109.2%	104.0%
DxCG	18.6%	18.9%	98.9%	98.4%
MARA	20.1%	18.9%	97.3%	98.4%
Truven	20.7%	20.5%	96.4%	96.7%
Wakely	17.0%	17.0%	100.5%	100.1%
Pharmacy-Only Models				
ACG System	11.6%	11.2%	102.7%	105.1%
DxCG	14.8%	15.2%	100.4%	99.6%
MARA	15.1%	15.1%	99.8%	99.5%
MedicaidRx	8.6%	9.6%	107.6%	105.2%
Wakely	9.9%	9.5%	103.9%	104.0%
Diagnosis-and-Pharmacy Models				
ACG System	17.2%	17.1%	97.6%	97.6%
CDPS+MRx	10.0%	15.8%	107.0%	101.5%
CRG	17.0%	16.8%	99.6%	99.1%
MARA	22.0%	20.7%	93.3%	94.6%
Wakely	18.5%	18.2%	97.1%	96.9%
Prior Cost Models				
ACG System	17.8%	17.8%	96.7%	96.8%
DxCG	23.8%	24.7%	91.2%	91.6%
MARA	24.8%	24.1%	91.8%	93.4%
SCIO	15.1%	15.2%	95.8%	96.2%

4.3 GROUP-LEVEL MEASURES OF FIT

While the R-Squared and MAE values presented above do provide some insight into the relative performance of the various models, they are notably removed from the business problem that risk scoring models are most typically employed to help solve: predicting healthcare expenditures for a *group* of insured individuals. We believe more useful measures for comparing models in this context are thus at the group level.

In this section, we present R-Squared and MAE values for randomly selected groups of both 1,000 and 10,000 individuals from our analytic sample. For each simulated group, we calculated the mean actual cost as well as the mean value for each of the evaluated risk models. We then computed the R-Squared and MAE across 1,000 sampled groups. In addition to these point statistics, we have also displayed the 95th percentile of the absolute error from across the simulated groups. This gives an indication of the level of error at which we can be 95 percent confident a group's risk score will be within, given the specified group size. We selected 1,000 groups to provide a reasonable sample size for this analysis. These results are shown in Table 4.3.1 and Table 4.3.2.

A key observation from this analysis is that the error levels are much more tightly clustered across models for these randomly selected groups than at the individual level. Especially with the prospective models, the range from the worst performing models to the best performing models is quite small. The implication of this is that the differences in predictive power that we have measured at the individual level become less meaningful as the results are aggregated across groups of individuals. Also, despite still-different R-Squared statistics, the error rates for the concurrent and prospective models are also similar at the group level. For groups of 10,000, the mean absolute error among concurrent diagnosis-only models ranges from 2.1 percent to 2.7 percent, while the error for those models' prospective counterparts ranges from 2.8 percent to 3.1 percent.

A 2012 SOA study found that the mean absolute error for groups of 1,000 using the prospective CMS-HCC model and measured against Medicare experience was 4.85 percent (we found a range from 6.9 percent to 8.2 percent). This research found that the error rate dropped to 2.16 percent for groups of 5,000. (Mehmud & Yi, 2012) The error rates we are finding are somewhat higher, which may suggest that commercial populations are somewhat less predictable than Medicare populations with a more stable base of long-term chronic disease burden.

The 95th percentile measurements do demonstrate that significant uncertainty can remain, even at the large group level. Across all concurrent models, we observe an error range of 16.1 percent to 21.8 percent for groups of 1,000 and a range of 5.2 percent to 7.2 percent for groups of 10,000. For prospective models, these values grow slightly to a range of 21.5 percent to 23.0 percent for groups of 1,000 and a range of 6.9 percent to 7.7 percent for groups of 10,000.

Table 4.3.1: R-Squared and MAE, Simulated Random Groups, Concurrent Models (Uncensored)

	R-Squared		MAE		95th Percentile of Error	
	1,000	10,000	1,000	10,000	1,000	10,000
Diagnosis-Only Models						
ACG System	43.7%	48.7%	7.3%	2.3%	17.2%	5.6%
CDPS	26.9%	30.5%	8.6%	2.7%	21.0%	6.5%
DxCG	49.4%	56.2%	6.9%	2.1%	16.2%	5.4%
HHS-HCC	40.2%	46.0%	7.7%	2.3%	18.8%	5.9%
MARA	51.0%	57.8%	6.8%	2.1%	16.0%	5.4%
Truven	49.6%	55.7%	6.8%	2.1%	16.8%	5.1%
Wakely	41.8%	48.5%	7.5%	2.3%	17.7%	6.0%
Pharmacy-Only Models						
DxCG	28.6%	33.3%	8.3%	2.6%	20.7%	6.4%
MARA	28.5%	32.8%	8.3%	2.6%	19.5%	6.5%
MedicaidRx	14.4%	15.4%	9.3%	3.0%	22.0%	7.1%
Wakely	19.2%	22.1%	9.0%	2.8%	21.5%	7.0%
Diagnosis-and-Pharmacy Models						
ACG System	46.4%	51.9%	7.1%	2.2%	17.1%	5.6%
CDPS-MRx	27.8%	32.3%	9.5%	2.7%	21.1%	6.5%
CRG	40.5%	42.8%	7.5%	2.4%	18.2%	5.8%
MARA	53.9%	60.0%	6.6%	2.0%	15.5%	5.2%
Wakely	43.0%	49.1%	7.6%	2.4%	17.4%	6.0%

Table 4.3.2: R-Squared and MAE, Simulated Random Groups, Prospective Models (Uncensored)

	R-Squared		MAE		95th Percentile of Error	
	1,000	10,000	1,000	10,000	1,000	10,000
Diagnosis-Only Models						
ACG System	12.1%	16.0%	9.4%	2.9%	22.5%	7.2%
CDPS	7.8%	9.6%	9.2%	3.1%	22.5%	7.3%
DxCG	14.5%	18.8%	9.2%	2.9%	21.9%	7.2%
Impact Pro	13.1%	18.1%	9.2%	2.9%	22.2%	6.9%
MARA	15.2%	19.8%	9.2%	2.9%	21.7%	7.1%
Truven	16.6%	20.7%	9.1%	2.8%	21.3%	7.3%
Wakely	13.8%	17.4%	9.2%	2.9%	21.8%	7.2%
Pharmacy-Only Models						
ACG System	10.6%	14.0%	9.4%	3.0%	22.2%	7.1%
DxCG	13.8%	14.4%	9.2%	3.0%	22.0%	7.2%
Impact Pro	13.8%	13.8%	9.2%	3.0%	21.9%	7.3%
MARA	14.0%	15.3%	9.2%	2.9%	22.1%	7.1%
MedicaidRx	8.8%	8.4%	9.5%	3.1%	22.6%	7.3%
Wakely	10.3%	9.7%	9.4%	3.0%	22.9%	7.5%
Diagnosis-and-Pharmacy Models						
ACG System	13.9%	18.5%	9.4%	2.9%	22.3%	7.1%
CDPS+MRX	9.7%	10.7%	9.5%	3.1%	22.6%	7.2%
CRG	11.8%	12.1%	9.3%	3.0%	22.2%	7.4%
Impact Pro	15.8%	20.6%	9.1%	2.9%	22.0%	6.8%
MARA	18.5%	21.6%	8.9%	2.8%	21.7%	7.0%
Wakely	16.0%	18.5%	9.1%	2.9%	21.4%	7.1%
Prior Cost Models						
ACG System	14.5%	20.0%	9.2%	2.9%	21.9%	6.9%
DxCG	22.1%	24.3%	8.8%	2.8%	23.0%	7.1%
MARA	22.4%	24.9%	8.7%	2.8%	21.9%	6.8%
SCIO	14.3%	15.8%	9.2%	2.9%	22.1%	7.2%

4.4 PREDICTIVE RATIOS

One of the more useful measures of predictive fit is the predictive ratio, defined here as the mean risk score for a group of individuals divided by the mean actual scaled cost for that same group   Predictive ratios closest to 100 percent indicate a very good fit for a particular subgroup. A predictive ratio in excess of 100 percent indicates that a model overestimates the risk level for that group, while a predictive ratio below 100 percent indicates that the model underestimates the risk level. One paper suggests that a predictive ratio within plus or minus 10 percent of 100 percent indicates a reasonable degree of accuracy for a subgroup (Kautter, Ingber, Pope, & Freeman, 2012). Of course, any threshold should necessarily vary based on the size and composition of the subgroup and also based on the potential consequences of bias in a particular application.

We have calculated the predictive ratios for each of the included models for groups defined by several independent variables: the presence of specific health conditions, age and gender, actual expenditure range, a measure of plan benefit richness, and geographic area. For the main body of the report, we have included predictive ratios for each of the models using offered weights and with censoring at $250,000. The predictive ratios for the uncensored models are shown in Appendices I.A and 0. We have chosen to include the censored versions of the model for the main body of the report because the influence of outliers could be more acute when isolating the results to specific subgroups of individuals.

4.4.1 Predictive Ratios by Health Condition

We calculated predictive ratios for persons with six specific health conditions: heart disease, mental illness, diabetes, low back pain, asthma, and arthritis. Each condition was defined as at least one instance (in any position) in the prediction year of one or more ICD-9 diagnosis codes as shown in Table 4.4.1. We used conditions indicated on any type of claim for the classification of individuals for this exercise. We included all costs for the identified individuals, not just those associated with the condition.

Table 4.4.1: Diagnoses included in Specific Health Conditions

Condition	Included ICD-9 Codes
Heart Disease	390-398, 402, 404-429
Mental Illness	290-298.9, 300-312.9
Diabetes	250.1, 250.10, 250.11, 250.12, 250.13, 648.0, 648.00, 648.01, 648.02, 648.03, 648.04, 648.8, 648.80,648.81, 648.82, 648.83, 648.84, 250.0, 250.00, 250.01, 250.02, 250.03, 250.2, 250.20, 250.21, 250.22,250.23, 250.3, 250.30, 250.31, 250.32, 250.33, 250.40, 250.41, 250.42, 250.43, 250.5, 250.50, 250.51,250.52, 250.53, 250.60, 250.61, 250.62, 250.63, 250.70, 250.71, 250.72, 250.73, 250.8,250.80, 250.81, 250.82, 250.83, 250.9, 250.90, 250.91, 250.92, 250.93, 362.0, 362.0, 362.01, 362.02,362.1, 775.1, 790.2, 790.21, 790.22, 790.29, 253.5
Low Back Pain	724-724.9
Asthma	493-493.9
Rheumatoid Arthritis and Other Inflammatory Polyarthropathies	714-714.9

Table 4.4.2 shows the predictive ratios by medical condition for the concurrent models with offered weights and censoring at $250,000. We have also included the unweighted average predictive ratio

across the included conditions to permit for a more general comparison. Several key observations emerge from this table. First, there is a clear advantage provided by the commercially-marketed models compared to the HHS-HCC and CDPS-based models in terms of accurately estimating costs associated with individuals with these specific medical conditions. Second, the predictive ratios for these specific health conditions are much closer to 1.0 for the diagnosis-only models than for those that include only pharmacy inputs. No firm conclusion can be drawn about whether the addition of pharmacy data to the diagnosis models provides a better fit for these conditions. When examining the MARA results, all six conditions' ratios move closer to 1.0 when moving from diagnosis-only to diagnosis-and-pharmacy. The ACG System and Wakely results are less clear. Third, it seems that of the selected conditions, the costs of persons with low back pain, asthma, and arthritis are more likely to be underpredicted than heart disease and diabetes.

Table 4.4.2: Predictive Ratios by Health Conditions (Concurrent; Offered Weights; $250,000 Censoring)

	Heart Disease	Mental Illness	Diabetes	Low Back Pain	Asthma	Arthritis	Average
Diagnosis-Only Models							
ACG System	110.4%	101.0%	105.6%	98.7%	96.6%	97.2%	101.6%
CDPS	71.7%	86.7%	79.5%	69.4%	97.4%	67.1%	78.6%
DxCG	101.9%	98.2%	100.1%	98.0%	90.0%	86.4%	95.8%
HHS-HCC	101.7%	85.7%	104.4%	74.5%	95.6%	98.3%	93.4%
MARA	94.4%	97.1%	95.3%	95.2%	92.6%	88.6%	93.8%
Truven	101.6%	101.4%	98.7%	102.9%	93.6%	95.0%	98.9%
Wakely	98.6%	98.3%	100.0%	94.7%	90.6%	90.5%	95.4%
Pharmacy-Only Models							
DxCG	83.1%	86.0%	93.9%	87.2%	80.0%	88.6%	86.5%
MARA	77.1%	86.5%	91.2%	86.1%	85.0%	89.1%	85.8%
MedicaidRx	49.9%	79.0%	72.4%	64.7%	79.7%	66.5%	68.7%
Wakely	70.5%	88.4%	92.7%	86.4%	79.4%	86.2%	83.9%
Diagnosis-and-Pharmacy Models							
ACG System	101.2%	98.8%	98.5%	95.8%	92.9%	88.9%	96.0%
CDPS-MRx	75.3%	93.4%	86.2%	72.2%	94.9%	73.0%	82.5%
CRG	92.8%	95.5%	98.1%	86.0%	93.6%	98.3%	94.0%
MARA	94.8%	97.5%	96.1%	95.7%	94.9%	90.9%	95.0%
Wakely	90.6%	95.3%	98.1%	92.7%	87.1%	89.6%	92.2%

Table 4.4.3 also shows predictive ratios for persons with specific conditions, but for *prospective* models with offered weights and $250,000 censoring. The most notable difference between these ratios and those for the concurrent models is, of course, that they are much lower. This is because the conditions were indicated in year 2 (i.e. – the prediction year). As a consequence, it is expected that the prospective models will underestimate costs for these individuals, since some of these individuals may not have had any indication of the condition in the year 1 diagnosis and pharmacy data. As with the concurrent models, CDPS results are not as good as the commercial risk scoring models. Comparing across

conditions, diabetes has the highest predictive ratio, indicating that the associated costs are most readily predicted from prior year claims data (consistent with a chronic condition that may result in recurring costs). Heart disease has the lowest predictive ratios, suggesting that related costs may be more sporadic in nature.

Table 4.4.3: Predictive Ratios by Health Conditions (Prospective; Offered Weights; $250,000 Censoring)

	Heart Disease	Mental Illness	Diabetes	Low Back Pain	Asthma	Arthritis	Average
Diagnosis-Only Models							
ACG System	62.4%	73.6%	86.9%	72.8%	71.5%	75.9%	73.8%
CDPS	43.6%	66.3%	64.9%	55.2%	71.2%	44.4%	57.6%
DxCG	65.5%	75.6%	87.8%	76.1%	72.2%	72.1%	74.9%
Impact Pro	61.4%	73.9%	84.3%	73.6%	76.1%	71.5%	73.5%
MARA	63.3%	75.7%	85.5%	76.0%	74.5%	74.3%	75.1%
Truven	63.7%	77.5%	86.9%	77.3%	77.3%	80.5%	77.2%
Wakely	62.4%	74.7%	87.0%	73.2%	70.8%	66.4%	72.4%
Pharmacy-Only Models							
ACG System	56.3%	72.9%	82.1%	69.5%	70.3%	75.2%	71.1%
DxCG	59.3%	73.1%	84.7%	70.5%	68.9%	80.0%	80.8%
Impact Pro	56.9%	73.5%	83.7%	70.6%	70.9%	72.4%	71.4%
MARA	58.8%	74.2%	83.5%	71.1%	72.2%	81.2%	73.5%
MedicaidRx	43.8%	65.8%	73.8%	60.8%	62.7%	56.2%	60.6%
Wakely	54.7%	73.2%	83.4%	69.5%	70.0%	69.7%	70.1%
Diagnosis-and-Pharmacy Models							
ACG System	63.5%	77.4%	87.3%	75.8%	74.9%	77.1%	76.0%
CDPS+MRX	46.9%	72.6%	72.5%	57.6%	70.9%	50.6%	61.9%
CRG	63.0%	74.0%	88.1%	68.4%	71.5%	80.4%	74.2%
Impact Pro	63.4%	78.5%	86.8%	76.3%	77.2%	75.4%	76.3%
MARA	64.8%	77.9%	86.9%	77.5%	77.5%	82.0%	78.2%
Wakely	65.0%	77.5%	88.9%	75.3%	75.2%	76.0%	76.3%
Prior Cost Models							
ACG System	64.9%	77.0%	88.9%	75.5%	75.6%	82.1%	77.3%
DxCG	71.6%	79.6%	90.2%	78.8%	77.8%	86.8%	80.8%
MARA	69.7%	79.9%	90.0%	79.2%	81.0%	84.3%	80.7%
SCIO	59.1%	80.6%	89.1%	81.2%	89.0%	83.8%	80.5%

4.4.2 Predictive Ratios by Age and Sex

We calculated predictive ratios for six age-sex groups for both concurrent and prospective models with offered weights and $250,000 censoring in Table 4.4.4 and Table 4.4.5. The other variants are presented in Appendix I, along with additional age-sex groups.

The predictive ratios for children show considerable variation across vendors. The CDPS-based models overestimate the risk associated with children, likely because these models are developed and specified using a Medicaid population, in which eligible children have a much higher rate of costly medical conditions than a commercial population such as is included in the MarketScan data. This effect is also present with the CRG scores, where there has been a historical focus on pediatric and Medicaid populations.

The HHS-HCC predictive ratios for children are below 1.0. The HHS-HCC model is calibrated to predict plan liability, not total allowed cost. Since children typically have lower healthcare expenditures, the effect of deductibles and copays is more significant and the model thus *appears* to be underpredicting risk for children. Although we have used the version of the model designed for predicting platinum plan liability to minimize this issue, the predictive ratios less than 1.0 are an expected result. We note that MARA (overestimation) and both ImpactPro and DxCG (underestimation) produce results similarly deviant from 1.0 for children. For the 0-6 age group, the exclusion of most newborns from the study population may contribute to some observed bias.

Among adults, strong patterns of systematic bias are not evident. Adult women under age 45, predictive ratios are consistently less than 1.0. While not universally true, this suggests some degree of underestimation for this subgroup of the population. The pattern is less strong among the models relying exclusively on pharmacy data, so there may be a source of healthcare expenditures for this population that is most readily captured by pharmacy data, possibly prescription contraceptives.

Table 4.4.4: Predictive Ratios by Age-Sex, Concurrent Models (Offered Weights; $250,000 Censoring)

	Children, 0-6	Children, Age 7-18	Males, Age 19-44	Males, Age 45-64	Females, Age 19-44	Females, Age 45-64
Diagnosis-Only Models						
ACG System	104.5%	90.0%	97.9%	102.9%	93.9%	104.2%
CDPS	219.1%	216.9%	120.5%	80.7%	88.6%	76.5%
DxCG	88.0%	89.4%	100.4%	103.8%	94.9%	103.6%
HHS-HCC	88.1%	88.8%	100.1%	105.8%	94.6%	102.4%
MARA	115.9%	104.5%	100.8%	98.7%	97.3%	100.0%
Truven	94.0%	91.9%	98.5%	99.8%	101.0%	102.6%
Wakely	108.7%	100.4%	101.0%	100.0%	98.2%	100.0%
Pharmacy-Only Models						
DxCG	92.2%	91.6%	102.0%	104.4%	99.3%	99.1%
MARA	105.5%	99.7%	103.8%	99.8%	100.2%	98.5%
MedicaidRx	229.9%	220.3%	123.5%	79.5%	94.2%	71.4%
Wakely	97.7%	101.9%	106.0%	98.8%	101.0%	98.2%
Diagnosis-and-Pharmacy Models						
ACG System	108.6%	99.3%	101.3%	98.9%	99.7%	100.1%
CDPS-MRx	202.2%	209.1%	119.7%	82.4%	89.6%	78.2%
CRG	122.2%	109.1%	95.7%	104.4%	91.4%	98.9%
MARA	113.9%	103.9%	101.6%	98.6%	98.7%	99.3%
Wakely	103.0%	98.8%	104.2%	97.8%	99.6%	100.8%

Table 4.4.5: Predictive Ratios by Age-Sex, Prospective Models (Offered Weights; $250,000 Censoring)

	Children, 0-6	Children, Age 7-18	Males, Age 19-44	Males, Age 45-64	Females, Age 19-44	Females, Age 45-64
Diagnosis-Only Models						
ACG System	99.8%	98.7%	108.7%	100.6%	88.7%	104.0%
CDPS	212.4%	226.0%	119.0%	76.3%	93.3%	75.9%
DxCG	87.1%	88.7%	99.8%	104.3%	96.4%	102.7%
Impact Pro	116.0%	107.9%	102.1%	99.5%	101.6%	95.4%
MARA	105.9%	96.7%	103.7%	102.7%	96.7%	99.1%
Truven	96.3%	99.2%	101.0%	100.4%	99.5%	100.2%
Wakely	102.0%	100.3%	103.9%	100.4%	99.1%	98.7%
Pharmacy-Only Models						
ACG System	106.8%	100.6%	111.0%	98.5%	91.0%	102.6%
DxCG	86.1%	90.3%	103.1%	104.6%	100.3%	98.8%
Impact Pro	111.2%	92.9%	99.0%	100.9%	100.0%	100.5%
MARA	101.1%	95.4%	106.8%	101.0%	101.2%	97.4%
MedicaidRx	147.0%	134.0%	116.3%	91.6%	100.5%	88.8%
Wakely	94.2%	101.0%	105.8%	100.6%	99.4%	98.3%
Diagnosis-and-Pharmacy Models						
ACG System	99.7%	97.3%	108.3%	99.2%	93.3%	102.8%
CDPS+MRX	199.5%	219.0%	117.5%	78.2%	93.5%	77.5%
CRG	139.1%	113.2%	73.9%	99.0%	112.4%	94.7%
Impact Pro	100.7%	99.9%	101.8%	100.1%	102.4%	97.8%
MARA	106.2%	98.2%	104.1%	102.2%	98.2%	98.1%
Wakely	98.8%	100.1%	103.9%	100.9%	99.1%	98.7%
Prior Cost Models						
ACG System	101.3%	97.9%	106.7%	100.3%	92.1%	102.8%
DxCG	95.2%	88.4%	98.4%	103.7%	100.7%	100.5%
MARA	106.0%	97.3%	104.5%	102.5%	98.3%	97.8%
SCIO	122.5%	105.1%	99.0%	87.6%	108.2%	102.1%

4.4.3 Predictive Ratios by Cost Range

We have calculated predictive ratios for subgroups of individuals classified by their costs in the target year. This is a useful calculation to help illustrate the tendency of risk scoring models to regress toward the mean. For example, individuals with the lowest healthcare expenditures tend to be vastly overpredicted; among the commercial prospective diagnosis-only models, the bottom two deciles of cost were overpredicted by a factor of between 8.9 and 10.0. By contrast, the individuals with the very highest expenditures are, as a group, greatly underpredicted. The same set of models estimated the risk of the top two percent of spenders at 23.7 to 31.0 percent of their actual cost. This effect is considerably more pronounced for the prospective model than for the concurrent models. For the main body of the report, Table 4.4.6 and Table 4.4.7 present the concurrent and prospective models with offered weights and no censoring. The other model variants' results are shown in Appendix I.

The values at the lower end of the distribution should be interpreted with considerable caution. The 25th percentile of annual healthcare expenditures in the MarketScan database is just over $200, so the individuals in the lowest quintile have both very low actual costs and very low risk scores, thus greatly increasing the volatility of the predictive ratio.

In order to facilitate comparison, we have calculated ratio of the predictive ratio for the 40th-60th percentile range to the predictive ratio for the 95th-98th percentile range. A ratio closer to 1.0 indicates a smaller degree of bias by cost range. This measure helps highlight the models that are more effective at limiting bias by cost range.

Table 4.4.6: Predictive Ratios by 2012 Cost Percentile, Concurrent Models (Offered Weights; No Censoring)

	0-20th Percentile	20-40th Percentile	40-60th Percentile	60-80th Percentile	80-90th Percentile	90-95th Percentile	95-98th Percentile	98th-99th Percentile	40-60 / 95-98
Diagnosis-Only Models									
ACG System	1427%	268%	197%	154%	124%	104%	84%	59%	2.3
CDPS	6827%	635%	333%	191%	116%	79%	54%	31%	6.1
DxCG	1489%	351%	239%	165%	118%	91%	73%	59%	3.3
HHS-HCC	7405%	526%	259%	148%	99%	79%	66%	57%	3.9
MARA	1438%	350%	232%	159%	116%	94%	78%	60%	3.0
Truven	751%	253%	200%	151%	117%	100%	83%	66%	2.4
Wakely	868%	359%	267%	184%	127%	92%	69%	49%	3.8
Pharmacy-Only Models									
DxCG	5459%	487%	265%	169%	121%	91%	72%	41%	3.7
MARA	4906%	467%	266%	174%	124%	92%	71%	40%	3.8
MedicaidRx	8205%	777%	373%	198%	114%	73%	50%	20%	7.5
Wakely	3277%	479%	297%	195%	134%	93%	66%	30%	4.5
Diagnosis-and-Pharmacy Models									
ACG System	778%	245%	189%	153%	130%	112%	88%	55%	2.1
CDPS-MRx	5301%	573%	321%	193%	121%	82%	59%	32%	5.5
CRG	2155%	504%	267%	163%	107%	77%	64%	58%	4.2
MARA	624%	290%	209%	153%	119%	99%	84%	63%	2.5
Wakely	1168%	383%	277%	190%	130%	93%	68%	43%	4.1

Table 4.4.7: Predictive Ratios by 2013 Cost Percentile, Prospective Models (Offered Weights; No Censoring)

	0-20th Percentile	20-40th Percentile	40-60th Percentile	60-80th Percentile	80-90th Percentile	90-95th Percentile	95-98th Percentile	98th-99th Percentile	40-60 / 95-98
Diagnosis-Only Models									
ACG System	9897%	755%	358%	193%	112%	70%	48%	24%	7.5
CDPS	13985%	990%	405%	191%	98%	58%	36%	16%	11.2
DxCG	9396%	713%	348%	193%	114%	72%	49%	26%	7.2
Impact Pro	9264%	715%	337%	180%	105%	67%	46%	31%	7.4
MARA	10013%	736%	346%	186%	110%	71%	50%	27%	6.9
Truven	8912%	670%	324%	184%	113%	75%	54%	31%	6.1
Wakely	9322%	748%	363%	195%	112%	69%	46%	25%	7.8
Pharmacy-Only Models									
ACG System	10703%	770%	359%	194%	115%	72%	48%	19%	7.5
DxCG	10773%	751%	352%	190%	112%	70%	50%	23%	7.1
Impact Pro	10451%	768%	361%	195%	113%	70%	47%	21%	7.7
MARA	10212%	738%	349%	191%	113%	72%	49%	24%	7.1
MedicaidRx	13176%	956%	408%	199%	105%	60%	37%	12%	11.1
Wakely	9965%	763%	369%	202%	117%	71%	46%	18%	8.1
Diagnosis-and-Pharmacy Models									
ACG System	8417%	680%	338%	192%	118%	77%	53%	26%	6.4
CDPS+MRX	12375%	920%	393%	195%	104%	62%	39%	17%	10.0
CRG	8083%	745%	352%	184%	106%	67%	46%	28%	7.7
Impact Pro	8089%	659%	326%	182%	109%	71%	50%	31%	6.5
MARA	8962%	671%	324%	183%	114%	76%	55%	30%	5.9
Wakely	8105%	674%	341%	194%	117%	74%	51%	27%	6.7
Prior Cost Models									
ACG System	8328%	672%	335%	191%	118%	77%	54%	26%	6.2
DxCG	7920%	599%	299%	174%	111%	76%	58%	40%	5.1
MARA	8626%	645%	311%	177%	112%	76%	57%	35%	5.5
SCIO	6268%	596%	326%	199%	127%	84%	58%	24%	5.6

4.4.4 Predictive Ratios by Benefit Richness

Because MarketScan contains experience from a wide range of benefit designs, we have calculated predictive ratios for three levels of benefit richness. While the specific benefit design is not specified in the Commercial Claims and Encounters database, a unique plan identifier is provided for some individuals. We tabulated the total paid amounts and total allowed amounts for each unique plan key and calculated each plan's paid-to-allowed ratio. Table 4.4.8 shows the distribution of months of observed experience by plan-level paid-to-allowed ratio. Finally, we grouped the ranges of paid-to-allowed ratios into three broad categories of benefit richness.

Table 4.4.8: Plan-Level Paid-to-Allowed Distribution

Paid-to-Allowed	Months	Percentage	Category
95%+	256,539	1%	HIGH
90-95%	1,783,478	7%	
85-90%	5,162,027	21%	
80-85%	7,329,528	30%	MEDIUM
75-80%	1,908,176	8%	
70-75%	6,572,266	26%	LOW
<70%	1,828,579	7%	

Table 4.4.9 shows the predictive ratios by paid-to-allowed range for the concurrent models with offered weights and $250,000 censoring, while Table 4.4.10 shows the same data for the prospective models. We had expected that perhaps the richer benefit designs would result in lower predictive ratios, since the less restrictive cost sharing would in theory result in higher utilization and costs for the same diagnostic profile. We were surprised to see the opposite result, with the plans with richer benefits resulting in the highest predictive ratios.

Table 4.4.9: Predictive Ratios by Benefit Richness, Concurrent Models (Offered Weights; $250,000 Censoring)

	Low	Medium	High
Diagnosis-Only Models			
ACG System	97.2%	99.0%	110.9%
CDPS	104.0%	96.1%	108.8%
DxCG	96.9%	97.0%	108.0%
HHS-HCC	103.5%	99.0%	103.4%
MARA	98.9%	99.2%	108.9%
Truven	96.0%	97.3%	105.7%
Wakely	96.0%	97.6%	108.9%
Pharmacy-Only Models			
DxCG	108.7%	94.7%	109.7%
MARA	109.7%	95.8%	110.0%
MedicaidRx	115.2%	96.1%	110.3%
Wakely	110.5%	96.4%	113.4%
Diagnosis-and-Pharmacy Models			
ACG System	99.7%	95.8%	112.6%
CDPS-MRx	104.3%	96.3%	110.0%
CRG	103.3%	99.4%	106.9%
MARA	101.2%	98.5%	109.9%
Wakely	100.2%	97.3%	110.8%

Table 4.4.10: Predictive Ratios by Benefit Richness, Prospective Models (Offered Weights; $250,000 Censoring)

	Low	Medium	High
Diagnosis-Only Models			
ACG System	98.2%	97.2%	98.3%
CDPS	108.5%	94.7%	100.8%
DxCG	97.1%	96.4%	97.1%
Impact Pro	103.2%	96.4%	101.9%
MARA	101.2%	97.8%	100.3%
Truven	100.9%	98.5%	99.7%
Wakely	97.2%	95.5%	96.6%
Pharmacy-Only Models			
ACG System	110.3%	98.6%	111.7%
DxCG	110.0%	97.5%	108.7%
Impact Pro	110.7%	98.6%	109.5%
MARA	110.5%	97.9%	109.0%
MedicaidRx	115.6%	96.8%	108.1%
Wakely	111.2%	97.6%	111.7%
Diagnosis-and-Pharmacy Models			
ACG System	101.1%	96.7%	107.7%
CDPS+MRX	108.3%	95.6%	103.5%
CRG	105.7%	97.6%	103.5%
Impact Pro	104.2%	97.9%	105.2%
MARA	104.5%	98.4%	106.0%
Wakely	101.0%	96.7%	103.4%
Prior Cost Models			
ACG System	101.6%	98.2%	105.0%
DxCG	95.1%	97.6%	101.5%
MARA	103.2%	98.5%	105.0%
SCIO	98.4%	101.7%	113.7%

Because these results seemed counterintuitive (that the richer plan designs would be more likely to be overpredicted), we looked more closely at the groups that underlie these data. Less than ten percent of the sampled individuals included a plan linkage in the MarketScan data, so these results are based only on that smaller sample. We determined that the paid-to-allowed groups are dominated by several large groups. 38 percent of the experience for the high benefit richness category comes from one group with an observed 88 percent paid-to-allowed ratio and 59 percent of the experience in the low benefit richness category comes from two groups with paid-to-allowed ratios just under the 75 percent cutoff for that category. The predictive ratio for a single employer group can be influenced by factors other than the level of benefit richness of plan benefit. In order to limit the effects of these larger groups, we calculated the predictive ratio for each specific group and plotted these values against each group's paid-to-allowed ratio. These results are shown in Figure 1 and indicate a slight negative correlation between predictive ratio and benefit richness. Each point in this figure represents a unique benefit plan design. While not an authoritative study, this does suggest that richer benefit plans may have a tendency toward being underpredicted due to the higher utilization associated with the richer benefit packages.[2]

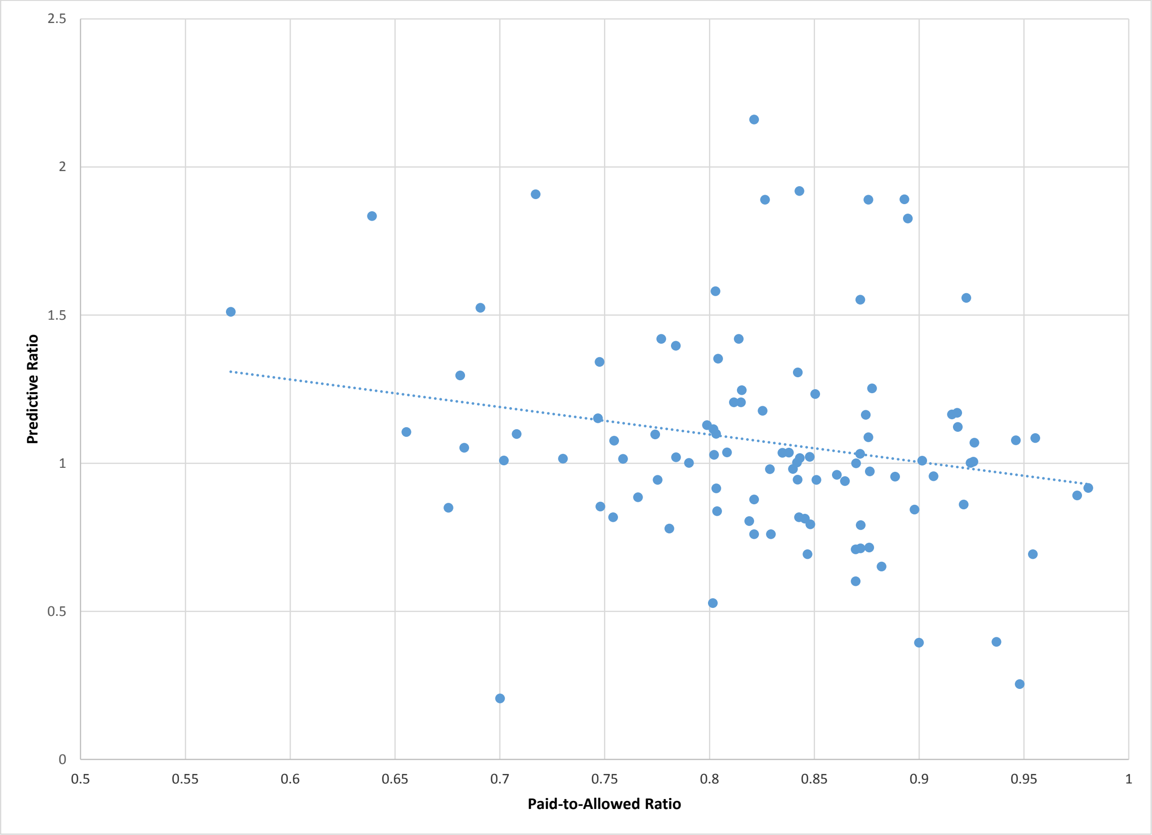


Figure 1: Predictive Ratios vs Paid-to-Allowed Ratio

4.4.5 Predictive Ratios by Geographic Area

We also calculated the predictive ratios within various regions of the United States. MarketScan provides several levels of geographic delineation, including region, state, and metropolitan statistical area. We aggregated at the region level to ensure statistical stability of the calculated values. The four regions are defined in MarketScan as follows:

- Northeast: Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont
- North Central: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin
- South: Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, Washington DC, West Virginia
- West: Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington

[2] The data points in Figure 1 were generated using the MARA Concurrent CX model.

These results are summarized in Tables 4.4.10 and 4.4.11.

The more interesting variation in this analysis is the differences across the regions rather than the differences across the various models. Generally speaking, we see a very similar pattern across the risk scoring models. However, the inclusion of pharmacy inputs seems to exert upward pressure on the predictive ratios in the south and downward pressure on the predictive ratios in the northeast and west. This means that the risk scores are generally overstated for the associated risk in the south and understated in the northeast and west. Differences in care patterns or overall medical cost levels by region could contribute to these differences. Based on this analysis, diagnosis-only models seem to be the most effective at minimizing regional bias.

Table 4.4.11: Predictive Ratios by Geographic Region, Concurrent Models (Offered Weights; $250,000 Censoring)

	Northeast	North Central	South	West
Diagnosis-Only Models				
ACG System	98.8%	101.2%	99.9%	97.4%
CDPS	100.8%	100.2%	98.8%	99.2%
DxCG	98.6%	101.3%	100.7%	96.2%
HHS-HCC	98.3%	101.3%	99.7%	99.1%
MARA	98.1%	101.1%	99.9%	98.6%
Truven	100.0%	101.3%	99.8%	96.5%
Wakely	103.6%	98.9%	99.6%	95.7%
Pharmacy-Only Models				
DxCG	90.3%	98.1%	107.3%	97.2%
MARA	89.2%	99.0%	107.7%	96.5%
MedicaidRx	88.7%	101.5%	105.3%	98.9%
Wakely	86.0%	98.3%	111.1%	94.1%
Diagnosis-and-Pharmacy Models				
ACG System	92.6%	100.9%	104.5%	96.3%
CDPS-MRx	98.9%	100.8%	100.3%	97.3%
CRG	95.7%	102.3%	101.5%	97.2%
MARA	95.6%	100.6%	102.4%	97.3%
Wakely	98.1%	98.4%	103.6%	94.7%

Table 4.4.12: Predictive Ratios by Geographic Region, Prospective Models (Offered Weights; $250,000 Censoring)

	Northeast	North Central	South	West
Diagnosis-Only Models				
ACG System	98.5%	99.7%	99.1%	102.2%
CDPS	97.7%	100.9%	98.2%	103.8%
DxCG	99.3%	99.8%	99.1%	101.3%
Impact Pro	97.6%	99.6%	100.1%	101.7%
MARA	97.8%	99.8%	99.7%	101.9%
Truven	99.8%	100.6%	98.5%	101.3%
Wakely	101.1%	99.1%	98.7%	100.8%
Pharmacy-Only Models				
ACG System	91.5%	98.7%	106.4%	99.4%
DxCG	91.8%	99.3%	105.7%	99.8%
Impact Pro	90.4%	99.6%	106.6%	99.2%
MARA	91.4%	99.3%	106.6%	98.7%
MedicaidRx	90.9%	100.3%	104.7%	101.5%
Wakely	90.4%	99.2%	107.6%	97.9%
Diagnosis-and-Pharmacy Models				
ACG System	93.6%	99.5%	104.4%	99.2%
CDPS+MRX	96.7%	101.4%	99.5%	102.1%
CRG	95.5%	101.1%	100.9%	101.0%
Impact Pro	96.6%	99.1%	101.7%	101.3%
MARA	95.7%	99.4%	103.0%	99.9%
Wakely	97.6%	99.0%	102.5%	98.7%
Prior Cost Models				
ACG System	97.1%	98.9%	102.4%	99.8%
DxCG	98.6%	99.7%	100.2%	101.5%
MARA	95.9%	99.3%	102.8%	100.4%
SCIO	102.5%	97.0%	102.4%	97.9%

4.5 TOLERANCE CURVES

In addition to the more conventional measures of predictive accuracy presented in previous sections, we have also developed a new visual and quantitative approach to compare the predictive accuracy of various models. This approach answers the following question: *for a given risk scoring methodology, what percentage of individuals are predicted accurately within an absolute error of X points?* For example, within the concurrent diagnosis-only models, we calculated that the CDPS risk score was accurate within 0.20 points[3] of the actual scaled cost level for 31.5 percent of individuals, but the HHS-HCC model was accurate to the same tolerance for 44.1 percent of individuals. We calculated this cumulative distribution of model error for each model and for each error tolerance level between 0.01 and 5.00. A larger percentage of individuals within each tolerance range indicates a more accurate model, by this metric.

Table 4.5.1 shows a selection of these cumulative distribution values for each of the concurrent diagnosis-only models. In addition to the tested risk scoring models, we have also included a naïve model baseline which is constructed using age and sex only (shown in Table 4.5.1 as "Age-Sex" model). The age-sex only approach was developed by computing the relative cost of each age and sex cell compared to the overall average. This is intended to represent the predictive power that can be achieved using traditional actuarial approaches without health-based risk scoring. Even an exceptionally naïve approach of assigning each individual a risk score of 1.0 would be very close for some percentage of individuals, so we believe that this inclusion of a baseline model is important for comparison purposes.

Table 4.5.1: Cumulative Distribution of Error – Concurrent Dx-Only Models

	Pr(|Error|) < 0.25	Pr(|Error|) < 0.50	Pr(|Error|) < 0.75	Pr(|Error|) < 1.00	Pr(|Error|) < 1.50	Pr(|Error|) < 2.00	Pr(|Error|) < 2.50
ACG System	58.8%	72.8%	79.7%	83.9%	88.9%	91.9%	93.8%
CDPS	43.4%	67.0%	75.8%	81.3%	87.6%	91.5%	93.5%
DxCG	58.6%	73.3%	80.9%	85.4%	90.5%	93.2%	94.9%
HHS-HCC	44.5%	68.3%	80.0%	83.6%	88.7%	91.2%	93.1%
MARA	59.1%	74.3%	82.2%	86.8%	91.7%	94.1%	95.4%
Truven	61.0%	74.7%	81.8%	86.1%	90.9%	90.5%	95.0%
Wakely	54.4%	69.6%	77.9%	83.1%	89.1%	92.2%	94.0%
Age-Sex	14.3%	40.7%	55.8%	67.1%	84.8%	92.5%	94.8%

Comparing model fit in this manner most closely aligns with the concept of the accuracy of individual risk estimation (i.e., what is the probability that the model accurately predicts the risk of an individual within a certain tolerance?). Most evident at the lower end of the tolerance spectrum, the models all provide a significant improvement to age-sex rating alone. This stems from risk scoring models' ability to differentiate very low risk individuals from higher cost individuals. That advantage wanes as the tolerance increases, with age-sex rating being approximately as accurate as the diagnosis-based models at predicting an individual's risk level within a tolerance of 2.0.

Although Table 4.5.1 was provided to help illustrate the mechanics of this concept, a more useful comparison would utilize all of the data points along the continuum. In Figure 2, we have plotted the percentage of individuals within each individual tolerance level for all of the concurrent diagnosis-only models. Similar graphs are provided for the other model comparisons in Appendix II. From Figure 2, it is clear that the models all provide a significant advantage over age-sex rating. At the lower end of the tolerance spectrum, the CDPS and HHS-HCC models underperform the other models. The models' performance is more similar at the higher end of the curve. We should note that the choice of an end

[3] By *points*, we are referring to risk score units, where the risk scores have been standardized to a mean of 1.0.

point for this graph is somewhat arbitrary, as the curves would not reach 100 percent until the maximum error is displayed, which can be quite high. For this illustration we selected 1.0 as the maximum tolerance threshold. The curves can be viewed as a continuous counterpart to the receiver-operator characteristic (ROC) curves that are frequently used for evaluation of binary classifications.

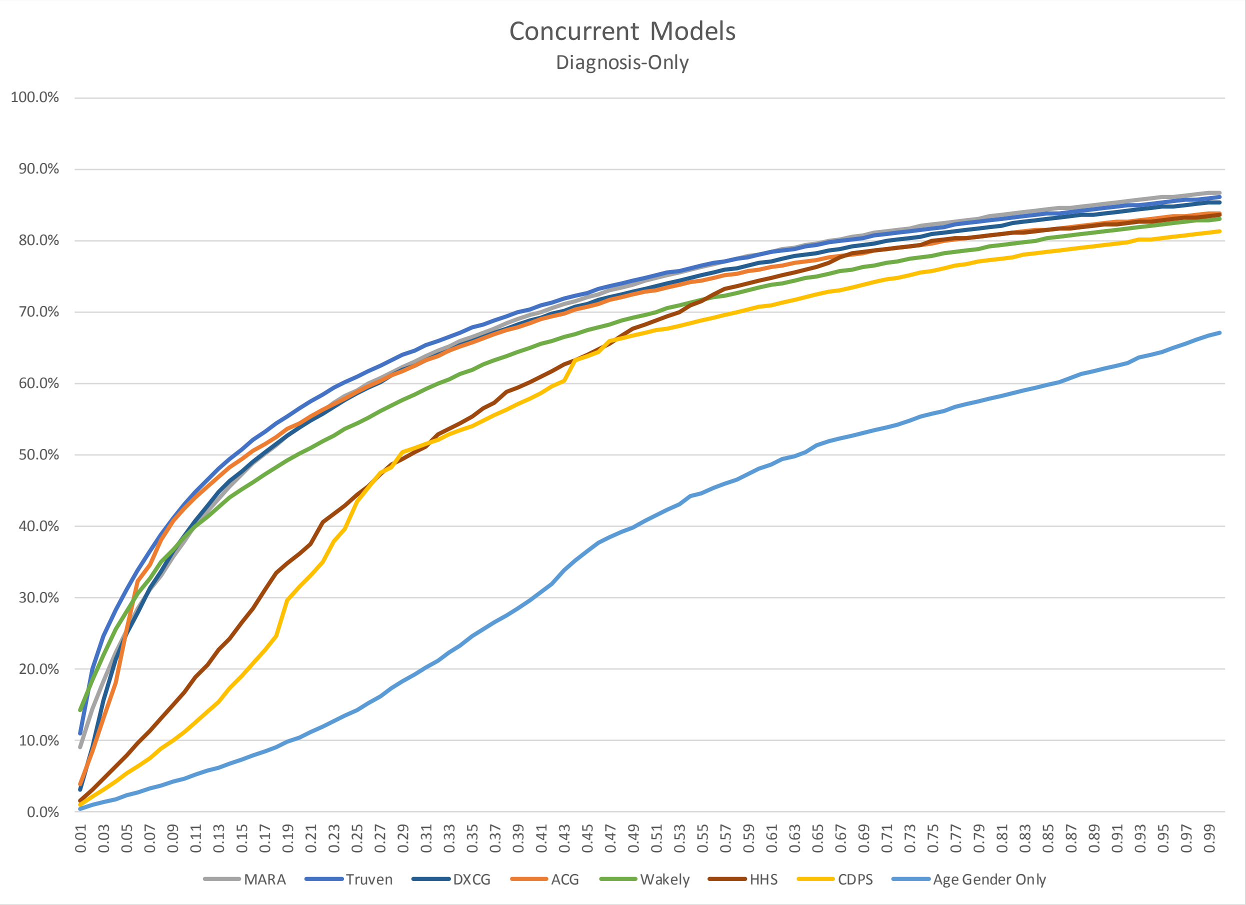


Figure 2: Tolerance Curves for Concurrent (DX) Models

In order to compare the models in a systematic fashion using this concept, we have calculated the area under each curve (presented in Tables 4.5.3 and 4.5.4). Again, the choice of a right-hand endpoint is arbitrary. Because of this, we have calculated the area under the curve separately through a maximum tolerance of 1.0 and through a maximum tolerance of 3.0. In the latter case, we have divided the area by 3.0 so that 100 percent would represent a perfect model – that is, one that exactly predicts each individual's risk. For comparison purposes, the age-sex only baseline model covered 36.3 percent of the area under the curve with a maximum tolerance of 1.0 and 71.5 percent of the area under the curve with a maximum tolerance of 3.0. All models are shown in their uncensored forms and with offered weights.

Table 4.5.2: Area Under Tolerance Curves (AUC), Concurrent Models (Offered Weights, Uncensored)

Diagnosis-Only Models		
	AUC through 1.0	AUC through 3.0
ACG System	66.8%	83.0%
CDPS	56.5%	79.0%
DxCG	67.1%	84.0%
HHS-HCC	59.5%	80.3%
MARA	68.1%	84.9%
Truven	69.1%	84.9%
Wakely	64.6%	82.4%
Pharmacy-Only Models		
DxCG	60.2%	80.7%
MARA	60.3%	80.6%
MedicaidRx	51.1%	77.0%
Wakely	58.8%	78.9%
Diagnosis-and-Pharmacy Models		
ACG System	68.0%	84.0%
CDPS+MRX	58.1%	79.5%
CRG	63.1%	82.1%
MARA	70.9%	86.3%
Wakely	63.9%	82.7%

Table 4.5.3: Area Under Tolerance Curves (AUC), Prospective Models (Offered Weights, Uncensored)

Diagnosis-Only Models		
	AUC through 1.0	AUC through 3.0
ACG System	50.8%	76.6%
CDPS	43.6%	74.4%
DxCG	51.1%	76.4%
Impact Pro	50.8%	77.2%
MARA	50.6%	77.2%
Truven	53.3%	77.5%
Wakely	50.5%	76.2%
Pharmacy-Only Models		
ACG System	48.9%	75.8%
DxCG	48.4%	75.8%
Impact Pro	48.7%	76.0%
MARA	49.7%	76.2%
MedicaidRx	41.6%	74.2%
Wakely	48.5%	75.0%
Diagnosis-and-Pharmacy Models		
ACG System	53.6%	77.4%
CDPS+MRX	46.0%	74.9%
CRG	51.0%	76.8%
Impact Pro	52.2%	77.7%
MARA	52.9%	78.1%
Wakely	52.6%	77.0%
Prior Cost Models		
ACG System	53.7%	77.6%
DxCG	56.3%	79.0%
MARA	54.2%	78.7%
SCIO	54.1%	77.8%

4.6 IDENTIFICATION OF HIGHEST EXPENDITURE INDIVIDUALS

In addition to the calculation of the average risk score for a group of individuals, risk scoring models are also often used to identify the very highest cost individuals. This functionality can be a component of care management programs or other efforts to identify and manage the costs of the highest utilizing individuals in a population.

A method for comparing the ability of various models in fulfilling this role was demonstrated in a 2014 Society of Actuaries Health Meeting presentation (Leida & Siegel, 2014). We have reproduced this demonstrated approach for each of the models included in this study. First, we identified the individuals whose healthcare expenditures were in the top one percent of individuals in our analytic sample. This equated to about 13 times the level of an average individual in the sample. Then, we constructed receiver operating characteristic, or ROC, curves for each included model. ROC curves are visual representations of the relationship between the specificity and sensitivity of predictions of individuals being within the top one percent. Specificity is the "true negative rate", or the percentage of individuals that are correctly identified as not being in the top one percent. Sensitivity is the "true positive rate", which is the percentage of individuals that are correctly identified as being among the top one percent. As an illustration, Figure 3 below shows the plotted ROC curve for the HHS-HCC model.

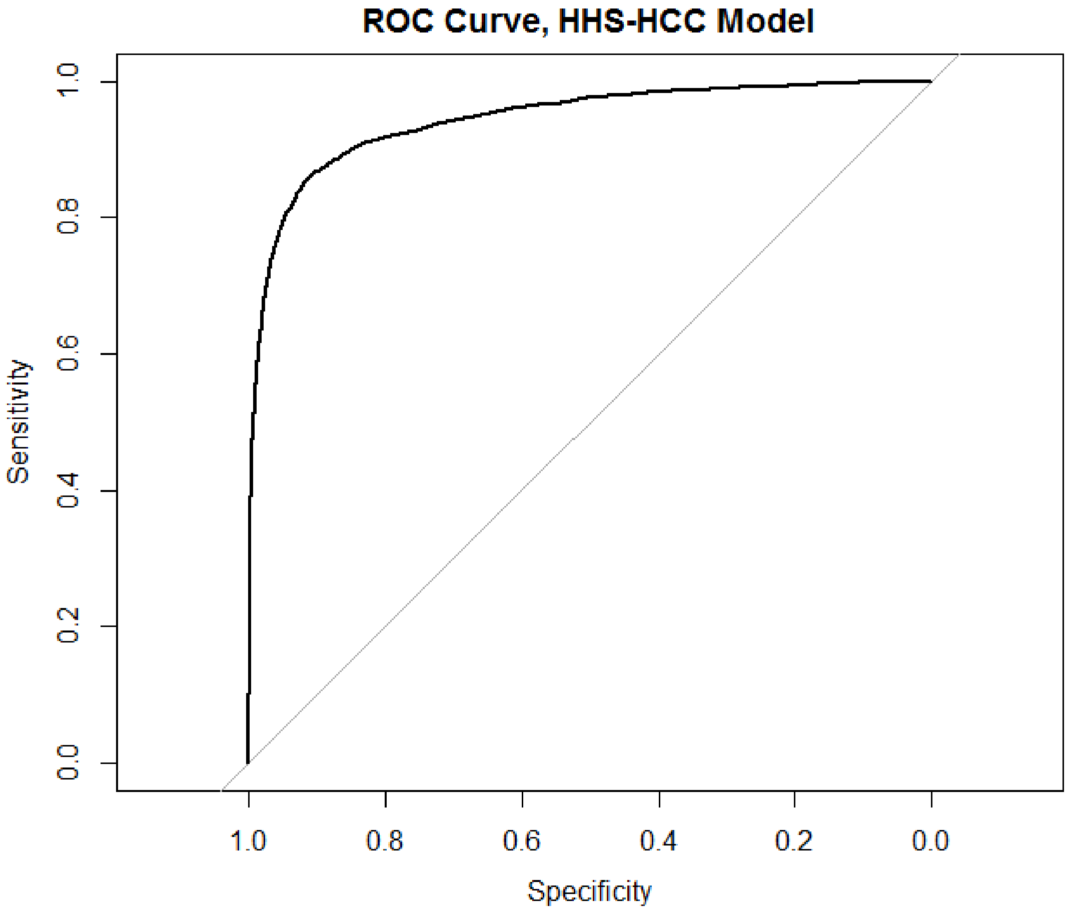


Figure 3: ROC Curve for HHS-HCC Model

ROC curves are typically compared by calculating the area under the curve. A perfect model would have an area under the ROC curve (AUC) of 1.0, compared to a naïve model of random guesses (represented by the faint line in Figure 3), which would have an AUC of 0.5. We have calculated the AUCs for all of the included models and shown the results below in Figure 4 and Figure 5.

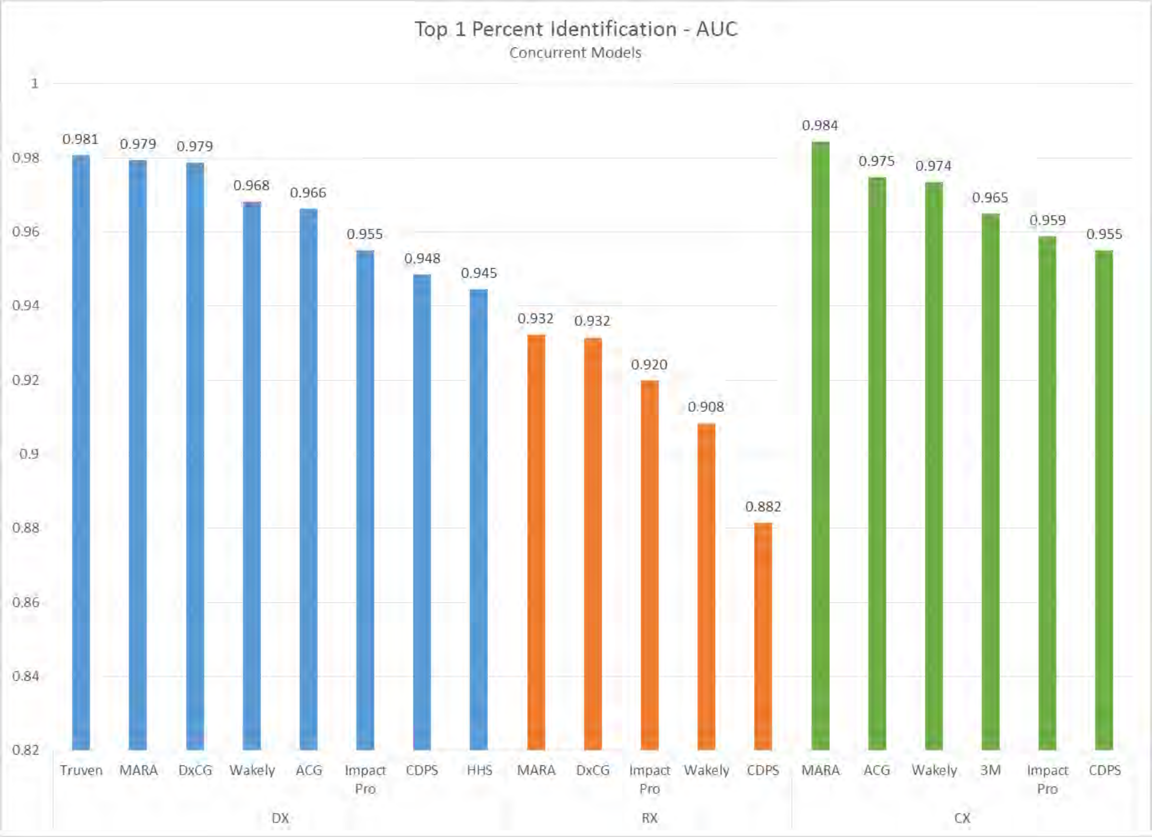


Figure 4: Area Under ROC Curves, Concurrent Models

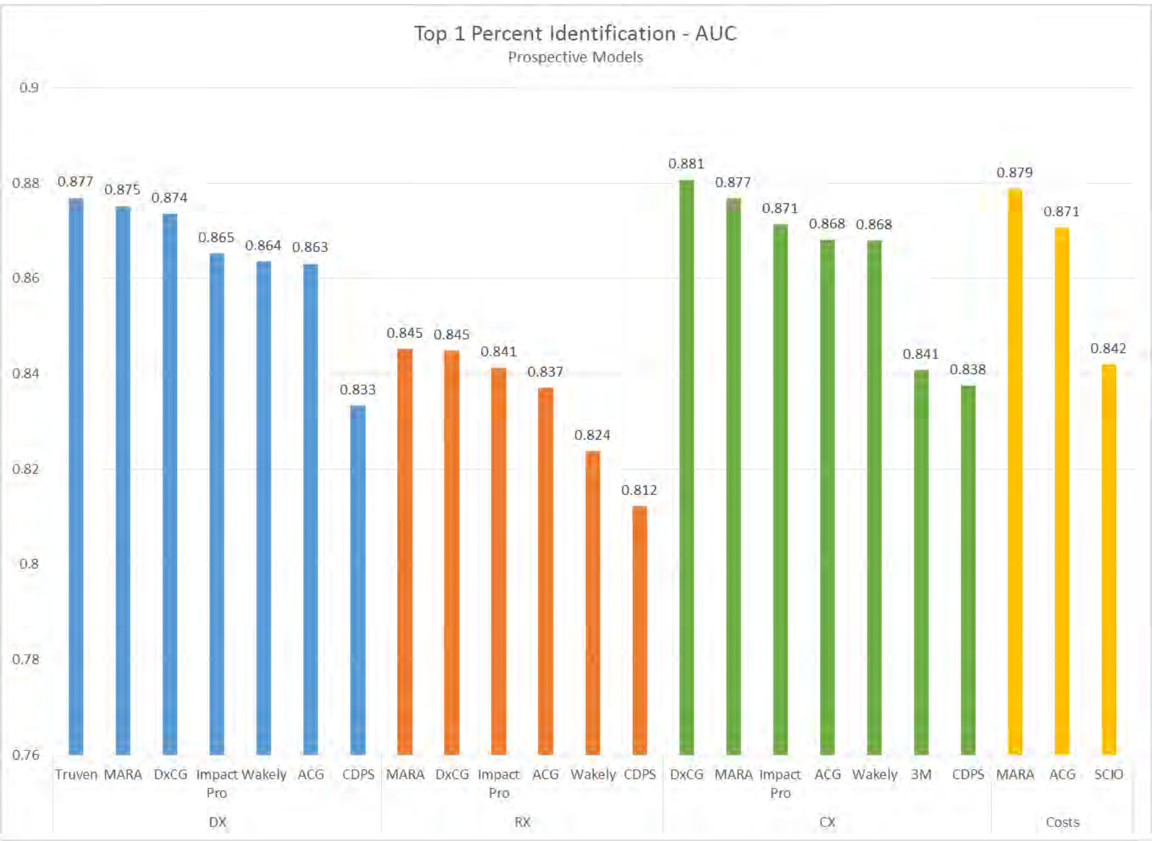


Figure 5: Area Under ROC Curves, Prospective Models

The AUC metrics displayed in the figures above generally mirror the clustering of R-Squared values with the various families of models. However, we note that the levels of the AUCs are quite high and indicate that all of the risk scoring models studied perform very well at identifying the very highest healthcare spenders. This long tail of high-cost individuals, which is partially responsible for the relatively low R-Squared values for risk scoring models in general, actually benefits the goal of trying to predict the very highest cost expenditures. These individuals have distinctive patterns of healthcare utilization that are readily captured by diagnosis- and pharmacy-based risk scoring methods.

4.7 ACCURACY IN PREDICTING BIASED GROUPS

As we demonstrated in Table 4.4.6 and Table 4.4.7, risk scoring models tend to underpredict expenditures for higher-cost individuals and overpredict expenditures for lower-cost individuals. This tendency is a result of the finite number of combinations of independent variables and the fact that there is more variation than the independent variables can explain. For a given set of values of independent variables in a risk scoring model (for example, a diagnostic and demographic profile), there is a distribution of possible cost outcomes that is associated with that profile. Since a single risk score value is associated with that diagnostic and demographic profile, the individuals who share that profile but have higher costs will be underestimated while the individuals with that profile but lower costs will be overestimated.

While this fact is easily understandable and may seem quite obvious, it leads to an important implication of using risk scoring models to normalize health plan premiums: risk scoring models cannot completely compensate for adverse selection. If costlier-than-average individuals select into a health plan, risk scoring models will produce risk scores above average, but the expected value of the risk score will be somewhat less than the actual associated risk. By contrast, if healthier-than-average individuals select into a health plan, the expected risk score will be somewhat higher than the actual risk. We have modeled the degree to which this is captured by each of the included risk scoring models.

In order to simulate the effect of adverse selection in a population, we have repeatedly drawn biased samples of individuals from our analytic sample. First, we drew 100 samples of 10,000 individuals representing a "moderately adverse" population with an average cost equal to 109 percent of the general sample average. For each of the 100 samples, we computed the average cost and the average risk score using each of the concurrent models. We used the diagnosis-only version of the models, where such a model was provided by a vendor. In this moderately adverse scenario, the Truven model produced a risk score that was 1.1 percent lower than the average risk score in the median case of the 100 iterations, while the MARA, DxCG, Wakely, and ACG System models achieved a median error within two percent. The median error for each model is displayed in Figure 6 below, along with the 10th and 90th percentile of the error.

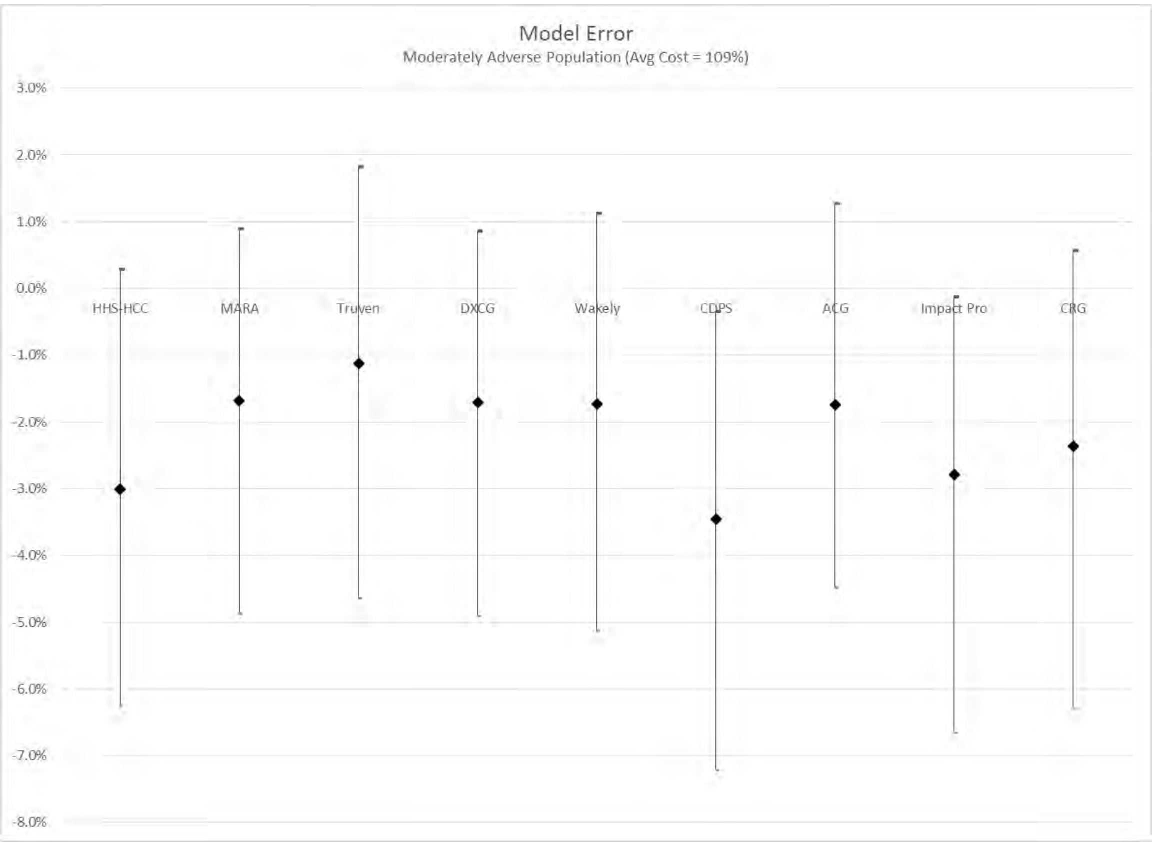


Figure 6: Calculate Error Distribution for a Moderately Adverse Population

We also drew 100 iterations of a sample of 10,000 individuals to produce a "highly adverse" population with average costs equal to 121 percent of the overall average. In this scenario, the Truven model (1.8

percent) and ACG System (2.1 percent) achieved the lowest median error rates. These results are shown below in Figure 7.

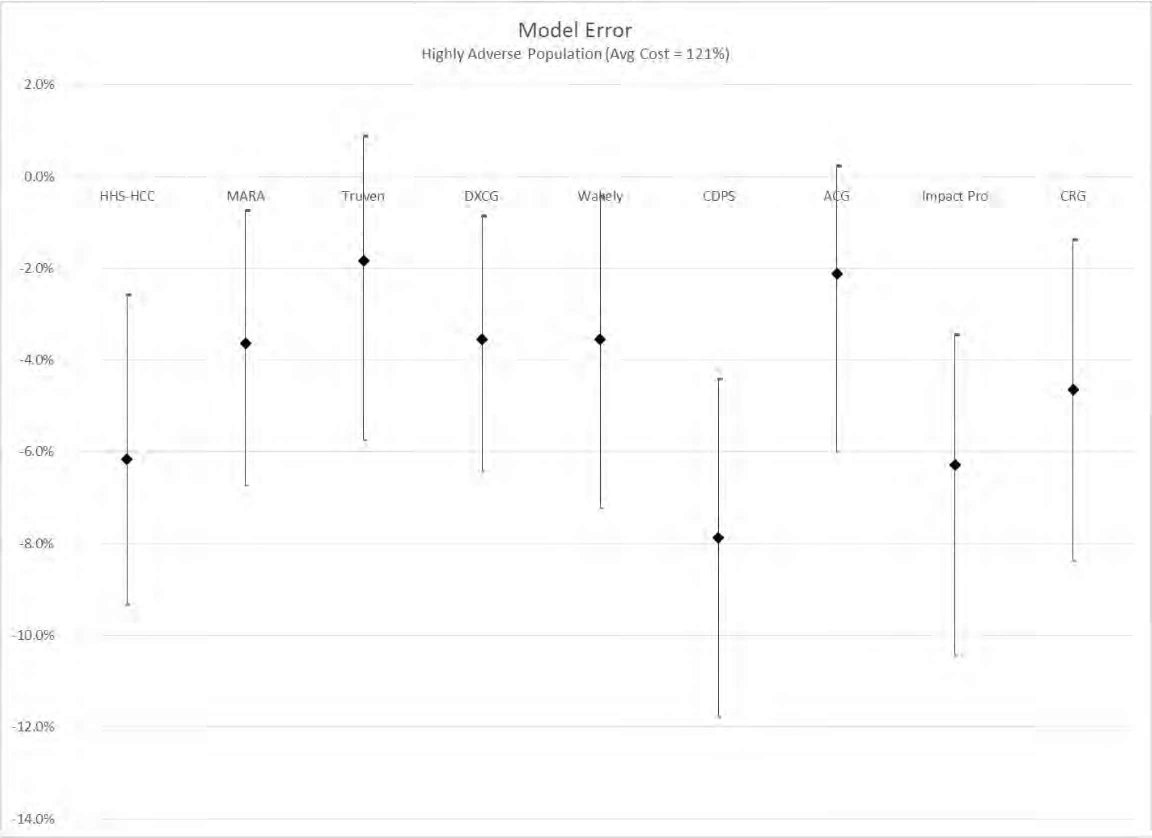


Figure 7: Calculate Error Distribution for a Highly Adverse Population

Both of these scenarios illustrate that risk scoring models are not likely to fully compensate for adverse selection. The best-fitting risk scoring models are able to close most of this gap, but some bias is likely to remain with any approach.

Section 5: Additional Analyses

5.1 COMPARISON TO HHS-HCC MODEL

Since the completion of the previous SOA comparison study (Winkelman & Mehmud, 2007), the most pervasive change in the healthcare system has been the passage in 2010 of the Affordable Care Act. Of particular relevance to the field of risk adjustment, the absence of medical underwriting for individual and small group health insurance resulted in the need for a risk adjustment mechanism to normalize plan payments across the commercial marketplace. This has been accomplished through the use of a risk scoring model developed by the Centers for Medicare and Medicaid Services and known as the HHS-HCC model. Additional details concerning this model and its development were published by CMS (Kautter, et al., 2014).

Given the importance of the HHS-HCC model and its scores in the current market, we have computed several metrics to compare the tested models' scores to those produced by the HHS-HCC model. For concurrent models, we first calculated the Pearson correlation coefficient of each model's risk score and the HHS-HCC risk scores in order to see how similar each model's risk scores are to those produced by the HHS-HCC model. These results are summarized in Table 5.1.1. While the correlations are quite high in general, there is a degree of variation from one model to the next. The highest correlations appear to be among the models that, like the HHS-HCC model, use only diagnoses as inputs. With the exceptions of CDPS and ImpactPro, the models that are available in both DX-only and DX+RX form have correlations about two percentage points higher in the DX-only version. These models are more closely aligned to the HHS-HCC model because it is also a diagnosis-only model.

Table 5.1.1: Correlation Coefficient Between HHS-HCC and Selected Models

Diagnosis-Only Models	
ACG System	0.851
CDPS	0.643
DxCG	0.821
MARA	0.844
Truven	0.813
Wakely	0.859
Pharmacy-Only Models	
DxCG	0.531
MARA	0.545
MedicaidRx	0.364
Wakely	0.447
Diagnosis-and-Pharmacy Models	
ACG System	0.811
CDPS-MRx	0.644
CRG	0.767
MARA	0.824
Wakely	0.834

For prospective models, we have focused instead on the potential for prospective risk scores to be used to identify individuals or groups of individuals who are likely to be overvalued or undervalued by the HHS-HCC model. In order to evaluate this, we computed the "HCC error" for each individual in our sample, defined as the difference between the HHS risk score and the actual scaled cost level. We then specified a linear regression model with the HHS error as the dependent variable and each model's prospective risk score as the lone independent variable. For every prospective model, our regression analysis produced a highly significant negative coefficient, which suggests that high prospective risk scores are likely to identify individuals for whom the HHS-HCC risk score will undervalue risk. A closer inspection reveals a more complex relationship. In Table 5.1.2, we have displayed the mean and median HHS-HCC error for various percentile ranges of the MARA prospective model[4] using diagnoses, pharmacy, and prior year costs. We note that we observed similar results with other risk scoring models.

[4] MARA was selected as an illustrative model, other models produced similar results.

Table 5.1.2: Mean and Median HHS-HCC Error for Various Percentiles of Prospective Risk Scores

Percentile Range (Risk Score Range)	Mean HHS Error	Median HHS Error
Bottom 20% (0-0.231)	0.07	0.13
20% to 40% (0.231-0.453)	0.07	0.14
40% to 60% (0.453-0.685)	0.12	0.21
60% to 80% (0.685-1.240)	0.19	0.31
Top 20% (1.240+)	-0.4	0.04
Top 10% (2.007+)	-0.75	-0.12
Top 5% (3.071+)	-1.25	-0.33
Top 2% (5.217+)	-2.25	-0.65
Top 1% (7.715+)	-3.42	-0.67
Top 0.5% (11.647+)	-5.26	-0.75

We then calculated the mean and median HHS error within each band of prospective risk score in increments of 0.1 risk score units. Figure 8 shows the relationship very clearly: individuals with low prospective risk scores are most likely to result in overpayment under the HHS-HCC model and those individuals with high prospective risk scores are most likely to result in underpayment. While HHS-HCC error generally decreases as the prospective risk score increases, the individuals with the largest difference between the HHS-HCC risk score and actual costs appear to be those with prospective risk scores between 0.45 and 0.75.

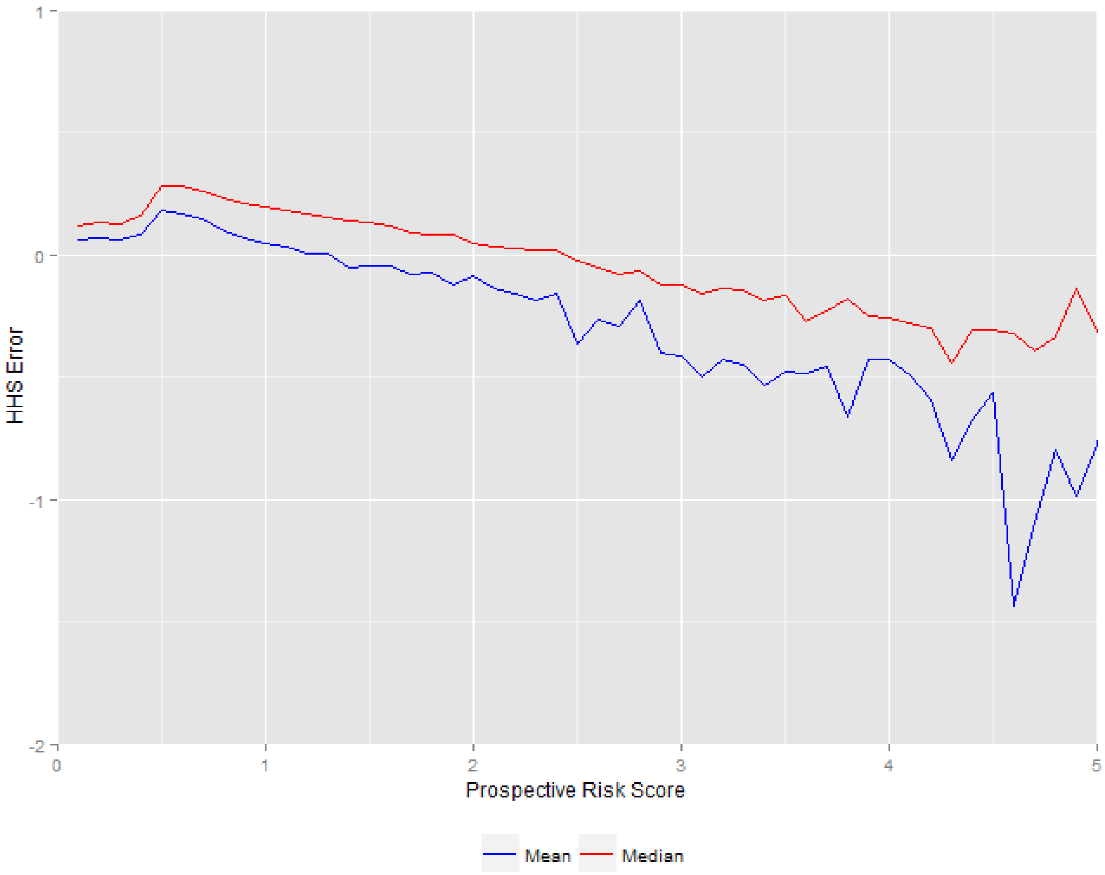


Figure 8: Relationship Between Prospective Risk Scores and HHS-HCC Error

5.2 ENSEMBLE MODELS

In the course of conducting this study, we were in the relatively unique position of having simultaneous access to risk scores produced by all of the participating models. As one additional analytic question, we explored whether an ensemble model built as a composite from these models could provide significant improvements in predictive power over the best fitting stand-alone models.

We have built two sets of simple ensemble models from the existing risk scores, two each for prospective and concurrent models. We only included the broadest set of inputs from each vendor. For example, the MARA prospective models are available with diagnoses only, pharmacy only, diagnoses and pharmacy, or diagnosis and pharmacy plus prior costs. We included only the prospective MARA model with diagnosis, pharmacy and cost inputs. Verisk did not provide a concurrent model with both diagnosis and pharmacy inputs, so we included both the diagnosis-based and pharmacy-based models as potential inputs for the ensemble models.

For the simplest blended models, we calculated the unweighted mean of the candidate models (there were seven concurrent models and eight prospective models used). With even this very simple approach, we were able to achieve R-Squared values at least as accurate as the most accurate standalone model. The concurrent mean ensemble model produced an R-Squared of 55.7 percent, compared to the best-fitting offered model's (MARA DX+RX) R-Squared of 55.4 percent. The prospective mean ensemble model equaled the best-fitting prospective model's (MARA DX+RX with Costs) R-Squared of 24.8 percent.

We then constructed a second pair of ensemble models by using a linear regression model to determine the weights for each of the models. To do this, we specified linear regression equations to predict the actual cost from the set of candidate models. Models with negative or insignificant coefficients were discarded in stepwise fashion. We also forced the intercept to zero, so that the resulting set of regression coefficients would provide weights for the models to be included in the ensemble model. The weights were specified on the same sample of one million records that was used for the recalibration process. The statistics were calculated on the sample of individuals that was used for all of the other model testing. Most of the candidate models were eliminated from the regression model through the stepwise process. Table 5.2.1 shows the final models and associated weights.

Table 5.2.1: Composition of Weighted Ensemble Models

Concurrent Models		Prospective Models	
Model	Weight	Model	Weight
MARA (DX+RX)	0.459	MARA (DX+RX+Costs)	0.425
Truven (DX)	0.298	Verisk (DX+RX)	0.298
Verisk (DX)	0.205	Truven (DX)	0.221
Verisk (RX)	0.038	SCIO (DX+RX+Cost)	0.056

Table 5.2.2 shows the R-Squared and MAE of both ensemble models, compared to the highest-scoring models as offered by the vendors.

Table 5.2.2: R-Squared and MAE of Ensemble Models

	Concurrent Models		Prospective Models	
	R-Squared	MAE	R-Squared	MAE
Best-Fit Single Model	55.4%	57.9%	24.8%	91.8%
Mean Ensemble	55.7%	63.1%	24.8%	92.5%
Weighted Ensemble	58.2%	57.6%	26.4%	90.3%

Despite the slight gain in predictive power above the offered models, applying such an ensemble is probably impractical in most applications due to the licensing and implementation costs of each model.

5.3 EXPLORATION OF A MACHINE LEARNING IMPLEMENTATION (Assisted by: Forecast Health)

In addition to the models provided by the vendors discussed throughout this paper, we also considered including machine learning models built by Forecast Health. Forecast provided a concurrent and a prospective model that had been trained on Medicare encounter data. Those Medicare-specific models did not perform as well on the MarketScan commercial data as Forecast had observed on Medicare data. As a consequence, Forecast elected to not include their models in the comparison study.

However, because their approach was so unique, we worked collaboratively with Forecast's team to specify a prospective model with our training dataset for the specific purpose of demonstrating a proof of concept for the application of this approach to risk scoring for a commercial population. The final specified model was not retained by Forecast, but used exclusively for calculating fit statistics in this report.

The model that we developed with Forecast used diagnoses, pharmacy data, and prior year cost to predict relative risk in the subsequent year. The development of this test model entailed the use of an ensemble model which combined results from several thousand independent models developed on bootstrapped samples of the training set. This technique is often called a random forest.

The results of this analysis were illuminating, both to the potential of this approach to risk scoring and to the perils of relying exclusively on R-Squared. We calculated the R-Squared for the Forecast model on both an uncensored and censored basis and found dramatically different results. For the uncensored model, the R-Squared was 1.3 percent, compared to 20.8 percent with censoring at $250,000. This compared to a range of 15.1 to 24.8 percent for the other prospective models with the same types of inputs on an uncensored basis and a range of 22.4 to 26.9 percent with censoring at $250,000. Based on R-Squared alone, the Forecast model appears to be quite poor without censoring and somewhat less accurate than the field with censoring.

The mean absolute error tells a very different story. With the Forecast approach, we found a MAE of 68.4% without censoring, compared to a minimum of 91.8% among the tested models. With censoring, we found a MAE of 77.5%, which compares to a minimum of 90.1% among the tested models. We were surprised that these two metrics could paint such different pictures of the potential predictive accuracy of this machine learning approach. Also by way of contrast with the R-Squared result, we calculated the AUC for the identification of the top one percent of spenders. The Forecast model produced an AUC value of 0.960, where none of the tested models shown in Figure 5 exceeded a level of 0.881.

In order to further understand the R-Squared result for the uncensored model, we divided the one-million-person test sample into ten equal-sized random subsets and calculated the R-Squared within each of these subsets. We found that the lowest R-Squared among those ten subsets was 0.5 percent, but that the remaining nine subsets had R-Squared values which ranged from 15.0 percent to 30.4 percent. Upon closer inspection, we found that a single data point with a predicted cost of over $37 million and actual costs of $267,000 was causing a tremendous influence on the R-Squared. In fact, removing that single observation from the dataset increased the overall R-Squared from 1.3 percent to 19.7 percent. While this is an extreme case that could easily be resolved by a small adjustment to the model logic (such as limiting the risk score to a specified threshold), it serves as a vivid illustration of the influence that outliers can have on the individual R-Squared measure.

We concluded that applications of machine learning to risk scoring, such as the concurrent and prospective risk models that Forecast has developed for Medicare, have great promise for commercial populations as well.

Section 6: Concluding Statements

Since the passage and implementation of the Affordable Care Act, risk scoring models have assumed a very important role in the health care financing environment. In the small group and individual marketplaces, the risk adjustment process has now replaced the historical practice of medical underwriting. As such, it is essential that health actuaries have a thorough understanding of the mechanics of these models, their strengths and weaknesses, their biases, and their ability to predict or explain healthcare expenditures.

This paper has examined the relative predictive abilities of over 40 such risk scoring models provided by eleven distinct vendors. We have used a variety of analytical techniques to quantify how closely these models are able to estimate actual healthcare expenditures for individuals and for groups of individuals. Through this process, some models appear to be better on one scale while others perform better on another scale.

We believe that it is essential that risk scoring models should be evaluated in a manner consistent with the application for which the models are being selected. For example, if a risk scoring model is to be used to identify the highest-cost individuals as candidates for case management, then the best measure to use would be one similar to our analysis of accuracy in identifying the top one percent of spenders. An actuary or underwriter who is using a risk scoring model to estimate prospective risk for a renewing employer group would likely want to consider the error distribution in estimating groups of similar sizes. At other times, some applications may call for predictive ratios that are close to 1.0, indicating minimal prediction bias.

We also believe that although the individual R-Squared statistic is a convenient single measure that does provide some information about the fit of a model, it should never be the sole determinant of a model's success or failure, either in developing or selecting a risk scoring model. As we illustrated in Section 5.3, such a reliance may lead a practitioner to completely dismiss a superior model that is simply not handling outliers in an efficient manner.

Section 7: Reliances and Limitations

Most critically, this project is reliant on the accuracy and completeness of the sampled MarketScan data. While we did not perform an audit of the provided data, Truven Health Analytics has a long and demonstrated track record of providing accurate data in its MarketScan databases. We performed reasonableness checks on the required data elements and shared intermediate results with each vendor to validate our findings.

While we have attempted to use a variety of approaches to compare the various risk scoring models, we note that any comparison is limited. These quantitative comparisons should not be interpreted as an authoritative ranking of the models or their usefulness in a variety of applications. The metrics we have calculated may vary across different populations, time periods, or actuarial applications.

Neither Kennell and Associates nor the specific authors of this study have any conflicts of interest regarding any of the vendors participating in the study. Kennell has purchased data from Truven Health Analytics for this and other efforts. Kennell has licensed a grouper from 3M Health Information Systems in the past for its work with the Defense Health Agency. Kennell has also subcontracted with Wakely Consulting Group on a different research project for the Society of Actuaries and for a project with the Defense Health Agency. None of these relationships affected the objectivity of the study. Additionally, no members of the Project Oversight Group are currently employed by any of the model vendors.

References

Actuarial Standards Board. (2012, January). *Actuarial Standard of Practice No. 45: The Use of Health Status Based Risk Adjustment Methodologies.*

Anscombe, F. (1973). Graphs in Statistical Analysis. *American Statistician*, 17-21.

Cumming, R., Knutson, D., Cameron, B., & Derrick, R. (2002). *A Comparative Analysis of Claims-Based Methods of Health Risk Assessment for Commercial Populations.* Society of Actuaries.

Dunn, D. L., Rosenblatt, A., Taira, D. A., Latimer, E., Bertko, J., Stoiber, T., . . . Busch, S. (1996). *A Comparative Analysis of Methods of Health Risk Assessment.* Society of Actuaries.

Hileman, G. (2015, December). A Comparison of Risk Scoring Recalibration Methods. *Predictive Analytics and Futurism Section Newsletter*. Retrieved from https://www.soa.org/Library/Newsletters/Predictive-Analytics-and-Futurism/2015/december/paf-iss12.pdf

Hileman, G. R., Rosenberg, M., & Mehmud, S. M. (2016). *Risk Scoring: A Primer.* Society of Actuaries.

Kautter, J., Ingber, M., Pope, G. C., & Freeman, S. (2012). Improvements in Medicare Part D Risk Adjustment: Beneficiary Access and Payment Accuracy. *Medical Care, 50*(12), 1102-1108. doi:10.1097/MLR.0b013e318269eb20

Kautter, J., Pope, G. C., & Keenan, P. (2014). Affordable Care Act Risk Adjustment: Overview, Context, and Challenges. *Medicare and Medicaid Research Review, 4*(3). doi:10.5600/mmrr.004.03.a02

Kautter, J., Pope, G. C., Ingber, M., Freeman, S., Patterson, L., Cohen, M., & Keenan, P. (2014). The HHS-HCC Risk Adjustment Model for Individual and Small Group Markets under the Affordable Care Act. *Medicare and Medicaid Research Review, 4*(3). doi:10.5600/mmrr.004.03.a03

Leida, H., & Siegel, J. (2014). ACA Risk Adjustment: Performance Measures, Potential Bias, and Strategy. *Society of Actuaries Spring Health.*

Mehmud, S. M., & Yi, R. (2012). *Uncertainty in Risk Adjustment.* Society of Actuaries. Retrieved from https://www.soa.org/research/research-projects/health/uncertainty-risk-adjustment.aspx

Parkes, S. &. (2015, July). Calibrating Risk Score Models with Partial Credibility. *Forecasting and Futurism Section Newsletter*. Retrieved from https://www.soa.org/Library/Newsletters/Forecasting-Futurism/2015/July/ffn-2014-iss10.pdf

Pope, G. C., Bachofer, H., Pearlman, A., Kautter, J., Hunter, E., Miller, D., & Keenan, P. (2014). Risk Transfer Formula for Individual and Small Group Markets Under the Affordable Care Act. *Medicare and Medicaid Research Review, 4*(3). doi:10.5600/mmrr.004.03.a04

Sawhney, T. G. (2013). *Health Insurance Risk Adjustment: Looking Beyond R-Squared.* Retrieved from https://www.soa.org/files/pd/2013/health-mtg/2013-maryland-health-mtg-62.pdf

Schone, E., & Brown, R. (2013). *Risk Adjustment: What is the Current State of the Art and How Can It Be Improved?* Robert Wood Johnson Foundation.

Winkelman, R., & Mehmud, S. (2007). *A Comparative Analysis of Claims-Based Tools for Health Risk Assesment.* Society of Actuaries. Retrieved from https://www.soa.org/research/research-projects/health/hlth-risk-assement.aspx

Appendix I Predictive Ratios

I.A Concurrent Models, No Censoring

Table I.A.1: Predictive Ratios by Health Conditions (Concurrent; Offered Weights; No Censoring)

	Heart Disease	Mental Illness	Diabetes	Low Back Pain	Asthma	Arthritis
Diagnosis-Only Models						
ACG System	104.3%	100.7%	103.4%	98.8%	97.8%	96.3%
CDPS	67.7%	86.5%	77.9%	69.5%	98.6%	66.5%
DxCG	98.6%	97.3%	98.4%	96.6%	89.8%	83.4%
HHS-HCC	96.1%	85.5%	102.2%	74.6%	96.8%	97.4%
MARA	89.1%	96.8%	93.3%	95.3%	93.7%	87.8%
Truven	101.1%	101.3%	97.9%	103.4%	94.6%	93.7%
Wakely	93.1%	98.0%	97.9%	94.8%	91.7%	89.7%
Pharmacy-Only Models						
DxCG	78.5%	85.7%	92.0%	87.3%	81.0%	87.8%
MARA	72.8%	86.2%	89.3%	86.2%	86.1%	88.3%
MedicaidRx	47.1%	78.8%	70.9%	64.8%	80.7%	65.9%
Wakely	66.6%	88.1%	90.7%	86.6%	80.4%	85.4%
Diagnosis-and-Pharmacy Models						
ACG System	95.6%	98.5%	96.5%	95.9%	94.1%	88.1%
CDPS-MRx	71.1%	93.1%	84.4%	72.3%	96.1%	72.4%
CRG	87.7%	95.2%	96.0%	86.1%	94.7%	97.4%
MARA	89.6%	97.2%	94.1%	95.8%	96.1%	90.1%
Wakely	85.6%	95.0%	96.1%	92.8%	88.2%	88.7%

Table I.A.2: Predictive Ratios by Age-Sex (Children; Concurrent; Offered Weights; No Censoring)

	Children, Age 0-1	Children, Age 2-6	Children, Age 7-18
Diagnosis-Only Models			
ACG System	99.9%	102.3%	89.7%
CDPS	262.9%	201.5%	216.2%
DxCG	79.8%	87.5%	88.8%
HHS-HCC	81.8%	86.9%	88.5%
MARA	107.2%	114.3%	104.1%
Truven	94.1%	94.6%	91.4%
Wakely	102.5%	106.8%	100.1%
Pharmacy-Only Models			
DxCG	81.8%	91.8%	91.3%
MARA	85.9%	107.0%	99.4%
MedicaidRx	343.0%	194.9%	219.5%
Wakely	85.0%	97.7%	101.6%
Diagnosis-and-Pharmacy Models			
ACG System	100.3%	107.1%	98.9%
CDPS-MRx	242.7%	185.9%	208.3%
CRG	118.7%	119.2%	108.7%
MARA	103.6%	112.9%	103.5%
Wakely	95.1%	101.6%	98.5%

Table I.A.3: Predictive Ratios by Age-Sex (Males; Concurrent; Offered Weights; No Censoring)

	Males, Age 19-22	Males, Age 23-24	Males, Age 25-29	Males, Age 30-34	Males, Age 35-39	Males, Age 40-44	Males, Age 45-49	Males, Age 50-54	Males, Age 55-59	Males, Age 60-64
Diagnosis-Only Models										
ACG System	96.4%	99.5%	95.6%	98.5%	101.6%	96.9%	100.1%	101.3%	101.5%	102.5%
CDPS	254.0%	184.2%	100.1%	99.0%	93.7%	85.4%	86.2%	81.9%	78.7%	75.0%
DxCG	98.8%	100.2%	100.6%	100.6%	104.1%	100.6%	102.0%	101.7%	103.7%	105.9%
HHS-HCC	104.1%	105.6%	97.0%	98.7%	100.4%	99.5%	99.8%	102.8%	106.0%	106.8%
MARA	103.1%	105.2%	101.0%	104.1%	101.2%	97.8%	96.9%	97.6%	97.9%	96.9%
Truven	100.9%	100.5%	99.9%	102.0%	101.5%	97.4%	97.7%	98.7%	100.8%	102.0%
Wakely	99.0%	104.1%	103.2%	105.4%	103.9%	97.5%	97.5%	98.7%	99.9%	98.1%
Pharmacy-Only Models										
DxCG	95.2%	99.4%	101.4%	108.8%	107.4%	99.7%	102.8%	101.2%	103.7%	103.9%
MARA	99.6%	99.7%	103.6%	112.1%	107.0%	101.5%	99.1%	100.0%	99.0%	96.5%
MedicaidRx	274.5%	191.5%	95.4%	101.1%	94.4%	84.5%	93.2%	84.9%	76.3%	67.1%
Wakely	103.2%	103.9%	104.2%	116.2%	109.7%	101.8%	100.1%	100.7%	99.2%	91.9%
Diagnosis-and-Pharmacy Models										
ACG System	97.6%	101.7%	100.8%	106.6%	105.0%	98.9%	98.6%	99.4%	97.2%	96.0%
CDPS-MRx	252.9%	183.8%	92.7%	95.9%	93.4%	87.5%	85.5%	83.0%	81.0%	78.0%
CRG	98.8%	100.1%	97.5%	100.2%	97.3%	90.4%	107.4%	104.2%	103.2%	99.6%
MARA	101.3%	104.4%	102.3%	106.4%	102.5%	98.9%	97.4%	97.9%	97.4%	96.8%
Wakely	100.6%	105.2%	105.9%	111.5%	108.2%	99.9%	97.1%	97.4%	97.5%	94.4%

Table I.A.4: Predictive Ratios by Age-Sex (Females; Concurrent; Offered Weights; No Censoring)

	Females, Age 19-22	Females, Age 23-24	Females, Age 25-29	Females, Age 30-34	Females, Age 35-39	Females, Age 40-44	Females, Age 45-49	Females, Age 50-54	Females, Age 55-59	Females, Age 60-64
Diagnosis-Only Models										
ACG System	95.8%	92.9%	93.0%	93.3%	97.4%	98.2%	103.2%	105.5%	103.1%	105.4%
CDPS	206.8%	136.9%	73.6%	73.2%	72.7%	71.7%	77.6%	76.3%	76.1%	76.8%
DxCG	96.7%	93.5%	92.5%	91.8%	98.0%	98.7%	100.7%	101.8%	102.3%	107.7%
HHS-HCC	96.5%	95.5%	93.0%	94.8%	97.7%	98.7%	99.6%	102.1%	102.7%	104.9%
MARA	105.4%	101.0%	96.7%	97.5%	99.1%	99.0%	99.2%	100.1%	99.9%	101.1%
Truven	100.9%	101.0%	103.9%	103.1%	100.6%	99.0%	101.5%	100.5%	100.9%	104.4%
Wakely	105.1%	101.1%	97.6%	98.8%	100.7%	99.9%	100.0%	101.0%	99.0%	100.4%
Pharmacy-Only Models										
DxCG	101.8%	102.1%	99.3%	100.2%	101.6%	102.3%	102.0%	98.7%	96.5%	100.5%
MARA	106.5%	105.2%	97.0%	101.6%	102.2%	102.9%	101.4%	100.0%	96.8%	97.1%
MedicaidRx	236.4%	166.1%	69.5%	72.3%	74.5%	76.7%	74.7%	72.9%	70.4%	68.9%
Wakely	106.1%	106.0%	99.2%	102.0%	102.3%	104.4%	102.1%	100.3%	97.0%	95.1%
Diagnosis-and-Pharmacy Models										
ACG System	104.4%	100.5%	99.3%	101.0%	102.7%	101.8%	102.7%	102.9%	98.2%	98.1%
CDPS-MRx	211.0%	140.9%	70.3%	72.2%	73.5%	74.7%	77.7%	77.6%	78.1%	79.6%
CRG	99.1%	95.2%	80.4%	82.4%	94.9%	103.6%	98.3%	99.1%	99.7%	98.7%
MARA	106.9%	101.4%	97.9%	99.8%	100.6%	100.2%	99.4%	99.8%	99.0%	99.3%
Wakely	107.5%	103.2%	97.4%	99.5%	101.6%	102.5%	102.4%	102.3%	99.6%	99.8%

Table I.A.5: Predictive Ratios by Cost Percentile (Concurrent; Offered Weights; No Censoring)

	0-20th Percentile	20-40th Percentile	40-60th Percentile	60-80th Percentile	80-90th Percentile	90-95th Percentile	95-98th Percentile	98th-99th Percentile
Diagnosis-Only Models								
ACG System	1427%	268%	197%	154%	124%	104%	84%	59%
CDPS	6827%	635%	333%	191%	116%	79%	54%	31%
DxCG	1489%	351%	239%	165%	118%	91%	73%	59%
HHS-HCC	7405%	526%	259%	148%	99%	79%	66%	57%
MARA	1438%	350%	232%	159%	116%	94%	78%	60%
Truven	751%	253%	200%	151%	117%	100%	83%	66%
Wakely	868%	359%	267%	184%	127%	92%	69%	49%
Pharmacy-Only Models								
DxCG	5459%	487%	265%	169%	121%	91%	72%	41%
MARA	4906%	467%	266%	174%	124%	92%	71%	40%
MedicaidRx	8205%	777%	373%	198%	114%	73%	50%	20%
Wakely	3277%	479%	297%	195%	134%	93%	66%	30%
Diagnosis-and-Pharmacy Models								
ACG System	778%	245%	189%	153%	130%	112%	88%	55%
CDPS-MRx	5301%	573%	321%	193%	121%	82%	59%	32%
CRG	2243%	524%	278%	170%	112%	80%	67%	54%
MARA	624%	290%	209%	153%	119%	99%	84%	63%
Wakely	1168%	383%	277%	190%	130%	93%	68%	43%

Table I.A.6: Predictive Ratios by Paid-to-Allowed (Concurrent; Offered Weights; No Censoring)

	Low	Medium	High
Diagnosis-Only Models			
ACG System	97.0%	98.5%	110.8%
CDPS	103.8%	95.7%	108.7%
DxCG	96.8%	96.1%	107.2%
HHS-HCC	103.3%	98.5%	103.3%
MARA	98.7%	98.7%	108.9%
Truven	95.6%	96.6%	104.6%
Wakely	95.8%	97.2%	108.9%
Pharmacy-Only Models			
DxCG	108.4%	94.3%	109.6%
MARA	109.5%	95.3%	110.0%
MedicaidRx	115.0%	95.7%	110.3%
Wakely	110.3%	96.0%	113.4%
Diagnosis-and-Pharmacy Models			
ACG System	99.5%	95.3%	112.5%
CDPS-MRx	104.1%	95.9%	109.9%
CRG	103.0%	98.9%	106.8%
MARA	101.0%	98.1%	109.8%
Wakely	100.0%	96.8%	110.7%

Table I.A.7: Predictive Ratios by Geographic Region (Concurrent; Offered Weights; No Censoring)

	Northeast	North Central	South	West
Diagnosis-Only Models				
ACG System	98.7%	101.2%	100.0%	96.9%
CDPS	100.7%	100.3%	98.9%	98.7%
DxCG	98.2%	101.5%	101.0%	95.6%
HHS-HCC	98.3%	101.3%	99.8%	98.6%
MARA	98.1%	101.2%	100.0%	98.1%
Truven	99.7%	101.4%	100.1%	95.9%
Wakely	103.5%	99.0%	99.7%	95.2%
Pharmacy-Only Models				
DxCG	90.2%	98.2%	107.5%	96.8%
MARA	89.2%	99.1%	107.9%	96.0%
MedicaidRx	88.6%	101.5%	105.4%	98.4%
Wakely	86.0%	98.4%	111.2%	93.6%
Diagnosis-and-Pharmacy Models				
ACG System	92.6%	101.0%	104.6%	95.8%
CDPS-MRx	98.8%	100.8%	100.5%	96.8%
CRG	95.6%	102.4%	101.6%	96.7%
MARA	95.6%	100.6%	102.5%	96.8%
Wakely	98.0%	98.5%	103.8%	94.2%

I.B Concurrent Models, $250,000 Censoring

Table I.B.1: Predictive Ratios by Health Conditions (Concurrent; Offered Weights; $250,000 Censoring)

	Heart Disease	Mental Illness	Diabetes	Low Back Pain	Asthma	Arthritis
Diagnosis-Only Models						
ACG System	110.4%	101.0%	105.6%	98.7%	96.6%	97.2%
CDPS	71.7%	86.7%	79.5%	69.4%	97.4%	67.1%
DxCG	101.9%	98.2%	100.1%	98.0%	90.0%	86.4%
HHS-HCC	101.7%	85.7%	104.4%	74.5%	95.6%	98.3%
MARA	94.4%	97.1%	95.3%	95.2%	92.6%	88.6%
Truven	101.6%	101.4%	98.7%	102.9%	93.6%	95.0%
Wakely	98.6%	98.3%	100.0%	94.7%	90.6%	90.5%
Pharmacy-Only Models						
DxCG	83.1%	86.0%	93.9%	87.2%	80.0%	88.6%
MARA	77.1%	86.5%	91.2%	86.1%	85.0%	89.1%
MedicaidRx	49.9%	79.0%	72.4%	64.7%	79.7%	66.5%
Wakely	70.5%	88.4%	92.7%	86.4%	79.4%	86.2%
Diagnosis-and-Pharmacy Models						
ACG System	101.2%	98.8%	98.5%	95.8%	92.9%	88.9%
CDPS-MRx	75.3%	93.4%	86.2%	72.2%	94.9%	73.0%
CRG	92.8%	95.5%	98.1%	86.0%	93.6%	98.3%
MARA	94.8%	97.5%	96.1%	95.7%	94.9%	90.9%
Wakely	90.6%	95.3%	98.1%	92.7%	87.1%	89.6%

Table I.B.2: Predictive Ratios by Age-Sex (Children; Concurrent; Offered Weights; $250,000 Censoring)

	Children, Age 0-1	Children, Age 2-6	Children, Age 7-18
Diagnosis-Only Models			
ACG System	102.6%	105.0%	90.0%
CDPS	270.0%	206.7%	216.9%
DxCG	81.5%	89.6%	89.4%
HHS-HCC	84.0%	89.1%	88.8%
MARA	110.1%	117.3%	104.5%
Truven	94.6%	93.9%	91.9%
Wakely	105.2%	109.6%	100.4%
Pharmacy-Only Models			
DxCG	84.0%	94.2%	91.6%
MARA	88.2%	109.8%	99.7%
MedicaidRx	352.3%	200.0%	220.3%
Wakely	87.3%	100.3%	101.9%
Diagnosis-and-Pharmacy Models			
ACG System	103.0%	109.9%	99.3%
CDPS-MRx	249.2%	190.7%	209.1%
CRG	121.9%	122.3%	109.1%
MARA	106.4%	115.8%	103.9%
Wakely	97.6%	104.3%	98.8%

Table I.B.3: Predictive Ratios by Age-Sex (Males; Concurrent; Offered Weights; $250,000 Censoring)

	Males, Age 19-22	Males, Age 23-24	Males, Age 25-29	Males, Age 30-34	Males, Age 35-39	Males, Age 40-44	Males, Age 45-49	Males, Age 50-54	Males, Age 55-59	Males, Age 60-64
Diagnosis-Only Models										
ACG System	96.3%	96.4%	95.0%	97.0%	101.7%	97.6%	101.7%	101.6%	102.3%	105.1%
CDPS	253.9%	178.4%	99.5%	97.5%	93.8%	86.0%	87.6%	82.2%	79.4%	76.8%
DxCG	97.1%	96.5%	99.0%	99.9%	104.1%	100.7%	102.9%	101.6%	103.8%	106.2%
HHS-HCC	104.1%	102.2%	96.4%	97.1%	100.5%	100.3%	101.4%	103.2%	106.8%	109.5%
MARA	103.1%	101.8%	100.4%	102.5%	101.4%	98.5%	98.5%	97.9%	98.7%	99.3%
Truven	97.4%	95.9%	98.4%	100.3%	101.3%	96.9%	98.2%	98.0%	100.1%	101.6%
Wakely	99.0%	100.8%	102.6%	103.7%	104.0%	98.2%	99.0%	99.1%	100.7%	100.5%
Pharmacy-Only Models										
DxCG	95.2%	96.3%	100.8%	107.1%	107.6%	100.4%	104.4%	101.6%	104.6%	106.5%
MARA	99.6%	96.5%	103.0%	110.3%	107.1%	102.2%	100.7%	100.3%	99.9%	98.9%
MedicaidRx	274.4%	185.4%	94.8%	99.5%	94.5%	85.1%	94.7%	85.2%	77.0%	68.8%
Wakely	103.1%	100.6%	103.6%	114.3%	109.8%	102.6%	101.7%	101.0%	100.0%	94.1%
Diagnosis-and-Pharmacy Models										
ACG System	97.5%	98.4%	100.2%	104.9%	105.2%	99.7%	100.1%	99.7%	97.9%	98.3%
CDPS-MRx	252.8%	178.0%	92.1%	94.4%	93.5%	88.1%	86.9%	83.3%	81.6%	79.9%
CRG	98.8%	96.9%	96.9%	98.6%	97.4%	91.1%	109.1%	104.5%	104.1%	102.1%
MARA	101.2%	101.1%	101.7%	104.7%	102.6%	99.6%	99.0%	98.2%	98.2%	99.2%
Wakely	100.5%	101.9%	105.2%	109.8%	108.3%	100.6%	98.6%	97.8%	98.3%	96.7%

Table I.B.4: Predictive Ratios by Age-Sex (Females; Concurrent; Offered Weights; $250,000 Censoring)

	Females, Age 19-22	Females, Age 23-24	Females, Age 25-29	Females, Age 30-34	Females, Age 35-39	Females, Age 40-44	Females, Age 45-49	Females, Age 50-54	Females, Age 55-59	Females, Age 60-64
Diagnosis-Only Models										
ACG System	93.9%	90.2%	90.8%	91.7%	95.6%	96.4%	102.9%	105.4%	103.1%	105.3%
CDPS	202.7%	133.0%	71.9%	71.9%	71.3%	70.4%	77.4%	76.2%	76.1%	76.8%
DxCG	95.7%	91.8%	91.1%	91.8%	96.7%	98.0%	101.3%	102.5%	102.7%	107.4%
HHS-HCC	94.6%	92.8%	90.8%	93.2%	95.9%	96.9%	99.3%	102.0%	102.7%	104.8%
MARA	103.3%	98.2%	94.4%	95.8%	97.3%	97.2%	98.8%	99.9%	99.9%	101.0%
Truven	101.2%	99.1%	103.8%	103.5%	100.3%	98.7%	103.0%	101.8%	101.6%	104.1%
Wakely	103.0%	98.2%	95.3%	97.1%	98.8%	98.1%	99.6%	100.9%	99.0%	100.4%
Pharmacy-Only Models										
DxCG	99.8%	99.2%	97.0%	98.5%	99.7%	100.5%	101.6%	98.5%	96.5%	100.5%
MARA	104.4%	102.2%	94.8%	99.9%	100.3%	101.0%	101.0%	99.8%	96.8%	97.0%
MedicaidRx	231.7%	161.4%	67.8%	71.0%	73.1%	75.3%	74.5%	72.8%	70.4%	68.9%
Wakely	103.9%	103.0%	96.9%	100.2%	100.4%	102.5%	101.7%	100.2%	97.0%	95.1%
Diagnosis-and-Pharmacy Models										
ACG System	102.3%	97.6%	96.9%	99.3%	100.8%	100.0%	102.4%	102.7%	98.2%	98.1%
CDPS-MRx	206.8%	136.9%	68.6%	71.0%	72.1%	73.4%	77.5%	77.6%	78.2%	79.5%
CRG	97.1%	92.5%	78.5%	81.0%	93.1%	101.8%	98.0%	98.9%	99.7%	98.6%
MARA	104.8%	98.5%	95.6%	98.0%	98.7%	98.4%	99.0%	99.7%	99.0%	99.3%
Wakely	105.4%	100.2%	95.1%	97.8%	99.7%	100.7%	102.0%	102.2%	99.6%	99.8%

Table I.B.5: Predictive Ratios by Cost Percentile (Concurrent; Offered Weights; $250,000 Censoring)

	0-20th Percentile	20-40th Percentile	40-60th Percentile	60-80th Percentile	80-90th Percentile	90-95th Percentile	95-98th Percentile	98th-99th Percentile
Diagnosis-Only Models								
ACG System	1371%	258%	189%	148%	119%	100%	81%	63%
CDPS	6559%	610%	320%	183%	112%	75%	52%	33%
DxCG	1449%	340%	232%	160%	115%	90%	73%	59%
HHS-HCC	7115%	505%	249%	142%	95%	76%	63%	61%
MARA	1382%	337%	223%	152%	111%	90%	75%	64%
Truven	793%	258%	203%	151%	116%	98%	79%	64%
Wakely	833%	345%	256%	177%	122%	89%	67%	52%
Pharmacy-Only Models								
DxCG	5245%	468%	255%	162%	116%	87%	69%	44%
MARA	4713%	449%	256%	167%	119%	88%	68%	43%
MedicaidRx	7883%	747%	358%	190%	110%	70%	48%	22%
Wakely	3148%	460%	285%	187%	129%	90%	64%	32%
Diagnosis-and-Pharmacy Models								
ACG System	747%	235%	182%	147%	125%	107%	85%	59%
CDPS-MRx	5093%	551%	308%	186%	116%	79%	56%	35%
CRG	2155%	504%	267%	163%	107%	77%	64%	58%
MARA	600%	279%	201%	147%	114%	95%	80%	67%
Wakely	1122%	368%	266%	182%	125%	89%	66%	46%

Table I.B.6: Predictive Ratios by Paid-to-Allowed (Concurrent; Offered Weights; $250,000 Censoring)

	Low	Medium	High
Diagnosis-Only Models			
ACG System	97.2%	99.0%	110.9%
CDPS	104.0%	96.1%	108.8%
DxCG	96.0%	97.3%	105.7%
HHS-HCC	103.5%	99.0%	103.4%
MARA	98.9%	99.2%	108.9%
Truven	96.0%	97.3%	105.7%
Wakely	96.0%	97.6%	108.9%
Pharmacy-Only Models			
DxCG	108.7%	94.7%	109.7%
MARA	109.7%	95.8%	110.0%
MedicaidRx	115.2%	96.1%	110.3%
Wakely	110.5%	96.4%	113.4%
Diagnosis-and-Pharmacy Models			
ACG System	99.7%	95.8%	112.6%
CDPS-MRx	104.3%	96.3%	110.0%
CRG	103.3%	99.4%	106.9%
MARA	101.2%	98.5%	109.9%
Wakely	100.2%	97.3%	110.8%

Table I.B.7: Predictive Ratios by Geographic Region (Concurrent; Offered Weights; $250,000 Censoring)

	Northeast	North Central	South	West
Diagnosis-Only Models				
ACG System	98.8%	101.2%	99.9%	97.4%
CDPS	100.8%	100.2%	98.8%	99.2%
DxCG	98.6%	101.3%	100.7%	96.2%
HHS-HCC	98.3%	101.3%	99.7%	99.1%
MARA	98.1%	101.1%	99.9%	98.6%
Truven	100.0%	101.3%	99.8%	96.5%
Wakely	103.6%	98.9%	99.6%	95.7%
Pharmacy-Only Models				
DxCG	90.3%	98.1%	107.3%	97.2%
MARA	89.2%	99.0%	107.7%	96.5%
MedicaidRx	88.7%	101.5%	105.3%	98.9%
Wakely	86.0%	98.3%	111.1%	94.1%
Diagnosis-and-Pharmacy Models				
ACG System	92.6%	100.9%	104.5%	96.3%
CDPS-MRx	98.9%	100.8%	100.3%	97.3%
CRG	95.7%	102.3%	101.5%	97.2%
MARA	95.6%	100.6%	102.4%	97.3%
Wakely	98.1%	98.4%	103.6%	94.7%

I.C Prospective Models, No Censoring

Table I.C.1: Predictive Ratios by Health Conditions (Prospective; Offered Weights; No Censoring)

	Heart Disease	Mental Illness	Diabetes	Low Back Pain	Asthma	Arthritis
Diagnosis-Only Models						
ACG System	58.9%	73.3%	85.1%	72.9%	72.4%	75.2%
CDPS	41.2%	66.1%	63.6%	55.3%	72.1%	44.0%
DxCG	63.4%	75.1%	86.5%	75.6%	72.8%	70.5%
Impact Pro	58.0%	73.7%	82.5%	73.7%	77.0%	70.8%
MARA	60.2%	75.6%	84.1%	76.3%	75.7%	73.7%
Truven	63.4%	77.6%	86.7%	77.6%	79.4%	79.4%
Wakely	58.9%	74.5%	85.2%	73.2%	71.7%	65.8%
Pharmacy-Only Models						
ACG System	53.2%	72.7%	80.4%	69.5%	71.2%	74.5%
DxCG	56.0%	72.9%	82.9%	70.6%	69.8%	79.2%
Impact Pro	53.8%	73.3%	82.0%	70.7%	71.8%	71.7%
MARA	55.5%	73.9%	81.8%	71.2%	73.1%	80.4%
MedicaidRx	41.4%	65.6%	72.3%	60.9%	63.5%	55.7%
Wakely	51.7%	73.0%	81.7%	69.6%	70.9%	69.0%
Diagnosis-and-Pharmacy Models						
ACG System	60.0%	77.2%	85.5%	75.9%	75.9%	76.4%
CDPS+MRX	44.3%	72.4%	71.0%	57.7%	71.8%	50.1%
CRG	59.5%	73.7%	86.2%	68.5%	72.4%	79.7%
Impact Pro	59.9%	78.3%	85.0%	76.4%	78.2%	74.7%
MARA	62.0%	78.1%	85.6%	77.9%	79.0%	81.3%
Wakely	61.4%	77.2%	87.1%	75.4%	76.1%	75.3%
Prior Cost Models						
ACG System	61.3%	76.8%	87.0%	75.6%	76.6%	81.3%
DxCG	70.2%	79.1%	88.3%	78.8%	78.3%	83.2%
MARA	65.9%	79.7%	88.2%	79.3%	82.0%	83.5%
SCIO	55.8%	80.3%	87.3%	81.3%	90.1%	83.0%

Table I.C.2: Predictive Ratios by Age-Sex (Children; Prospective; Offered Weights; No Censoring)

	Children, Age 0-1	Children, Age 2-6	Children, Age 7-18
Diagnosis-Only Models			
ACG System	86.6%	99.9%	98.4%
CDPS	178.8%	213.9%	225.2%
DxCG	77.6%	85.9%	87.8%
Impact Pro	123.5%	110.5%	107.5%
MARA	115.9%	100.1%	96.4%
Truven	90.0%	96.6%	100.2%
Wakely	106.1%	97.8%	100.0%
Pharmacy-Only Models			
ACG System	98.9%	105.4%	100.2%
DxCG	79.7%	84.9%	90.0%
Impact Pro	102.6%	109.7%	92.5%
MARA	101.7%	97.8%	95.1%
MedicaidRx	112.1%	150.9%	133.6%
Wakely	87.8%	92.8%	100.6%
Diagnosis-and-Pharmacy Models			
ACG System	90.2%	98.8%	97.0%
CDPS+MRX	168.9%	200.7%	218.3%
CRG	157.6%	130.2%	112.8%
Impact Pro	106.2%	96.1%	99.5%
MARA	113.2%	101.1%	97.9%
Wakely	103.2%	94.6%	99.8%
Prior Cost Models			
ACG System	94.1%	99.9%	97.5%
DxCG	114.0%	89.0%	88.6%
MARA	115.7%	100.2%	97.0%
SCIO	143.2%	113.5%	104.7%

Table I.C.3: Predictive Ratios by Age-Sex (Males; Prospective; Offered Weights; No Censoring)

	Males, Age 19-22	Males, Age 23-24	Males, Age 25-29	Males, Age 30-34	Males, Age 35-39	Males, Age 40-44	Males, Age 45-49	Males, Age 50-54	Males, Age 55-59	Males, Age 60-64
Diagnosis-Only Models										
ACG System	112.2%	121.2%	112.9%	111.4%	113.9%	99.6%	107.5%	98.2%	101.2%	93.8%
CDPS	240.0%	173.6%	111.8%	106.4%	94.7%	81.0%	91.6%	81.0%	71.9%	65.0%
DxCG	88.3%	100.8%	102.7%	103.3%	106.8%	98.6%	106.6%	100.9%	104.9%	104.4%
Impact Pro	110.0%	116.7%	105.2%	104.9%	99.4%	96.1%	96.3%	100.3%	98.0%	98.0%
MARA	109.4%	102.7%	103.6%	107.5%	105.0%	99.6%	100.0%	102.0%	103.3%	99.6%
Truven	102.6%	106.6%	102.3%	101.8%	102.1%	99.2%	102.9%	100.7%	100.8%	98.3%
Wakely	102.0%	107.7%	104.9%	107.1%	107.0%	101.1%	102.9%	100.9%	100.3%	94.4%
Pharmacy-Only Models										
ACG System	113.2%	120.0%	115.8%	119.2%	116.9%	99.8%	107.2%	96.0%	101.5%	88.2%
DxCG	93.3%	104.3%	103.6%	109.0%	110.0%	100.5%	107.1%	100.6%	104.6%	101.9%
Impact Pro	95.0%	100.9%	97.7%	101.9%	108.3%	94.1%	103.9%	92.1%	108.8%	94.1%
MARA	112.6%	105.8%	104.8%	113.6%	108.4%	101.7%	99.5%	100.5%	101.0%	97.9%
MedicaidRx	101.4%	108.7%	138.4%	137.9%	122.3%	103.9%	115.2%	99.8%	86.0%	73.4%
Wakely	104.6%	107.1%	104.4%	111.4%	107.5%	103.6%	103.1%	101.7%	101.7%	92.9%
Diagnosis-and-Pharmacy Models										
ACG System	109.0%	117.0%	111.6%	114.8%	113.3%	100.0%	105.4%	98.3%	99.1%	92.1%
CDPS+MRX	239.2%	173.6%	103.7%	101.8%	93.5%	82.2%	90.7%	82.0%	74.6%	68.5%
CRG	84.2%	80.4%	75.5%	76.5%	73.8%	67.1%	110.2%	102.1%	95.2%	89.8%
Impact Pro	102.1%	109.1%	101.8%	106.2%	102.4%	98.5%	98.7%	101.3%	98.7%	97.1%
MARA	106.1%	102.2%	103.1%	110.5%	106.5%	100.0%	100.4%	101.5%	102.5%	98.8%
Wakely	102.7%	107.4%	104.3%	107.9%	106.4%	100.9%	102.9%	100.6%	100.3%	96.1%
Prior Cost Models										
ACG System	106.6%	115.0%	109.0%	111.7%	112.0%	99.5%	105.5%	98.7%	100.6%	94.1%
DxCG	99.1%	101.5%	100.8%	97.8%	103.5%	95.3%	106.9%	99.3%	103.4%	105.9%
MARA	108.6%	109.9%	102.9%	109.1%	106.0%	99.8%	100.5%	101.2%	102.8%	100.0%
SCIO	91.2%	96.4%	95.7%	106.2%	104.6%	97.9%	93.2%	90.5%	85.6%	80.2%

Table I.C.4: Predictive Ratios by Age-Sex (Females; Prospective; Offered Weights; No Censoring)

	Females, Age 19-22	Females, Age 23-24	Females, Age 25-29	Females, Age 30-34	Females, Age 35-39	Females, Age 40-44	Females, Age 45-49	Females, Age 50-54	Females, Age 55-59	Females, Age 60-64
Diagnosis-Only Models										
ACG System	109.3%	98.4%	81.3%	80.7%	91.3%	92.9%	104.2%	100.2%	106.5%	105.0%
CDPS	226.4%	153.5%	76.1%	74.2%	75.0%	74.3%	82.4%	77.4%	73.3%	72.8%
DxCG	98.6%	102.9%	95.8%	94.6%	97.6%	98.6%	102.2%	100.1%	101.1%	107.0%
Impact Pro	110.5%	102.0%	104.4%	107.1%	104.2%	98.0%	98.5%	96.1%	91.8%	96.9%
MARA	95.6%	104.9%	96.6%	97.8%	100.7%	98.3%	101.0%	99.2%	96.6%	100.8%
Truven	103.3%	112.5%	97.3%	97.7%	99.5%	99.4%	101.6%	99.0%	97.8%	100.9%
Wakely	105.8%	110.9%	103.0%	99.6%	99.0%	99.0%	101.1%	98.9%	97.2%	99.0%
Pharmacy-Only Models										
ACG System	108.7%	99.6%	82.2%	85.0%	95.1%	94.6%	104.5%	98.3%	106.1%	101.8%
DxCG	104.5%	111.1%	101.9%	100.2%	101.4%	102.2%	103.3%	97.6%	96.0%	100.0%
Impact Pro	124.8%	114.4%	92.9%	94.9%	101.5%	100.8%	103.8%	96.8%	102.8%	99.3%
MARA	99.4%	114.8%	101.1%	104.7%	103.6%	102.1%	101.7%	98.7%	93.7%	97.4%
MedicaidRx	154.9%	138.0%	91.0%	91.5%	94.8%	95.4%	98.5%	92.0%	85.4%	82.7%
Wakely	106.8%	114.3%	104.2%	99.1%	97.8%	100.0%	101.7%	99.2%	96.6%	97.1%
Diagnosis-and-Pharmacy Models										
ACG System	107.1%	99.9%	85.9%	88.5%	97.5%	97.4%	104.7%	101.2%	103.9%	102.2%
CDPS+MRX	229.8%	156.1%	72.4%	72.7%	75.2%	76.1%	81.9%	78.7%	75.4%	75.9%
CRG	131.4%	133.6%	114.6%	112.2%	111.2%	109.0%	97.6%	95.4%	93.1%	94.2%
Impact Pro	106.6%	100.5%	102.6%	108.6%	106.2%	100.9%	100.5%	98.8%	94.8%	98.8%
MARA	97.4%	106.9%	97.6%	100.0%	101.7%	100.0%	101.5%	98.9%	95.1%	98.3%
Wakely	105.4%	110.9%	103.2%	99.9%	99.2%	98.9%	100.4%	98.3%	96.8%	100.3%
Prior Cost Models										
ACG System	106.5%	99.6%	86.7%	88.5%	95.4%	94.6%	102.9%	100.4%	104.5%	103.7%
DxCG	102.3%	96.2%	102.9%	102.9%	102.3%	99.7%	100.7%	98.3%	96.8%	103.8%
MARA	98.1%	106.6%	98.5%	100.2%	101.8%	99.4%	100.8%	98.8%	94.8%	98.4%
SCIO	118.8%	114.3%	102.1%	109.1%	111.1%	110.7%	105.7%	104.3%	99.8%	100.2%

Table I.C.5: Predictive Ratios by Cost Percentile (Prospective; Offered Weights; No Censoring)

	0-20th Percentile	20-40th Percentile	40-60th Percentile	60-80th Percentile	80-90th Percentile	90-95th Percentile	95-98th Percentile	98th-99th Percentile
Diagnosis-Only Models								
ACG System	9897%	755%	358%	193%	112%	70%	48%	24%
CDPS	13985%	990%	405%	191%	98%	58%	36%	16%
DxCG	9396%	713%	348%	193%	114%	72%	49%	26%
Impact Pro	9642%	744%	351%	187%	109%	70%	48%	29%
MARA	10013%	736%	346%	186%	110%	71%	50%	27%
Truven	8912%	670%	324%	184%	113%	75%	54%	31%
Wakely	9322%	748%	363%	195%	112%	69%	46%	25%
Pharmacy-Only Models								
ACG System	10703%	770%	359%	194%	115%	72%	48%	19%
DxCG	10773%	751%	352%	190%	112%	70%	50%	23%
Impact Pro	10451%	768%	361%	195%	113%	70%	47%	21%
MARA	10212%	738%	349%	191%	113%	72%	49%	24%
MedicaidRx	13176%	956%	408%	199%	105%	60%	37%	12%
Wakely	9965%	763%	369%	202%	117%	71%	46%	18%
Diagnosis-and-Pharmacy Models								
ACG System	8417%	680%	338%	192%	118%	77%	53%	26%
CDPS+MRX	12375%	920%	393%	195%	104%	62%	39%	17%
CRG	8413%	775%	366%	192%	110%	70%	48%	26%
Impact Pro	8420%	686%	340%	189%	114%	74%	52%	29%
MARA	8962%	671%	324%	183%	114%	76%	55%	30%
Wakely	8105%	674%	341%	194%	117%	74%	51%	27%
Prior Cost Models								
ACG System	8328%	672%	335%	191%	118%	77%	54%	26%
DxCG	7920%	599%	299%	174%	111%	76%	58%	40%
MARA	8626%	645%	311%	177%	112%	76%	57%	35%
SCIO	6268%	596%	326%	199%	127%	84%	58%	24%

Table I.C.6: Predictive Ratios by Paid-to-Allowed (Prospective; Offered Weights; No Censoring)

	Low	Medium	High
Diagnosis-Only Models			
ACG System	98.21%	97.21%	98.30%
CDPS	108.49%	94.74%	100.82%
DxCG	97.09%	96.38%	97.10%
Impact Pro	102.99%	95.97%	101.88%
MARA	101.13%	97.72%	100.27%
Truven	100.85%	98.53%	99.70%
Wakely	97.16%	95.52%	96.61%
Pharmacy-Only Models			
ACG System	110.26%	98.61%	111.72%
DxCG	110.02%	97.52%	108.66%
Impact Pro	110.42%	98.12%	109.46%
MARA	110.45%	97.93%	108.99%
MedicaidRx	115.57%	96.79%	108.14%
Wakely	111.22%	97.56%	111.68%
Diagnosis-and-Pharmacy Models			
ACG System	101.08%	96.70%	107.72%
CDPS+MRX	108.32%	95.57%	103.55%
CRG	105.48%	97.18%	103.44%
Impact Pro	104.00%	97.49%	105.09%
MARA	104.28%	98.24%	105.90%
Wakely	101.05%	96.70%	103.38%
Prior Cost Models			
ACG System	101.64%	98.17%	105.04%
DxCG	95.13%	97.58%	101.49%
MARA	103.20%	98.49%	105.03%
SCIO	98.38%	101.69%	113.72%

Table I.C.7: Predictive Ratios by Geographic Region (Prospective; Offered Weights; No Censoring)

	Northeast	North Central	South	West
Diagnosis-Only Models				
ACG System	98.5%	99.8%	99.2%	101.7%
CDPS	97.7%	101.0%	98.3%	103.3%
DxCG	99.1%	99.9%	99.3%	100.8%
Impact Pro	97.5%	99.7%	100.2%	101.2%
MARA	97.8%	99.9%	99.8%	101.3%
Truven	99.7%	100.8%	98.5%	100.8%
Wakely	101.1%	99.2%	98.8%	100.3%
Pharmacy-Only Models				
ACG System	91.4%	98.8%	106.6%	98.9%
DxCG	91.8%	99.4%	105.9%	99.3%
Impact Pro	90.4%	99.6%	106.7%	98.7%
MARA	91.3%	99.4%	106.7%	98.2%
MedicaidRx	90.8%	100.3%	104.8%	101.0%
Wakely	90.3%	99.2%	107.7%	97.4%
Diagnosis-and-Pharmacy Models				
ACG System	93.5%	99.6%	104.5%	98.7%
CDPS+MRX	96.6%	101.5%	99.6%	101.5%
CRG	95.4%	101.2%	101.0%	100.5%
Impact Pro	96.5%	99.1%	101.8%	100.8%
MARA	95.6%	99.5%	103.1%	99.2%
Wakely	97.6%	99.1%	102.6%	98.2%
Prior Cost Models				
ACG System	97.0%	99.0%	102.5%	99.3%
DxCG	98.8%	99.8%	100.0%	101.6%
MARA	95.8%	99.3%	103.0%	99.9%
SCIO	102.4%	97.0%	102.5%	97.4%

I.D Prospective Models, $250,000 Censoring

Table I.D.1: Predictive Ratios by Health Conditions (Prospective; Offered Weights; $250,000 Censoring)

	Heart Disease	Mental Illness	Diabetes	Low Back Pain	Asthma	Arthritis
Diagnosis-Only Models						
ACG System	62.4%	73.6%	86.9%	72.8%	71.5%	75.9%
CDPS	43.6%	66.3%	64.9%	55.2%	71.2%	44.4%
DxCG	65.5%	75.6%	87.8%	76.1%	72.2%	72.1%
Impact Pro	61.4%	73.9%	84.3%	73.6%	76.1%	71.5%
MARA	63.3%	75.7%	85.5%	76.0%	74.5%	74.3%
Truven	63.7%	77.5%	86.9%	77.3%	77.3%	80.5%
Wakely	62.4%	74.7%	87.0%	73.2%	70.8%	66.4%
Pharmacy-Only Models						
ACG System	56.3%	72.9%	82.1%	69.5%	70.3%	75.2%
DxCG	59.3%	73.1%	84.7%	70.5%	68.9%	80.0%
Impact Pro	56.9%	73.5%	83.7%	70.6%	70.9%	72.4%
MARA	58.8%	74.2%	83.5%	71.1%	72.2%	81.2%
MedicaidRx	43.8%	65.8%	73.8%	60.8%	62.7%	56.2%
Wakely	54.7%	73.2%	83.4%	69.5%	70.0%	69.7%
Diagnosis-and-Pharmacy Models						
ACG System	63.5%	77.4%	87.3%	75.8%	74.9%	77.1%
CDPS+MRX	46.9%	72.6%	72.5%	57.6%	70.9%	50.6%
CRG	63.0%	74.0%	88.1%	68.4%	71.5%	80.4%
Impact Pro	63.4%	78.5%	86.8%	76.3%	77.2%	75.4%
MARA	64.8%	77.9%	86.9%	77.5%	77.5%	82.0%
Wakely	65.0%	77.5%	88.9%	75.3%	75.2%	76.0%
Prior Cost Models						
ACG System	64.9%	77.0%	88.9%	75.5%	75.6%	82.1%
DxCG	71.6%	79.6%	90.2%	78.8%	77.8%	86.8%
MARA	69.7%	79.9%	90.0%	79.2%	81.0%	84.3%
SCIO	59.1%	80.6%	89.1%	81.2%	89.0%	83.8%

Table I.D.2: Predictive Ratios by Age-Sex (Children; Prospective; Offered Weights; $250,000 Censoring)

	Children, Age 0-1	Children, Age 2-6	Children, Age 7-18
Diagnosis-Only Models			
ACG System	88.9%	102.5%	98.7%
CDPS	183.7%	219.5%	226.0%
DxCG	79.8%	88.9%	88.7%
Impact Pro	126.8%	113.3%	107.9%
MARA	119.1%	102.7%	96.7%
Truven	93.2%	97.0%	99.2%
Wakely	108.9%	100.3%	100.3%
Pharmacy-Only Models			
ACG System	101.6%	108.1%	100.6%
DxCG	81.9%	87.1%	90.3%
Impact Pro	105.4%	112.6%	92.9%
MARA	104.5%	100.3%	95.4%
MedicaidRx	115.1%	154.8%	134.0%
Wakely	90.1%	95.2%	101.0%
Diagnosis-and-Pharmacy Models			
ACG System	92.7%	101.4%	97.3%
CDPS+MRX	173.4%	205.9%	219.0%
CRG	161.8%	133.6%	113.2%
Impact Pro	109.1%	98.6%	99.9%
MARA	116.3%	103.8%	98.2%
Wakely	106.0%	97.1%	100.1%
Prior Cost Models			
ACG System	96.6%	102.5%	97.9%
DxCG	116.0%	90.1%	88.4%
MARA	118.8%	102.8%	97.3%
SCIO	147.1%	116.5%	105.1%

Table I.D.3: Predictive Ratios by Age-Sex (Males; Prospective; Offered Weights; $250,000 Censoring)

	Males, Age 19-22	Males, Age 23-24	Males, Age 25-29	Males, Age 30-34	Males, Age 35-39	Males, Age 40-44	Males, Age 45-49	Males, Age 50-54	Males, Age 55-59	Males, Age 60-64
Diagnosis-Only Models										
ACG System	112.1%	117.4%	112.2%	109.7%	114.1%	100.3%	109.2%	98.5%	102.0%	96.1%
CDPS	239.9%	168.1%	111.1%	104.8%	94.8%	81.6%	93.1%	81.2%	72.4%	66.6%
DxCG	87.0%	96.5%	102.1%	102.0%	107.3%	99.4%	107.9%	100.6%	104.5%	104.9%
Impact Pro	109.9%	113.0%	104.6%	103.2%	99.5%	96.8%	97.8%	100.6%	98.8%	100.4%
MARA	109.3%	99.4%	103.0%	105.8%	105.1%	100.3%	101.7%	102.3%	104.1%	102.1%
Truven	100.8%	102.4%	100.1%	101.2%	102.7%	99.8%	103.3%	100.7%	100.3%	98.5%
Wakely	102.0%	104.3%	104.3%	105.4%	107.1%	101.8%	104.6%	101.2%	101.1%	96.8%
Pharmacy-Only Models										
ACG System	113.2%	116.2%	115.1%	117.3%	117.0%	100.6%	108.9%	96.3%	102.3%	90.4%
DxCG	93.2%	101.0%	103.0%	107.3%	110.1%	101.2%	108.9%	100.9%	105.4%	104.4%
Impact Pro	95.0%	97.7%	97.1%	100.3%	108.4%	94.8%	105.5%	92.4%	109.6%	96.4%
MARA	112.5%	102.4%	104.2%	111.9%	108.5%	102.4%	101.1%	100.8%	101.8%	100.3%
MedicaidRx	101.4%	105.3%	137.6%	135.7%	122.4%	104.7%	117.1%	100.2%	86.7%	75.2%
Wakely	104.5%	103.7%	103.8%	109.7%	107.7%	104.4%	104.8%	102.1%	102.5%	95.3%
Diagnosis-and-Pharmacy Models										
ACG System	108.9%	113.2%	111.0%	113.0%	113.4%	100.7%	107.1%	98.6%	99.9%	94.3%
CDPS+MRX	239.1%	168.1%	103.1%	100.2%	93.6%	82.8%	92.1%	82.3%	75.2%	70.2%
CRG	84.1%	77.9%	75.0%	75.3%	73.8%	67.6%	112.0%	102.4%	96.0%	92.0%
Impact Pro	102.0%	105.6%	101.2%	104.6%	102.5%	99.3%	100.3%	101.6%	99.5%	99.6%
MARA	106.0%	98.9%	102.5%	108.8%	106.6%	100.7%	102.0%	101.8%	103.4%	101.3%
Wakely	102.7%	104.0%	103.7%	106.2%	106.5%	101.6%	104.5%	101.0%	101.1%	98.5%
Prior Cost Models										
ACG System	106.5%	111.4%	108.4%	109.9%	112.1%	100.3%	107.2%	99.1%	101.4%	96.4%
DxCG	96.0%	95.8%	100.3%	97.4%	104.4%	95.7%	108.0%	99.0%	102.8%	106.0%
MARA	108.5%	106.4%	102.3%	107.3%	106.1%	100.6%	102.1%	101.5%	103.7%	102.5%
SCIO	91.1%	93.3%	95.2%	104.5%	104.7%	98.6%	94.7%	90.8%	86.3%	82.2%

Table I.D.4: Predictive Ratios by Age-Sex (Females; Prospective; Offered Weights; $250,000 Censoring)

	Females, Age 19-22	Females, Age 23-24	Females, Age 25-29	Females, Age 30-34	Females, Age 35-39	Females, Age 40-44	Females, Age 45-49	Females, Age 50-54	Females, Age 55-59	Females, Age 60-64
Diagnosis-Only Models										
ACG System	107.1%	95.6%	79.4%	79.3%	89.6%	91.2%	103.8%	100.1%	106.5%	105.0%
CDPS	221.9%	149.1%	74.3%	72.9%	73.6%	73.0%	82.2%	77.3%	73.3%	72.8%
DxCG	97.8%	100.7%	94.2%	93.9%	96.7%	97.9%	103.0%	100.9%	101.2%	106.0%
Impact Pro	108.3%	99.1%	101.9%	105.2%	102.2%	96.3%	98.1%	96.0%	91.8%	96.8%
MARA	93.7%	101.9%	94.3%	96.1%	98.8%	96.6%	100.7%	99.1%	96.6%	100.7%
Truven	102.1%	110.8%	96.5%	98.2%	98.7%	99.1%	102.4%	99.7%	98.7%	100.8%
Wakely	103.7%	107.7%	100.6%	97.9%	97.2%	97.3%	100.7%	98.8%	97.2%	98.9%
Pharmacy-Only Models										
ACG System	106.5%	96.8%	80.2%	83.5%	93.3%	92.9%	104.2%	98.2%	106.1%	101.7%
DxCG	102.5%	107.9%	99.5%	98.4%	99.5%	100.4%	102.9%	97.5%	96.0%	100.0%
Impact Pro	122.3%	111.1%	90.7%	93.3%	99.6%	99.0%	103.4%	96.7%	102.8%	99.3%
MARA	97.4%	111.5%	98.7%	102.9%	101.7%	100.3%	101.3%	98.6%	93.8%	97.4%
MedicaidRx	151.9%	134.1%	88.9%	89.9%	93.0%	93.7%	98.2%	91.9%	85.4%	82.6%
Wakely	104.7%	111.1%	101.8%	97.4%	96.0%	98.2%	101.3%	99.1%	96.6%	97.1%
Diagnosis-and-Pharmacy Models										
ACG System	104.9%	97.1%	83.8%	87.0%	95.7%	95.7%	104.3%	101.1%	104.0%	102.1%
CDPS+MRX	225.2%	151.7%	70.7%	71.5%	73.8%	74.8%	81.6%	78.6%	75.4%	75.9%
CRG	128.8%	129.9%	111.9%	110.3%	109.2%	107.1%	97.2%	95.3%	93.1%	94.2%
Impact Pro	104.4%	97.6%	100.2%	106.7%	104.2%	99.1%	100.2%	98.7%	94.8%	98.7%
MARA	95.5%	103.9%	95.3%	98.3%	99.8%	98.2%	101.1%	98.8%	95.2%	98.3%
Wakely	103.3%	107.7%	100.7%	98.2%	97.3%	97.2%	100.0%	98.2%	96.8%	100.2%
Prior Cost Models										
ACG System	104.4%	96.8%	84.7%	86.9%	93.6%	92.9%	102.6%	100.3%	104.5%	103.6%
DxCG	101.8%	95.1%	101.2%	102.4%	101.1%	99.4%	101.5%	99.4%	97.5%	104.0%
MARA	96.2%	103.6%	96.2%	98.5%	99.9%	97.7%	100.5%	98.7%	94.8%	98.3%
SCIO	116.4%	111.1%	99.7%	107.2%	109.1%	108.7%	105.3%	104.1%	99.8%	100.1%

Table I.D.5: Predictive Ratios by Cost Percentile (Prospective; Offered Weights; $250,000 Censoring)

	0-20th Percentile	20-40th Percentile	40-60th Percentile	60-80th Percentile	80-90th Percentile	90-95th Percentile	95-98th Percentile	98th-99th Percentile
Diagnosis-Only Models								
ACG System	9508%	725%	344%	185%	108%	67%	46%	25%
CDPS	13436%	951%	389%	183%	94%	56%	35%	17%
DxCG	9040%	691%	337%	186%	110%	70%	47%	26%
Impact Pro	9264%	715%	337%	180%	105%	67%	46%	31%
MARA	9620%	707%	332%	178%	105%	69%	48%	29%
Truven	8657%	659%	321%	180%	110%	72%	50%	30%
Wakely	8956%	719%	349%	187%	108%	66%	45%	27%
Pharmacy-Only Models								
ACG System	10282%	739%	345%	186%	110%	69%	46%	20%
DxCG	10350%	722%	339%	182%	108%	68%	48%	24%
Impact Pro	10041%	737%	347%	187%	109%	68%	45%	22%
MARA	9812%	709%	335%	183%	109%	69%	47%	25%
MedicaidRx	12659%	918%	392%	191%	101%	58%	35%	13%
Wakely	9574%	733%	354%	194%	112%	68%	44%	19%
Diagnosis-and-Pharmacy Models								
ACG System	8087%	653%	325%	184%	113%	74%	51%	27%
CDPS+MRX	11889%	884%	378%	187%	100%	60%	38%	18%
CRG	8083%	745%	352%	184%	106%	67%	46%	28%
Impact Pro	8089%	659%	326%	182%	109%	71%	50%	31%
MARA	8564%	641%	310%	175%	109%	73%	53%	33%
Wakely	7787%	647%	327%	186%	112%	71%	49%	29%
Prior Cost Models								
ACG System	8002%	645%	322%	184%	114%	74%	52%	28%
DxCG	7638%	584%	293%	171%	109%	74%	57%	39%
MARA	8287%	619%	299%	170%	107%	73%	54%	38%
SCIO	6022%	573%	314%	191%	122%	80%	56%	26%

Table I.D.6: Predictive Ratios by Paid-to-Allowed (Prospective; Offered Weights; $250,000 Censoring)

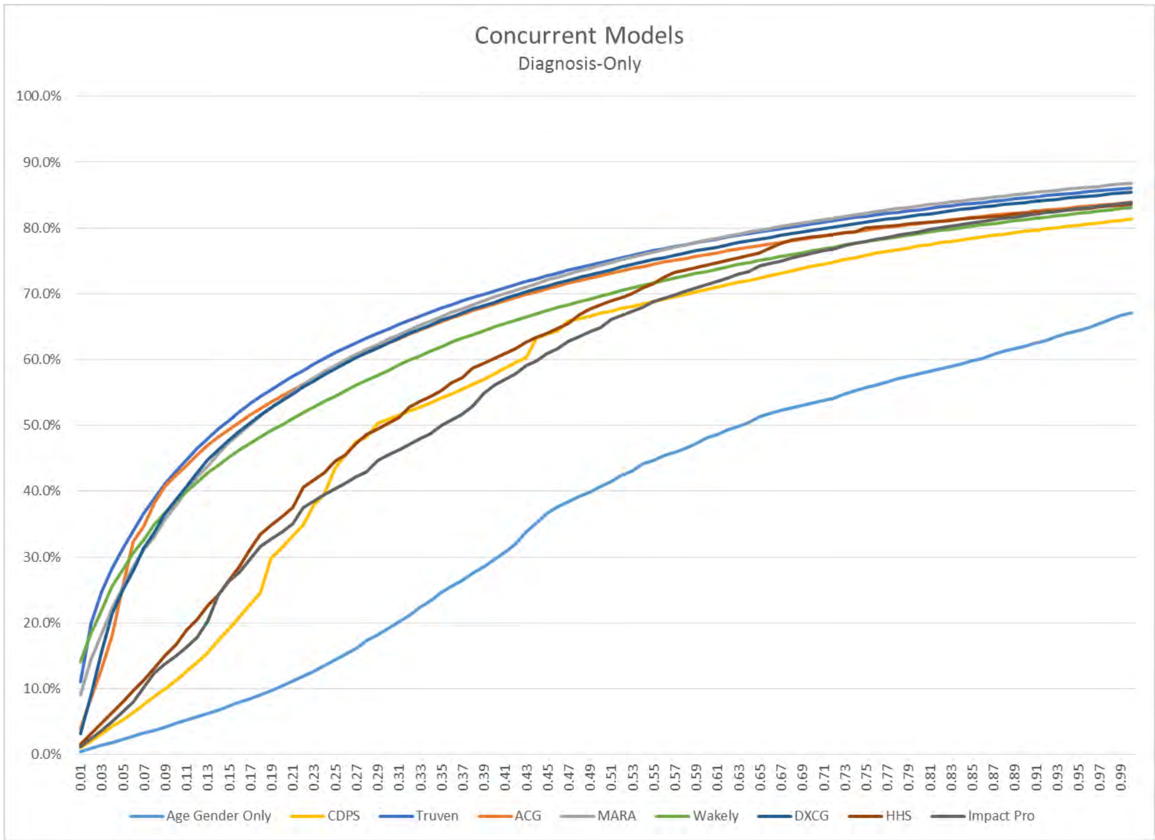
	Low	Medium	High
Diagnosis-Only Models			
ACG System	98.41%	97.66%	98.36%
CDPS	108.72%	95.18%	100.89%
DxCG	97.41%	96.72%	97.32%
Impact Pro	103.20%	96.41%	101.95%
MARA	101.35%	98.17%	100.33%
Truven	101.29%	98.83%	99.94%
Wakely	97.36%	95.96%	96.66%
Pharmacy-Only Models			
ACG System	110.50%	99.06%	111.79%
DxCG	110.26%	97.97%	108.72%
Impact Pro	110.66%	98.57%	109.53%
MARA	110.69%	98.39%	109.05%
MedicaidRx	115.82%	97.24%	108.21%
Wakely	111.45%	98.01%	111.75%
Diagnosis-and-Pharmacy Models			
ACG System	101.30%	97.14%	107.79%
CDPS+MRX	108.55%	96.01%	103.61%
CRG	105.70%	97.63%	103.50%
Impact Pro	104.23%	97.94%	105.16%
MARA	104.51%	98.70%	105.96%
Wakely	101.26%	97.15%	103.45%
Prior Cost Models			
ACG System	101.86%	98.63%	105.10%
DxCG	96.22%	97.75%	102.00%
MARA	103.42%	98.95%	105.10%
SCIO	98.58%	102.16%	113.79%

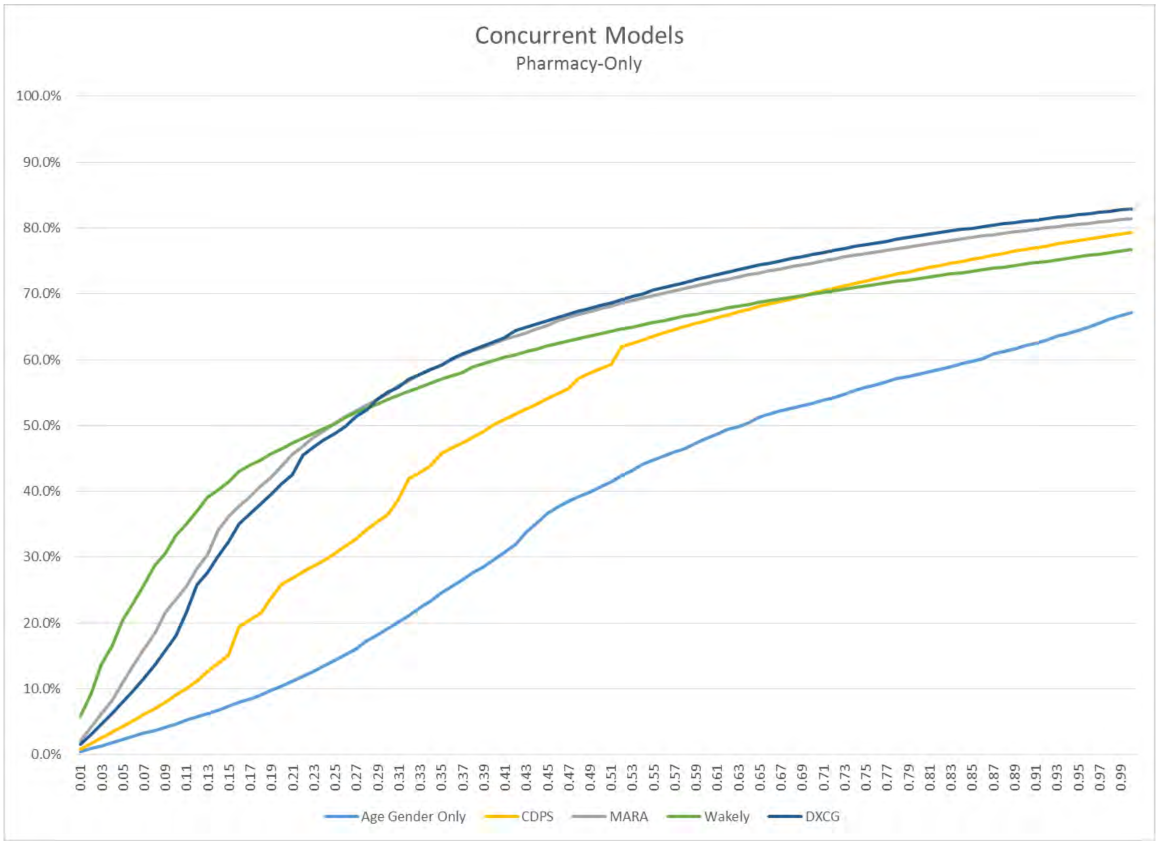
Table I.D.7: Predictive Ratios by Geographic Region (Prospective; Offered Weights; $250,000 Censoring)

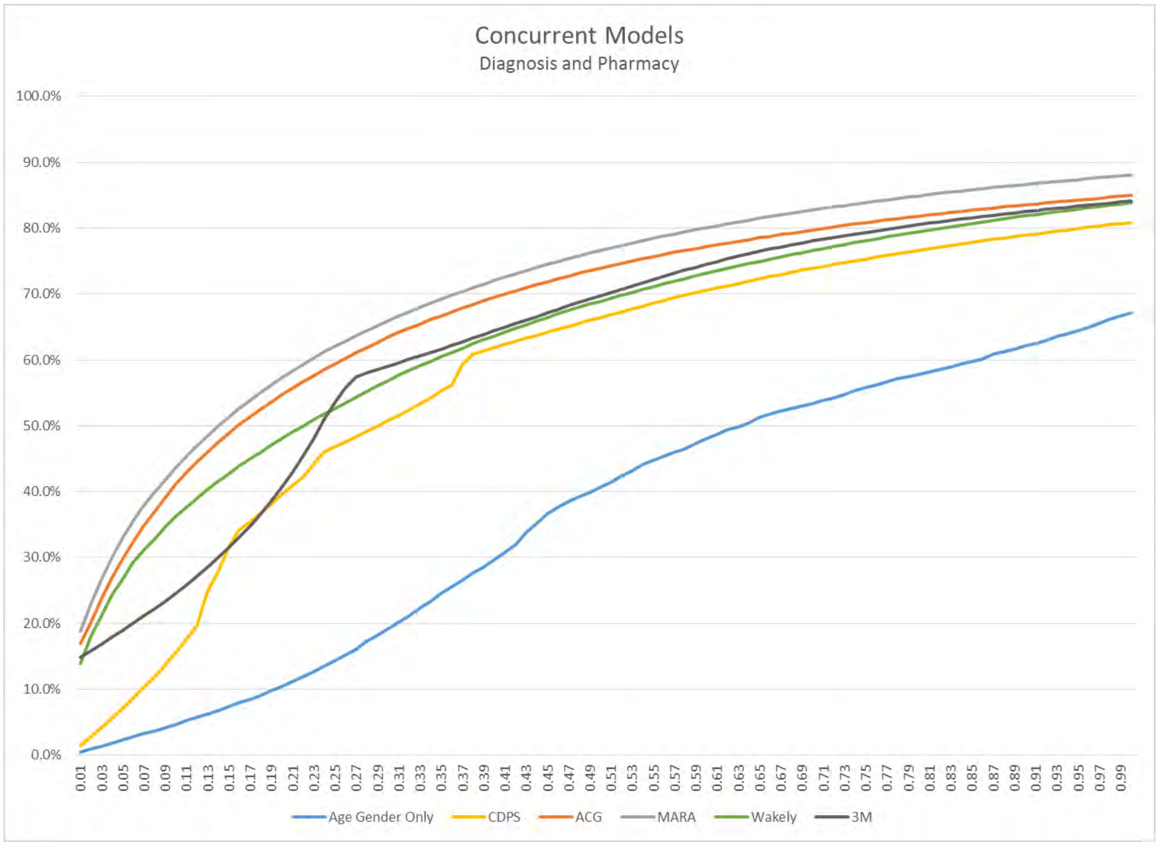
	Northeast	North Central	South	West
Diagnosis-Only Models				
ACG System	98.5%	99.7%	99.1%	102.2%
CDPS	97.7%	100.9%	98.2%	103.8%
DxCG	99.3%	99.8%	99.1%	101.3%
Impact Pro	97.6%	99.6%	100.1%	101.7%
MARA	97.8%	99.8%	99.7%	101.9%
Truven	99.8%	100.6%	98.5%	101.3%
Wakely	101.1%	99.1%	98.7%	100.8%
Pharmacy-Only Models				
ACG System	91.5%	98.7%	106.4%	99.4%
DxCG	91.8%	99.3%	105.7%	99.8%
Impact Pro	90.4%	99.6%	106.6%	99.2%
MARA	91.4%	99.3%	106.6%	98.7%
MedicaidRx	90.9%	100.3%	104.7%	101.5%
Wakely	90.4%	99.2%	107.6%	97.9%
Diagnosis-and-Pharmacy Models				
ACG System	93.6%	99.5%	104.4%	99.2%
CDPS+MRX	96.7%	101.4%	99.5%	102.1%
CRG	95.5%	101.1%	100.9%	101.0%
Impact Pro	96.6%	99.1%	101.7%	101.3%
MARA	95.7%	99.4%	103.0%	99.9%
Wakely	97.6%	99.0%	102.5%	98.7%
Prior Cost Models				
ACG System	97.1%	98.9%	102.4%	99.8%
DxCG	98.6%	99.7%	100.2%	101.5%
MARA	95.9%	99.3%	102.8%	100.4%
SCIO	102.5%	97.0%	102.4%	97.9%

Appendix II Tolerance Curves

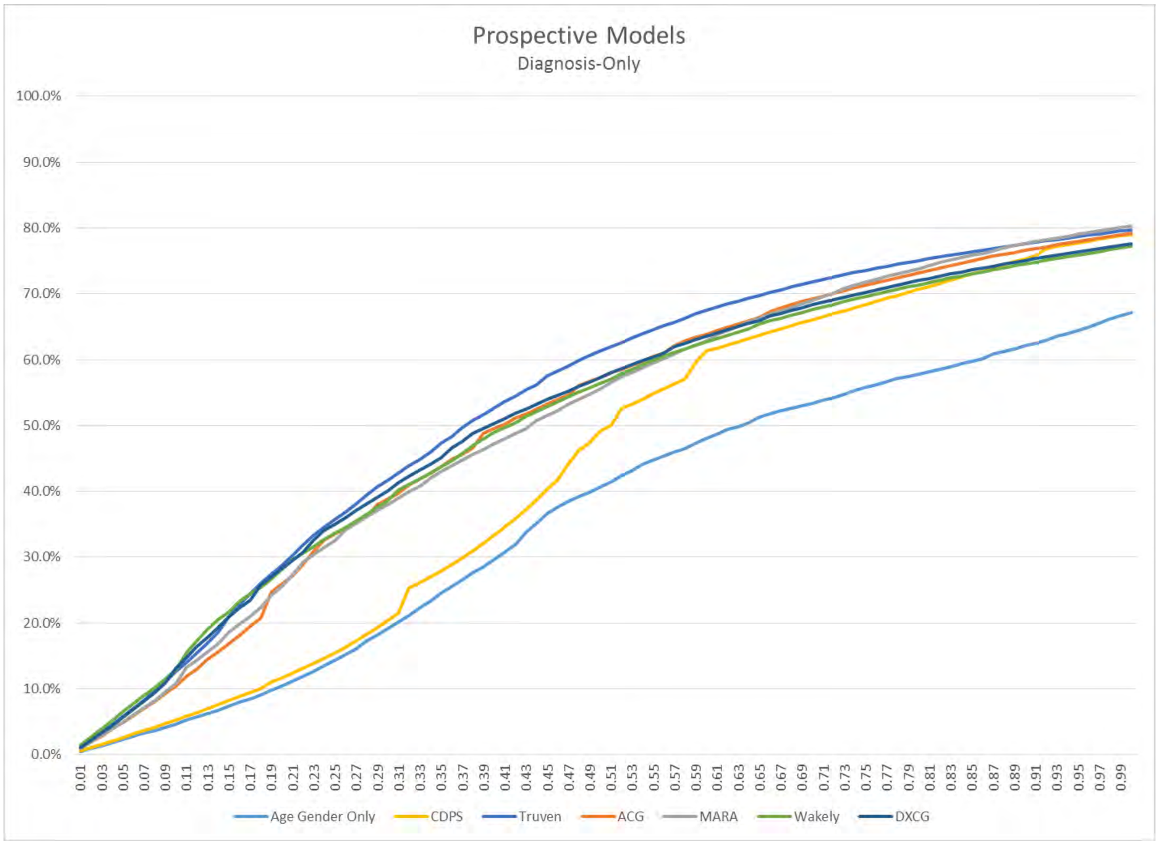
II.A Tolerance Curves, Concurrent Models

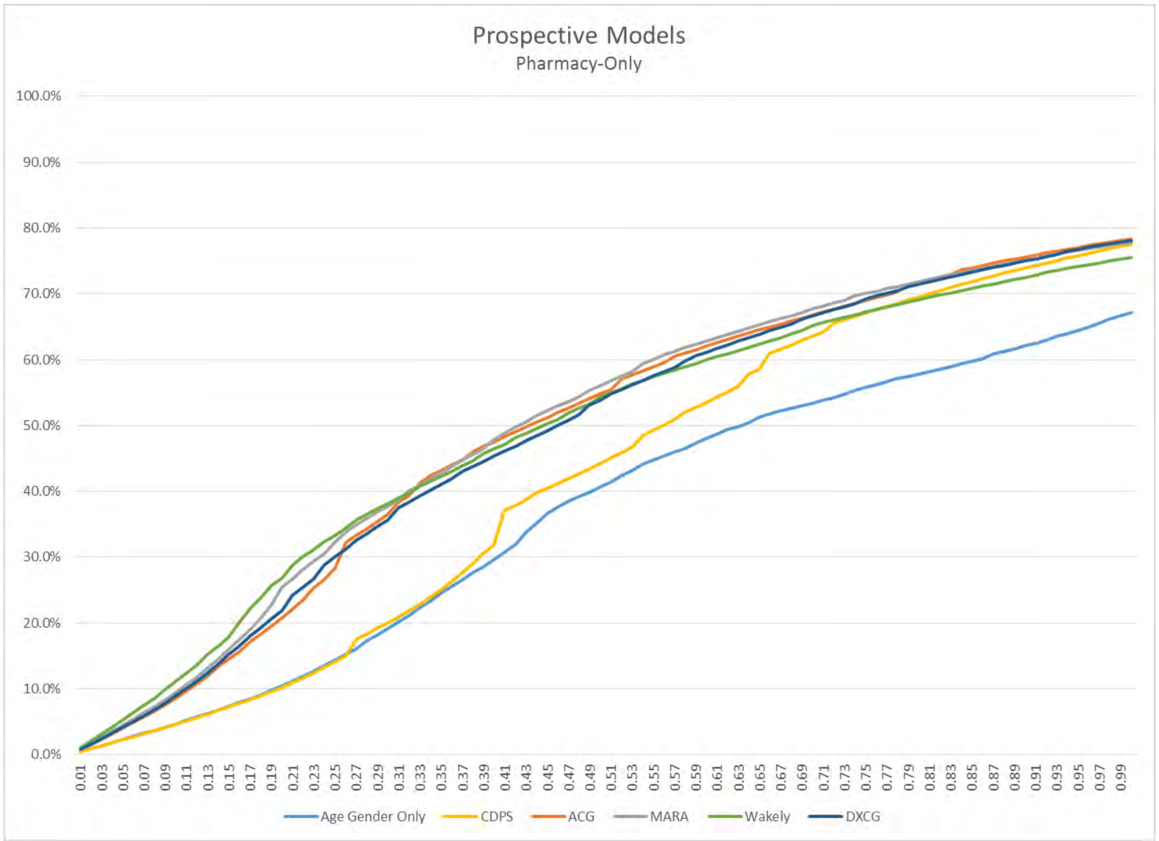


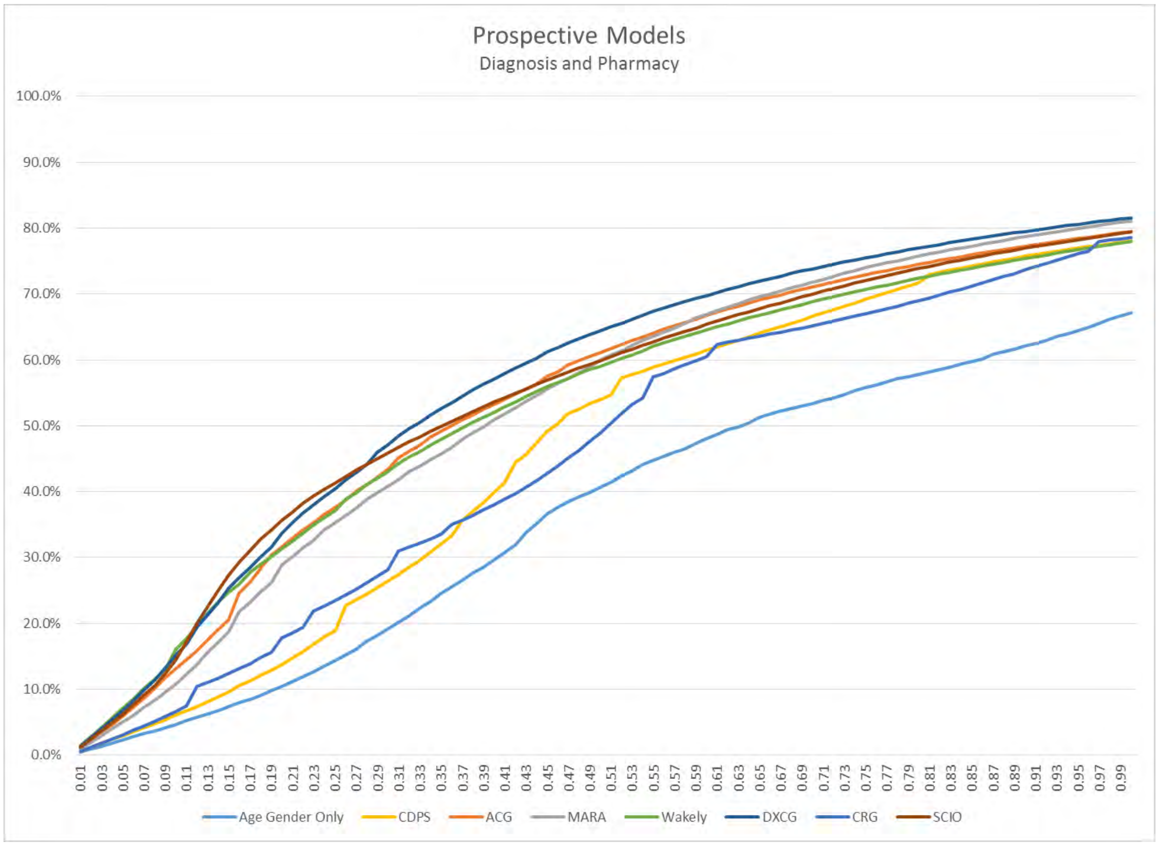

Concurrent Models
Pharmacy-Only
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0.01 0.03 0.05 0.07 0.09 0.11 0.13 0.15 0.17 0.19 0.21 0.23 0.25 0.27 0.29 0.31 0.33 0.35 0.37 0.39 0.41 0.43 0.45 0.47 0.49 0.51 0.53 0.55 0.57 0.59 0.61 0.63 0.65 0.67 0.69 0.71 0.73 0.75 0.77 0.79 0.81 0.83 0.85 0.87 0.89 0.91 0.93 0.95 0.97 0.99
Age Gender Only
CDPS
MARA
Wakely
DXCG


Concurrent Models
Diagnosis and Pharmacy
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0.01 0.03 0.05 0.07 0.09 0.11 0.13 0.15 0.17 0.19 0.21 0.23 0.25 0.27 0.29 0.31 0.33 0.35 0.37 0.39 0.41 0.43 0.45 0.47 0.49 0.51 0.53 0.55 0.57 0.59 0.61 0.63 0.65 0.67 0.69 0.71 0.73 0.75 0.77 0.79 0.81 0.83 0.85 0.87 0.89 0.91 0.93 0.95 0.97 0.99
Age Gender Only
CDPS
ACG
MARA
Wakely
3M

II.B Tolerance Curves, Prospective Models




Prospective Models
Pharmacy-Only
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
0.21
0.23
0.25
0.27
0.29
0.31
0.33
0.35
0.37
0.39
0.41
0.43
0.45
0.47
0.49
0.51
0.53
0.55
0.57
0.59
0.61
0.63
0.65
0.67
0.69
0.71
0.73
0.75
0.77
0.79
0.81
0.83
0.85
0.87
0.89
0.91
0.93
0.95
0.97
0.99
Age Gender Only
CDPS
ACG
MARA
Wakely
DXCG


Prospective Models
Diagnosis and Pharmacy
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
0.17
0.19
0.21
0.23
0.25
0.27
0.29
0.31
0.33
0.35
0.37
0.39
0.41
0.43
0.45
0.47
0.49
0.51
0.53
0.55
0.57
0.59
0.61
0.63
0.65
0.67
0.69
0.71
0.73
0.75
0.77
0.79
0.81
0.83
0.85
0.87
0.89
0.91
0.93
0.95
0.97
0.99
Age Gender Only
CDPS
ACG
MARA
Wakely
DXCG
CRG
SCIO