Section II. Study Design

Risk Assessment Models Evaluated: Diagnosis and Pharmacy-based Models

There are a number of approaches that can be used for health risk assessment. This study focuses on methods that use medical diagnosis codes and/or pharmacy codes in administrative claim data to drive the risk assessment. For this study, seven health risk assessment models were evaluated, including three diagnosis-based models, three pharmacy-based models, and one model based on diagnosis and pharmacy data.

Specifically, the following models were evaluated:

- Adjusted Clinical Groups (ACGs) Version 4.5
- Chronic Illness and Disability Payment System (CDPS) Version 1.7
- Diagnostic Cost Groups (DCGs) Version 5.1
- Medicaid Rx
- RxGroups Version 1.0
- RxRisk
- Episode Risk Groups (ERGs) Version 4.2

The ACGs, CDPS, and DCGs are based on diagnosis data available from administrative claim records. Medicaid Rx, RxGroups, and RxRisk use pharmacy data. The ERGs use diagnosis and pharmacy data, and, to a small extent, some surgical procedure code data. The model versions referenced above were the most recently available when the study began in May of 2001.

The following section provides a brief description of each of the risk adjusters. For a more detailed description see Section VI of this report. Section VI also discusses some of the diagnosis and pharmacy-based risk assessment models that were not included in this study.

Adjusted Clinical Groups

Adjusted Clinical Groups (ACGs) is a diagnosis-based risk assessment model developed by Jonathan Weiner and other researchers at Johns Hopkins University. ACG Version 4.5, released in 2000, was used for this study. The ACGs classifies each member into one of 81 categories based on inpatient and ambulatory diagnosis codes for the member. ACGs differ from the other models in this study in that the ACG categories are mutually exclusive; that is, a member is classified into only one category. Many of the ACGs also reflect age/gender characteristics; thus, there are no separate age/gender variables in the model. The ACGs are also unique among the models included in this study in that they do not provide a set of standard risk weights.

Chronic Illness and Disability Payment System

The Chronic Illness and Disability Payment System (CDPS) is a diagnosis-based risk assessment model developed by Richard Kronick and other researchers at the

University of California, San Diego. CDPS Version 1.7, released in 2000, was used for this study. This model was originally developed for use with Medicaid populations, including disabled and Temporary Aid for Needy Families (TANF) populations. The CDPS model is an update and expansion of a prior model developed by Kronick and published in 1996 called the Disability Payment System (DPS). The DPS model was developed for the Medicaid disabled population.

The CDPS model assigns each member to one or more of 67 possible medical condition categories based on diagnosis codes. Each member is also assigned to one of 16 age/gender categories. For each member, the model predicts total medical costs based on the medical condition categories and age/gender category assigned. The model provides two sets of risk weights – one set calibrated for a TANF population and another set calibrated for a disabled population. In this analysis, the weights for the TANF population were used, since a TANF population is more similar to the commercial population used for this analysis. The model also provides different sets of risk weights for adults and children, both of which were used for this analysis.

Diagnostic Cost Groups

The Diagnostic Cost Groups (DCGs) model is a diagnosis-based risk assessment model originally developed by researchers including Randall Ellis and Arlene Ash at Boston University. The DCG models include a number of variations depending on the type of population being analyzed (commercial, Medicaid, Medicare), the source of the diagnosis data (inpatient only versus all encounters) and the purpose of the model (payment versus explanation).

For this analysis, DCG Version 5.1 of the commercial all-encounter model, released in 2000, was used. For the prospective analysis, the payment version of the model was used. For the concurrent analysis, the explanation version of the model was used (since DxCG Inc. does not offer a concurrent model designed for payment purposes). The commercial DCG models can predict both medical expenses including pharmacy spending and medical expenses excluding pharmacy expenses. For this analysis, the predictions included both medical expenses and pharmacy spending.

The DCG model assigns each member to one or more of 136 possible medical condition categories (called hierarchical condition categories (HCCs)) based on diagnosis codes. Each member is also assigned to one of 32 age/gender categories. Based on these medical condition and age/gender categories, the model predicts the total medical costs for each member.

Medicaid Rx

Medicaid Rx is a pharmacy-based risk assessment model developed by Todd Gilmer and other researchers at the University of California San Diego. This model was developed and released in 2000. The model was originally designed and intended for a Medicaid population and is an update and expansion of the Chronic Disease Score model developed by researchers at Group Health Cooperative of Puget Sound.

The Medicaid Rx model assigns each member to one or more of 45 medical condition categories based on the prescription drugs used by each member and to one of 11 age/gender categories. Based on the medical conditions and age/gender categories, the model predicts the overall medical costs for each member. The model includes separate sets of risk weights for adults and children.

RxGroups

RxGroups is a pharmacy-based risk assessment model developed by DxCG Inc in conjunction with Kaiser Permanente and clinicians from CareGroup and Harvard Medical School. RxGroups Version 1.0 released in 2001 was used. The RxGroups model can be used alone to predict total medical costs for each member or it can be used in conjunction with hospital inpatient diagnosis codes.

The RxGroups model will assign each member to one or more of 127 drug therapy categories and to one of 32 age/gender categories. RxGroups is somewhat different than the other pharmacy-based risk adjusters, in that it uses drug therapy categories as opposed to medical condition categories.

RxRisk

RxRisk is a pharmacy-based risk assessment model developed by Paul Fishman at Group Health Cooperative of Puget Sound. This model was developed and released in 2001. RxRisk is a combination of the original Chronic Disease Score model, designed for adults, and the Pediatric Chronic Disease Score model.

The RxRisk model assigns each member to one or more of 27 medical condition categories (for adults) or to one or more of 42 medical condition categories (for children). The model also assigns each member to one of 22 age/gender categories. Based on these categories the model predicts total medical costs for each member.

Episode Risk Groups

The Episode Risk Groups (ERGs) is a risk assessment model developed by Symmetry Health Data Systems. The ERGs are based on the Episode Treatment Groups (ETGs) model also developed by Symmetry which group medical services into episodes of care. These groupings are used for provider profiling. The ERGs were developed and released in 2001. The ERGs used in this analysis are based on Version 4.2 of the ETGs.

The ERG model assigns each member to one or more of 119 possible medical condition categories (called episode risk groups). Since the ERG output did not include a set of age/gender indicator variables, 22 age/gender categories were added when the risk weights were recalibrated for this analysis. The medical condition categories assigned to a member depend primarily on that member's diagnosis codes and pharmacy data. In a small number of cases, the ERGs assigned to a member depend on the presence of a defining surgery code. This differs from the other risk adjusters included in this study, which do not depend on the whether a particular procedure was performed.

The ERGs provide two sets of risk weights depending on whether the input data includes pharmacy information.

Data Used for Study: Commercial Group Population

The data used for this study includes claim and enrollment information for commercial employer group business. The data is limited to those members continuously enrolled from January 1, 1998 to December 31, 1999 for which medical and pharmacy claim data and enrollment information, including age and gender, are available. The data includes a nationwide mix of both Preferred Provider Organization (PPO) and Health Maintenance Organization (HMO) business.

The claim expenditure data is reported after provider discounts but before member cost sharing is deducted (i.e., it reflects total payments to health care providers). The data used permits up to 15 diagnoses per inpatient admission and up to 2 diagnoses per outpatient claim. For this analysis, all of the reported diagnoses are used.

The data was reviewed for general reasonableness and any categories of business that appeared to have data issues were removed. For any categories of business that included a significant number of encounter claims, the number of claims and dollar amounts by type of claim were reviewed for reasonableness. Mental illness and pharmacy claims were tested for completeness by examining the number and dollar amount of mental health and pharmacy claims. The percentage of non-users based on the pharmacy and medical claims data was examined as well.

The final data set used for this analysis included 749,145 members.

Study Methodology: 50/50 Split Design with Offered & Recalibrated Weights

Each risk adjuster was analyzed using three applications:

1. Prospective Model with Offered Risk Weights.
2. Prospective Model with Recalibrated Risk Weights.
3. Concurrent Model with Recalibrated Risk Weights.

These applications represent different approaches to implementing the risk adjuster model. The following section describes the differences in the three applications.

Prospective vs. Concurrent

A *prospective* application of a risk adjuster involves using claims data from a prior period of time to project medical claim costs for a future period. A *concurrent* (sometimes called retrospective) application involves using claims data from a period of time to project medical claim costs for that same period. In this study, the prospective models use diagnosis and pharmacy data from 1998 to predict total medical claim costs for each member for 1999. The concurrent model uses diagnosis and pharmacy data from 1999 to predict total medical claim costs for each member for 1999.

Offered vs. Recalibrated Risk Weights

For each risk adjuster, there is a *risk weight* for each medical condition category. The risk weight reflects an estimate of the marginal cost for a given medical condition relative to the base cost for individuals with no medical conditions. The *offered* risk weights are the standard risk weights that are provided with the risk adjuster software. The *recalibrated* risk weights were developed as part of this study and are based on the data set described above.

As mentioned previously, the ACGs do not include a standard set of risk weights with the software, since they expect that users will want to recalibrate the risk weights to reflect their own situation. (Since the ACGs is a categorical model, it is easier to recalibrate the risk weights since they can be calculated directly, without performing a regression analysis.)

Claim Truncation

For each application, the results were analyzed using three scenarios for truncating large claims: truncate large claims at $50,000, truncate large claims at $100,000, and no truncation. The truncation applies to total claim dollars for a given member for 1999.

Truncation of large claims is common when analyzing the predictive accuracy of risk adjusters for a variety of reasons, including:

1. Truncation limits the impact of outliers. This should provide more stability in the results when recalibrating the models and when analyzing predictive accuracy.
2. Large claims for a given person are generally not predictable. Accordingly, some researchers argue that they should be removed or limited when doing the analysis.
3. Truncation simulates the impact of reinsurance or stop loss at those levels.
4. Some measures of predictive accuracy are overly sensitive to large claims.

Steps in Study Methodology

The analysis for the offered weight application consists of three steps:

1. Separation of the data set into two equal-sized subsets: (1) a calibration subset and (2) a validation subset.
2. Assignment of individual scores for each member in the validation data subset using each risk adjuster and the offered weights (the score for a particular member reflects an estimate of the relative cost for that member).
3. Analysis of predictive accuracy using the validation data set to compare the score (i.e., predicted claims) of each member or group of members to actual claim dollars.

The analysis for the recalibration applications consists of five steps:

1. Separation of the data set into two equal-sized subsets: (1) a calibration subset and (2) a validation subset.
2. Assignment of medical condition categories (including drug therapy categories) and age/gender categories to each member using each risk adjuster.
3. Performance of a linear regression using the calibration data subset to determine the recalibrated risk weights.
4. Use of the recalibrated risk weights to assign scores for each member in the validation data subset.

5. Analysis of predictive accuracy using the validation data set to compare the score (i.e., predicted claims) of each member or groups of members to the actual claim dollars.

Each of these steps is described below.

Step 1. Separation of Data into Calibration and Validation Data Subsets

To allow for development and testing of recalibrated risk weights, a 50/50 split design was used for the study. Specifically, each member was randomly assigned into one of two subsets: (1) the calibration data subset and (2) the validation data subset, placing half of the population in each subset. The split design was used to avoid overfitting the data which could exaggerate the goodness of fit and various other measures of predictive accuracy.

Step 2. Grouping Each Member Using each Risk Adjuster

Each member is grouped (i.e., assigned to certain medical condition categories, including drug therapy categories, and age/gender categories) by each risk adjuster model. Each risk adjuster model produces a set of indicator variables (0 or 1) representing the condition and age/gender categories assigned. For the prospective analysis, the indicator variables are based on 1998 diagnosis and pharmacy data. For the concurrent analysis, the indicator variables are based on 1999 diagnosis and pharmacy data.

The risk adjuster software was used to group each member for each of the risk adjusters. Milliman researchers ran the software for each of the risk adjusters, except for the ERGs. For the ERGs, Symmetry grouped the members into medical condition categories. (For the ERGs, the rest of the analysis, including recalibration and measurement of predictive performance, was done by Milliman using the same methodology as used for the other risk adjusters.)

Step 3. Calculation of Recalibrated Risk Weights

The calibration data subset was used to develop a new set of risk weights using the study data. In general, to calculate the risk weights for a particular risk adjuster, the following multivariate linear regression model is used:

$$P \quad = \quad \sum_i (RWMCC_i \times MCC_i) + \sum_k (RWAG_k \times AG_k)$$

Where:

P = total claim payments for 1999 (including medical and pharmacy)
$RWMCC_i$ = risk weight for medical condition category i
MCC_i = indicator variable (0 or 1) for medical condition category i
$RWAG_k$ = risk weight for age/gender category k
AG_k = indicator variable (0 or 1) for age/gender category k

A linear regression is performed to determine a set of risk weights that best fit the calibration data set.

A separate calibration analysis was performed for each level of claim truncation. Also, separate calibrations are performed for the prospective and concurrent applications. Accordingly, there are six sets of recalibrated risk weights for each risk adjuster.

In this analysis, the initial results included some negative risk weights for some of the risk adjusters. This can occur due to noise in the data or, in some cases, there may be a clinical explanation. The majority of the negative risk weights were not statistically significant.

For this study, any negative risk weights in the initial results were set to 0 in order to determine the final set of risk weights. This adjustment had very little impact on the results of the study. Negative risk weights are typically removed when developing a payment model for actual implementation since, according to Richard Kronick, "it would be awkward to reduce plan payments because of additional diagnoses" (Kronick et al, 2000). Similarly, negative risk weights might create a financial incentive to avoid treatment or coding of treatment for certain medical conditions. It should be noted that Kronick includes negative risk weights in his general analysis of risk adjustment models for Medicaid populations. Kronick states that "... we included a number of ADGs that have statistically significant, negative parameter estimates and that would likely be excluded if an ADG payment model were implemented..."

A number of other adjustments are commonly used in developing a final set of risk weights for a payment model for actual implementation. These other adjustments can include: removing variables that are not statistically significant, smoothing the age/gender risk weights, blending the developed risk weights with the "offered" risk weights, combining various variables in the payment model, recalibrating the risk weights after removing any variables, clinical review of the relationships, testing the stability of the risk weights with different claim truncation levels, and testing the stability of the risk weights using subsets of the data. This study does *not* include any of these further adjustments.

The ACG model does not have a separate set of age/gender variables since age/gender is built into the ACG categories. The structure of the ACG methodology, which places each individual into exactly one category, allows a direct calculation of risk weights, rather than the use of a linear regression to develop them.

Step 4. Assignment of Score for each Member in the Validation Data Subset

Each member in the validation data subset is scored using the indicator variables described in Step 2 and the recalibrated risk weights from Step 3.

Step 5. Analysis of Predictive Accuracy

In the final step, the predictive performance of the models is analyzed by comparing the risk scores with the actual claim dollars incurred. This comparison is done for both individuals and groups of individuals as described below.

Measures Used to Analyze Predictive Accuracy: Individual and Non-Random Groups

A variety of measures were used to compare the predictive accuracy of the risk adjusters examined in this study. In general, these measures compare actual claim dollars with predictions from the risk adjuster models. This comparison is performed on two levels: (1) by individual and (2) by group.

Measures of Predictive Accuracy- Individual Level

The individual measures of predictive accuracy include:

1. Individual R-squared,
2. Mean absolute prediction error, and
3. A new measure, derived from mean absolute prediction error. (This new measure is presented and discussed separately in Section VII of this report.)

Individual R-squared is described as the percentage of the variation in medical claim costs explained by the risk adjuster model. Variation refers to the difference in medical costs for a given individual compared to the average medical cost for all individuals. The formula for R-squared is:

$$R^2 \quad = \quad 1 - (\sum_i (a_i - \hat{a}_i)^2) / (\sum_i (a_i - \bar{a})^2)$$

Where:

a_i = actual claim dollars for person i
$\hat{a}_i$ = predicted claim dollars for person i (based on a regression model)
$\bar{a}$ = mean of the actual claim dollars
i goes from 1 to n, where n is the number of people

Mean absolute prediction error is calculated as follows. First, the prediction error for each individual is determined by calculating the difference between predicted medical costs and actual medical costs. Next, the absolute value of each of these prediction errors is calculated, and, finally, the mean of the absolute prediction error across all individuals is determined. The formula for mean absolute prediction error (MAPE) is:

$$\text{MAPE} = \quad (\sum_i | a_i - \hat{a}_i |) / n$$

Where:

a_i = actual claim dollars for person i
$\hat{a}_i$ = predicted claim dollars for person i
i goes from 1 to n, where n is the number of people

Different arguments are made regarding the merits of alternative methods for measuring goodness of fit. Individual R-squared is a standard statistical measure for assessing model results. It is commonly used for measuring predictive accuracy of risk adjusters. It is a single summary measure on a standardized scale of 0 to 1, where 0 indicates that the model explains 0% of the variation in cost among the individuals and 1 indicates that the model explains 100% of the variation i.e., 100% accuracy in the predictions. The standardized scale helps with comparability between studies. However, there still are many potential issues associated with comparing individual R-squared from one study with another that may make the comparisons inappropriate or invalid. These issues include differences in the data sets, study design, and data quality.

Individual R-squared has certain drawbacks. Because it squares each prediction error, it tends to be overly sensitive to the prediction error for individuals with large claims. According to the prior Society of Actuaries (SOA) study, "because R^2 squares the errors of prediction, it can be greatly affected by a relatively small number of cases with very large prediction errors. Given the typical distribution of health expenditures across individuals, where a small number of individuals have relatively large expenditures, this is a concern for our analysis." (Dunn, et al., 1995) This is one of the reasons for truncating large claims when individual R-squared is used as a measure of predictive accuracy. The prior SOA study generally presents results with claims truncated at $25,000.

Another concern with individual R-squared is that it might give the appearance of poor performance. For example, individual R-squared is typically around 10% to 20% for prospective applications. As a result, health care decision makers may question the value of risk adjustment i.e. "Why invest in an expensive and complicated process that explains at most 15% of the variation in claims?" In fact, the key issue for most risk adjustment applications is the accuracy of the predictions for groups of people, rather than for each individual. As a result, many researchers also look at group level measures, such as those described below. One study showed that a diagnosis based risk adjuster that explained only 9% of the variation in claims across individuals, explained over 80% of the variation across certain groups. (Ash, et al, 1998) This result may vary significantly based on how the groups are defined.

The mean absolute prediction error is also a single summary measure of predictive accuracy. On the positive side, it does not square the prediction errors and, so, is not overly sensitive to large claims. However, it is not expressed on a standardized scale, so comparisons across studies are difficult to make.

Measures of Predictive Accuracy – Group Level

A group level measure of predictive accuracy involves adding up the total predicted claims for a group of individuals and comparing that value to the actual claims for the same group. This comparison gives a *predictive ratio.* A predictive ratio that is closer to 1.0 indicates a better fit. The predictive ratio is the reciprocal of the common actual-to-expected (A to E) actuarial ratio.

The group level measures differ in terms of how the groups are determined. There are two general approaches: (1) *non-random groups* and (2) *random groups.* Non-random refers to grouping individuals based on selected criteria. The common criteria used for analyzing risk adjusters include groups based on medical condition or amount of claim dollars. Non-random groups can also be defined based on other criteria, such as a being part of a particular employer group. This is sometimes referred to as using *real groups.* Random groups refer to groups created by selecting individuals at random from the study data set.

Non-Random Groups used for This Study

This study uses non-random groups based on the following criteria:

1. Medical condition in 1998,
2. Medical condition in 1999,
3. Quintiles based on medical claim dollars for 1999, and
4. Ranges of medical claim dollars for 1999.

The medical conditions used for this study include: breast cancer, congestive heart failure, asthma, depression, and HIV. As is common in these types of studies, the medical conditions are determined using medical diagnosis codes. It should be noted that this approach might create a fundamental bias in favor of risk adjusters that are based on diagnosis data. This reflects that a risk adjuster which distinguishes among people based on particular criteria (e.g., diagnosis codes) will naturally tend to perform better when predicting expenditures for groups of people determined using the same type of criteria.

Note: For different medical conditions, the performance of the risk adjuster models may change significantly. For a given medical condition, a risk adjuster will naturally tend to perform better on this test if it has a medical condition category that matches more closely with the definition of the medical condition used in this study.

Grouping Individuals using Base Year vs. Prediction Year Information

There are two alternate approaches in determining the non-random groups. One approach uses claim information from the base year (i.e., 1998) to define the group. The other approach uses claim information from the prediction year (i.e., 1999) to define the group. For medical conditions, the groups were constructed using both approaches. For claim dollars, the groups were constructed based on 1999 claim dollars.

Predictive ratios for groups based on claim information from the base year (e.g., medical condition in 1998) will naturally tend to be closer to 1 than predictive ratios for groups based on claim information from the prediction year (e.g., medical condition in 1999). This can occur for two reasons: (1) the tendency for health care expenditures to "regress toward the mean" for a given group of people and/or (2) the difficulty in predicting claim levels, based on historical claim information, for people that are newly diagnosed with a medical condition.

Measures that use groups based on claim information from the prediction year may be more useful when analyzing risk adjusters for applications such as underwriting/rating, identification of people for case or disease management, provider profiling, and provider payment. These types of measures help us answer questions such as: How well can the risk adjuster predict people's claims for the next year? How well can the models predict who will have a large claim next year? How well do the models adjust for those people that have a particular medical condition next year?

Measures that use groups based on claim information from the base year may be more useful when analyzing risk adjusters for applications such as health plan payment. These types of measures help us answer questions such as: If a health plan, directly or indirectly, selected members based on their claim history (i.e., past medical conditions or expenditures), would the health plan receive a fair payment for the upcoming year?

Section III. Results

Individual Level Results

General Findings

- For prospective risk assessment, the pharmacy-based models perform at a level similar to the diagnosis-based models. The pharmacy-based models perform slightly better when using the mean absolute prediction error as the performance measure. The diagnosis-based models perform slightly better when using R-squared as the performance measure.
- For concurrent risk assessment, the diagnosis-based models outperform the pharmacy-based models.
- For prospective risk assessment, the R-squared performance of the models varies from 9.8% to 19.3%, with offered weights and claims truncated at $100,000.
- For prospective risk assessment, the R-squared performance of the models varies from 14.0% to 19.8%, with recalibrated weights and claims truncated at $100,000.
- For concurrent risk assessment, the R-squared performance of the models varies from 29.2% to 54.7%, with recalibrated weights and claims truncated at $100,000.
- The risk adjusters originally developed and calibrated for Medicaid populations (CDPS and Medicaid Rx) showed significant improvement in their predictive performance when the risk weights were recalibrated. The performance of CDPS, as measured using R-squared, increased from 12.5% to 18.6%, with claims truncated at $100,000. The performance of Medicaid Rx increased from 9.8% to 16.5%.
- The general performance of the other risk adjusters increased slightly after recalibration, as measured by R-squared. The increase in performance varied from a 2.9% increase in R-squared for DCGs (with claims truncated at $50,000) to a 0.1% decrease in R-squared for RxGroups (with claims not truncated).
- Recalibration tended to result in a greater increase in performance when claims are truncated at $50,000 and a smaller increase in performance when claims are not truncated. (This is true even when the increase in R-squared is expressed on a relative or percentage basis.)
- As one would expect, the concurrent models significantly outperform the prospective models.
- It appears that the performance of the diagnosis-based risk adjusters has improved significantly since the 1995 Society of Actuaries (SOA) study. This improvement likely results from a combination of more detailed data reporting and refinement of the risk assessment models. (Note: the prior SOA study used only the primary diagnosis code and a number of the risk adjusters in the prior SOA study used only inpatient or only ambulatory diagnosis codes.)

Note: In most real-life prospective applications, the performance of the pharmacy-based models, relative to the diagnosis-based models, would be better than shown in this study due to shorter time lags for receiving pharmacy claim data compared to medical claim data.

The following section provides a more detailed presentation of the study results.

Prospective Model – Offered Weights

Table 3.1 summarizes R-squared and mean absolute prediction error for each risk adjuster when used for a prospective application with the offered weights. A higher R-squared indicates better predictive accuracy. A lower mean absolute prediction error indicates better predictive accuracy. Results for the ACG method are not available (NA) since the ACGs do not come with offered weights. The table shows the type of risk adjuster based on what data is used for the risk assessment: diagnosis data (diag), pharmacy data (Rx), or diagnosis and pharmacy data (Diag+Rx).

Table 3.1: Summary of R-squared and Mean Absolute Prediction Error – Prospective Model with Offered Weights

Risk Adjuster	Type of Risk Adjuster	R-Squared with claims truncated at:			Mean Absolute Prediction Error with claims truncated at:		
		$50,000	$100,000	None	$50,000	$100,000	None
ACG	Diag	NA	NA	NA	NA	NA	NA
CDPS	Diag	.134	.125	.103	2095	2210	2299
DCG	Diag	.195	.180	.143	1987	2098	2187
Medicaid Rx	Rx	.116	.098	.071	2103	2222	2310
RxGroups	Rx	.206	.181	.134	1916	2027	2113
RxRisk	Rx	.175	.148	.111	1988	2108	2200
ERG	Diag+Rx	.218	.193	.146	1875	1987	2082

As shown in Table 3.1, the ERGs perform well on each of the six measures. This is not surprising given that the ERGs use more information than any of the other risk adjusters included here. As described previously, the ERGs use diagnosis, pharmacy, and, in a small number of cases, certain surgery procedure codes. The other risk adjusters use either diagnosis or pharmacy data, but not both. Many of the risk assessment models specifically do not consider the treatment that an individual receives so that the risk scores are not biased by the practice patterns of the health care providers. This is a concern when using risk adjusters for health plan payment or provider payment. However, when using risk adjusters for underwriting/rating or case management, this is not an issue.

The CDPS and Medicaid Rx models do not perform as well as the other models. This is not surprising, given that these models were originally designed and calibrated for Medicaid populations. As the results below show, when these models are recalibrated for a commercial population, their performance improves significantly.

In general, the performance of the pharmacy based risk adjusters is similar to the performance of the diagnosis based risk adjusters. The pharmacy based risk adjusters perform better, relative to the diagnosis based risk adjusters, when using mean absolute prediction error. The diagnosis based risk adjusters perform better, relative to the pharmacy based risk adjusters, when using R-squared. Also, the relative performance of the pharmacy based risk adjusters tends to improve when using lower levels for truncating large claims. This would seem to indicate that the diagnosis based risk adjusters tend to do a relatively better job in predicting for large claims.

The level of claim truncation used by the developers of the risk adjusters to determine the offered weights could affect the results shown in Table 3.1. For example, suppose that the developers of the ERGs determined the offered weights using a $100,000 claim truncation level. If the developers re-determined the offered weights using untruncated claims, then one might expect the R-squared for the ERGs to increase at the untruncated claim level and decrease at the $100,000 claim truncation level.

Prospective Model – Recalibrated Weights

Table 3.2 summarizes R-squared and mean absolute prediction error for each risk adjuster when used for a prospective application with the recalibrated weights.

Table 3.2: Summary of R-squared and Mean Absolute Prediction Error – Prospective Model with Recalibrated Weights

Risk Adjuster	Type of Risk Adjuster	R-Squared with claims truncated at:			Mean Absolute Prediction Error with claims truncated at:		
		$50,000	$100,000	None	$50,000	$100,000	None
ACG	Diag	.172	.140	.099	1972	2100	2193
CDPS	Diag	.208	.186	.149	1944	2070	2164
DCG	Diag	.224	.198	.154	1902	2032	2133
Medicaid Rx	Rx	.200	.165	.119	1931	2062	2159
RxGroups	Rx	.222	.185	.132	1882	2014	2113
RxRisk	Rx	.188	.154	.111	1960	2091	2187
ERG	Diag+Rx	.230	.197	.148	1854	1983	2079

When interpreting and using the results shown in Table 3.2, keep in mind that R-squared can be overly sensitive to large claims. As mentioned in the prior section, this becomes a more significant issue when claims are truncated at higher limits (i.e., $100,000 or no truncation). This is not a concern with the mean absolute prediction error, since it does not square the prediction error.

As shown in Table 3.2, the ACGs do not perform as well as some of the other risk adjusters. This may reflect that the ACGs use mutually exclusive medical condition categories, while all of the other models are additive. That is, the other models can assign an individual to multiple medical condition categories and then add together the risk weight for each such condition to develop a prediction for each individual. The additive models allow much more flexibility in describing the overall medical condition of a given individual since you can use virtually any combination of the different medical condition categories. (Note that some of the additive risk adjusters use hierarchical designs that limit, to some degree, the possible combinations of medical condition categories.)

In comparing the performance of various risk adjusters, one should consider how the models will be implemented. For example, the ACGs do not come with a standard set of weights since the expectation is that the user will calibrate the model. However, the other risk adjusters do come with a standard set of risk weights. Accordingly, health plans might typically use the DCGs with the standard set of weights, rather than go through the process of recalibration. (Note: The recalibration of the ACGs, since it uses mutually exclusive categories rather than additive categories, is more straightforward and more likely to give reasonable results than the recalibration of the other risk adjusters.) So, for this scenario, it might be more appropriate to compare the performance of the recalibrated ACGs to the performance of the DCGs with offered weights. Based on the mean absolute prediction error with claims truncated at $100,000, the performance of the two models is nearly identical (the mean absolute prediction error for the ACGs with recalibrated weights is 2100 and the mean absolute prediction error for the DCGs with offered weights is 2098).

The pharmacy based models tend to perform better, relative to the diagnosis based models, when using the mean absolute prediction error as the measure, whereas, the diagnosis based risk adjusters tend to perform better, relative to the pharmacy based models, when using R-squared. For example, when comparing related products (i.e., DCG & RxGroups from DxCG Inc. and CDPS & Medicaid Rx from the University of California, San Diego researchers) the diagnosis based product outperforms the pharmacy based product based on R-squared whereas the pharmacy based product outperforms the diagnosis based product based on mean absolute prediction error.

Concurrent Model – Recalibrated Weights

Table 3.3 summarizes R-squared and mean absolute prediction error for each risk adjuster when used for a concurrent application with the recalibrated weights.

Table 3.3: Summary of R-squared and Mean Absolute Prediction Error – Concurrent Model with Recalibrated Weights

Risk Adjuster	Type of Risk Adjuster	R-Squared with claims truncated at:			Mean Absolute Prediction Error with claims truncated at:		
		$50,000	$100,000	None	$50,000	$100,000	None
ACG	Diag	.429	.376	.282	1487	1599	1685
CDPS	Diag	.440	.418	.355	1576	1697	1799
DCG	Diag	.564	.547	.466	1394	1509	1618
Medicaid Rx	Rx	.372	.328	.244	1661	1797	1909
RxGroups	Rx	.420	.376	.279	1569	1707	1823
RxRisk	Rx	.339	.292	.213	1724	1854	1956
ERG	Diag+Rx	.474	.427	.347	1441	1582	1700

As can be seen from Table 3.3, the diagnosis based models outperform the pharmacy based models when used for concurrent risk assessment.

Comparison of Results with and without Recalibration

Table 3.4 compares the performance of the risk adjustment models with and without recalibration of the risk weights. By far, the largest gains in performance occurred for the CDPS and Medicaid Rx risk adjusters.

Table 3.4: Comparison of Performance of Risk Adjustment Models with and without Recalibration of Risk Weights – Prospective Models

Risk Adjuster	Type of Risk Adjuster	R-Squared with claims truncated at $100,000 with:		Mean Absolute Prediction Error with claims truncated at $100,000 with:	
		Offered Weights	Recalibrated Weights	Offered Weights	Recalibrated Weights
ACG	Diag	NA	.140	NA	2100
CDPS	Diag	.125	.186	2210	2070
DCG	Diag	.180	.198	2098	2032
Medicaid Rx	Rx	.098	.165	2222	2062
RxGroups	Rx	.181	.185	2027	2014
RxRisk	Rx	.148	.154	2108	2091
ERG	Diag+Rx	.193	.197	1987	1983

Table 3.5 shows the increase in performance due to recalibration of the risk weights for the prospective model. Specifically, the table shows the increase in R-squared between the prospective model with the recalibrated weights and the prospective model with the offered weights.

Table 3.5: Increase in Performance due to Recalibration – Prospective Model

Risk Adjuster	Type of Risk Adjuster	Increase in R-Squared due to Recalibration with claims truncated at:		
		$50,000	$100,000	None
ACG	Diag	NA	NA	NA
CDPS	Diag	.074	.062	.046
DCG	Diag	.029	.018	.012
Medicaid Rx	Rx	.084	.067	.047
RxGroups	Rx	.015	.004	-.001
RxRisk	Rx	.014	.005	.001
ERG	Diag+Rx	.012	.003	.002

The CDPS and Medicaid Rx models show a very significant increase in performance due to recalibration. This might be expected since the offered weights for both of these models have been calibrated for Medicaid populations. The DCGs show a moderate improvement in performance. The other models show somewhat smaller increases in performance.

It is interesting to note that the increase in performance tends to decline when there is less claim truncation. (This occurs even when the increase is expressed on a relative or percentage basis, rather than additive basis.) One possible explanation for this pattern is that, although recalibrated risk weights provide a better fit, when the risk weights are based on untruncated claims it is likely that there will be more anomalies in the resulting risk weights that may require review and smoothing. (In this analysis, any negative risk weights were removed, but no review or smoothing beyond that occurred.) Another possible factor that might explain some of this pattern relates to the level of claim truncation used by the developers to determine the offered risk weights. For example, if the developers used no claim truncation, then the offered weights will fit the data better at that level of claim truncation and a smaller increase in performance due to recalibration would be expected.