contains diagnoses may require four or more months to be adequately complete. Prescription data has the benefit of no provider-required data submission to a plan, so that a potential barrier is eliminated.

The Time Required for Eligibility Data to be Updated

It may require two months to receive updates of changes in eligibility status of plan members from purchaser. For some large employers, the retroactive adjustment for new enrollment, enrollment status changes, or terminations may take even longer.

The Time to Execute the Risk Scoring and the Frequency of Risk Scoring

Purchasers can control how often and how fast they compute and assign risk scores. Combined with the usual claims run-out lag, the range can be from a minimum of six-months up to 24 months.

Data delays are an implementation problem for any risk adjustment model. For individual-level prospective models, the enrollee must be continuously eligible for 6-12 months in the assessment period, 6-18 months in the claims delay period, and 1-12 months in the payment period for a health plan to be paid for the risk of that enrollee. This continuous enrollment requirement can remove up to 40% to 50% of any currently enrolled Medicaid population from the clinical condition risk assessment (e.g., all new enrollees), thus dramatically reducing the predictive performance of the total capitation system. Therefore, it is important to know the extent to which the delay has reduced the performance of the model compared to its "laboratory" tested results that often included no delay.

Data Issues

Implementation will be more challenging if there is not some early testing and data handling in the planning phase. A simulation may be the first time the purchaser will be handling massive amounts of data, especially the encounter data. It is wise to expect a great deal of last minute processing of encounter data.

The critical data quality issues for risk adjustment are not necessarily those that are captured in a fee-for-service edit system. It will be necessary to selectively bypass some of these fee-for-service edits.

Data should be examined for reasonableness. Examining the frequency distributions of various data elements will help identify incomplete encounter data. Although there are no norms, there is some information about what non-contact percentages to expect. Data may be missing because of sub-capitation or because of carve-outs. A common problem is missing mental health provider data for a program that covers mental health services. Each person should have similar benefits such as prescription drugs, co-insurance, or deductible levels.

Different types of plans have different types of data problems. Staff model HMOs that have limited experience with fee-for-service billing will have concerns about data layout

for encounters and the bundling of services. Plans whose systems truncate the number of diagnosis codes per record will raise concerns about the number of diagnoses.

Data quality can be an issue at the plan level and also at the provider level. Data concerns at the plan level revolve around completeness, while data issues at the provider level include both completeness and accuracy.

For diagnosis data, the concern at the plan level is to capture all diagnoses already recorded by the provider. Plans may be missing diagnoses for two reasons:

- They may be missing encounter data from some providers.
- They may be truncating the number of diagnoses per encounter supplied by the provider.

The California Joint Purchaser study of data quality for risk adjustment found that diagnosis data quality, as measured by the number of diagnoses per encounter and other indicators, varies significantly across medical groups. Plans that rely on data from a limited number of medical groups may have their risk underestimated.

Plans whose payments have been adjusted by purchasers using diagnosis-based risk adjusters, such as those participating in the Colorado and Maryland Medicaid programs, have in many cases made significant improvements in addressing plan-level problems with data completeness.

Prescription data is complete and accurate at the plan-level for most significant conditions and does not involve data transfer from providers.

For diagnosis coding at the provider level, there are three possible activities that can change the number and distribution of diagnoses and can increase the measured risk for a population when, in fact, the underlying morbidity of the population may be stable:

- Diagnostic discovery -- Increased number and severity of diagnoses are reported, all of which are appropriate. The correction of previous underreporting will reduce the problem of lack of persistence of diagnoses and will more fairly represent the illness burden of the population.
- Diagnostic creep -- Increased number and severity of diagnoses for cases where the diagnosis is uncertain. This represents an upward bias in response to payment incentives. Many groupers try to minimize this problem by bundling related diagnoses and by excluding ill-defined codes.
- Tentative diagnoses -- Represents a potential source of error when a diagnosis is appropriately used to justify a diagnostic procedure (rule-out) or to signal the need to treat a person without confirmatory diagnostic tests as if the patient has the disease (presumptive), because delay in treatment is harmful. Here too, the groupers have rules for excluding codes that are highly likely to be tentative.

Purchasers have so far not detected significant changes in provider-level coding patterns, but it is important to keep looking and to set up monitoring and auditing systems that examine coding practices.

Some purchasers have begun medical record audits and some have not. One strategy develops linkages with other measurement activities such as quality assurance. Others

seek to automate data-quality monitoring through clinical edits and audits of encounter data for illogical combinations or changes in the relationship between diagnoses and services provided.

Model Calibration Issues

Experience has taught that imported risk weights can be sufficiently valid and stable for many applications if they are based on a similar population with similar covered costs. For some applications described above, however, it may be preferable to calculate weights on the user's population. This requires both a sufficiently large population and adequate data. Whether a user imports or calculates its own, weights must be updated at regular intervals to account for changes in practice patterns, coding changes, and significant changes in benefit design. Because prescribing patterns change much more rapidly than general treatment patterns, prescription-based models will age more rapidly and will need more frequent updates.

Although many of the diagnosis-based models are calibrated, of necessity, on fee-for-service data, and experience has taught us that these weights are reasonably valid for managed care applications, there is a desire to move, when possible, to encounter-based weights. There may be some gain in validity from encounter-based weights that reflect the clinical and coding practices of a managed care environment.

Using encounter data for weights requires the highest standard for completeness. Although duplications of diagnoses can be tolerated in the risk assessment, duplications of charges could cause significant errors when establishing proper weights. Another issue to consider in developing weights is how to apply charges to encounter data.

Section VI. Description of Risk Adjusters

This section briefly describes the background and key features of the risk adjusters evaluated in the study.

Diagnosis-based Models

The Johns Hopkins Adjusted Clinical Groups (ACGs) was developed by Jonathan Weiner and Barbara Starfield at Johns Hopkins University in the mid-1980s. This method was initially developed for epidemiological research on primary care. The system logic began with a clinical focus and was later modified to explain variation in total medical expenditures. This focus led ACG developers to be concerned with the total morbidity rather than specific diseases. ACGs was developed and tested on data from a few commercial HMOs and two state Medicaid data sets. ACGs was the first system to be used by health plans, primarily for profiling.

Most diagnoses are assigned to one of about 30 Adjusted Diagnosis Groups (ADGs). The assignment is based on clinical criteria such as severity, chronic or acute, and prognosis. ADGs, compared with the building blocks of DCG models, generally are not defined by specific disease but combine diseases with similar clinical management issues. ADGs are then combined with age and gender to produce mutually exclusive ACGs based on an analysis of clusters of ADGs. The approach emphasizes the number and severity of co-morbidities. An individual can have many ADGs, but only one ACG. Payment weights can be derived for ADGs in an additive model or, as is most common, for ACGs as defined rate cells. Under a recent contract with CMS, ACGs were revised and calibrated for the Medicare population. ACGs is licensed by Computer Science Corporation, Inc. ACGs are used by two Medicaid programs, a few employers and a number of health plans.

The Chronic Disease and Disability Payment System (CDPS) was developed by Richard Kronick and Tony Dreyfus at the University of California – San Diego in the mid-1990s as a demonstration project for providing managed care to a disabled population. Then called the Disability Payment System (DPS), CDPS is a new version that has been revised and expanded for the entire Medicaid population by refining or adding diagnosis categories important to a TANF population, e.g., pregnancy.

Most medium to high-cost chronic illness diagnoses are used to assign risk scores. Diagnoses are initially assigned to chronic condition categories. These categories retain the identity of the disease by diagnosis categories. The chronic illness categories are arranged into hierarchies. Only the highest cost category in a disease hierarchy is used to produce an individual's total risk score. An individual's risk score is computed by adding the weights for the age and gender category and any medical categories across the hierarchies. Within the hierarchies, only the highest cost category identified is used to assess risk. In this way CDPS is similar to HCCs.

Currently, seven states Medicaid managed care programs are using CDPS. Under a recent contract with CMS, CDPS was revised and calibrated for a Medicare population.

May 24, 2002

Before this study, CDPS had not been formally modeled for a commercial population. CDPS is available for essentially a no cost license by contacting the developers at University of California-San Diego.

Diagnostic Cost Groups (DCGs) was developed in the mid-1980s as an inpatient-data model for Medicare data. The original models were developed by Arlene Ash and Randall Ellis at Boston University. A number of models followed, including Hierarchical Condition Categories (HCCs), a comprehensive diagnosis model. The models have been refined over the years, and HCC models have now been developed for Medicaid and commercial populations in addition to Medicare.

The method assigns most diagnosis codes to categories called DxGroups. These categories are similar to disease categories. The DxGroups are then combined with related diseases into Condition Categories. A number of categories are arranged in hierarchies of diseases of similar type, primarily the same body system. Within hierarchies, only the weight of the highest cost category is used to assess to risk. An individual's risk score is computed by adding the weights of age and gender category and of each Hierarchical Condition Category identified. The DCG system is licensed by DxCG, Inc.

The Principle Inpatient Diagnosis model of DCGs is currently used to risk adjust a portion of payments to health plans in the Medicare+Choice program and a CMS customized version of the HCC model has been selected for the Medicare + Choice program for implementation in 2004. Employers and health plans are also using DCGs.

Pharmacy-based Models

Medicaid Rx was developed by the researchers who developed CDPS. The model was developed and validated for a Medicaid population. The prescription risk assessment logic is based on the Chronic Disease Score (CDS) model developed by researchers at Group Health Cooperative of Peugeot Sound. Medicaid Rx was created by revising CDS to include primarily chronic conditions prevalent in the Medicaid population. The Medicaid Rx model uses prescription data (NDC codes) to indicate the presence of a chronic disease. Prescriptions with multiple uses are often excluded. Medicaid Rx in a few instances adds some prescriptions that are typically prescribed for acute illnesses if the prescription is long standing, e.g. antibiotics for chronic infections. Additional information on Medicaid Rx is available from the CDPS developers at the University of California – San Diego.

RxRisk, formerly the Chronic Disease Score (CDS), was developed by researchers Paul Fishman and Michael Von Korpf at Group Health Cooperative of Puget Sound. RxRisk was developed from the research, modeling CDS for different populations. CDS was one of the first prescription data models to be developed and tested. RxRisk uses outpatient pharmacy data (NDC codes) to classify patients into disease categories. An individual's risk score is computed by adding the weights for age and gender categories with the weights for any identified disease category.

RxGroups were developed by the DxCG researchers in cooperation with Kaiser Permanente. Prescription data (NDC codes) are assigned to RxGroups. RxGroups are then combined to create Aggregated RxGroups. These Aggregated RxGroups are arranged into hierarchies, and a hierarchical additive model that includes age and

gender factors is used to compute an individual's total risk score. Another version of RxGroups combines inpatient diagnoses with ambulatory prescriptions. Additional information on RxGroups is available from DxCG, Inc.

Models based on Diagnosis and Pharmacy Data

Episode Risk Groups (ERGs) were developed in 2001 by Dan Dunn and researchers at Symmetry Health Data Systems, Inc. An ERG is a derivative of the Episode Treatment Groups (ETGs), an episode of care analysis system. Over 600 episodes from the ETG system are combined to produce ERGs. A surgical episode and medical episode for the same condition are combined in most instances, reducing the problem of risk adjustment for the care provided rather than for health status. The ERGs are then used to calculate a person's risk score by adding the weights for each identified ERG. ERGs are licensed by Symmetry Health Data Systems, Inc. ERGs are currently being distributed to customers of the ETCG system.

Other Models – Not Included in this Study

Other new risk assessment models are currently being tested and should be considered for future studies as they become more widely distributed. These include Clinical Risk Groups (CRGs) developed by 3M Health Systems. CRGs uses diagnoses and a selected set of non-discretionary procedures to calculate a risk score. Ingenix offers several predictive models that also rely on claims data, including one that uses only prescription drug data. In addition, CMS recently announced a Selected Condition derivative of HCCs to be used for Medicare+Choice in 2004. This model includes 61 condition categories and requires only about 3400 ICD-9 codes. The original HCC model included over 100 condition categories and used most of the over 15,000 ICD-codes.

Section VII. A New Measure of Predictive Accuracy

The researchers have developed a new measure of predictive accuracy. The researches believe that this new measure has advantages over existing, commonly used measures. The new measure quantifies predictive accuracy at the individual level. The following defines the new measure, and compares the new measure to some commonly used measures.

A New Measure of Predictive Accuracy

The researchers have developed a new measure of predictive accuracy, called *Cumming's Prediction Measure (CPM)*. (Since the researchers have not seen this measure defined or promoted in the research literature dealing with risk adjusters, it has been named, for the time being, after the developer. This should help to indicate that this is a newly developed measure, which has not yet been well studied, at least in the area of risk adjuster research.)

Cumming's Prediction Measure

Cumming's Prediction Measure (CPM) is calculated as shown below:

CPM = 1 – (Mean Absolute Prediction Error) / (Mean Absolute Deviation from Average)

The *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 *mean absolute deviation from average* is calculated as follows. First, the deviation from average for each individual is determined by calculating the difference between the actual medical costs for that individual and the average medical costs across all individuals. Next, the absolute value of each of these deviations is calculated, and, finally, the mean of the absolute deviation across all individuals is determined.

Comparison with Other Measures

The commonly used measures of predictive accuracy on an individual level include R-squared and mean absolute prediction error. These measures have certain advantages and disadvantages, as discussed in Section II of this report.

Cumming's Prediction Measure (CPM) combines the best qualities of Individual R-squared and mean absolute prediction error. CPM is a single, summary statistic of goodness of fit. Like individual R-squared, CPM is expressed 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. However, CPM uses the absolute value of the prediction errors rather than the square of the

prediction errors and, so, is not overly sensitive to large claims. In this respect, it is similar to the mean absolute prediction error.

Both CPM and R-squared can be described as the percentage of the variation in cost among individuals that is explained by the model. The difference is that R-squared measures variation using the square of each prediction error, whereas, CPM measures variation using the absolute value of each prediction error.

The sensitivity of R-squared to large prediction errors is a concern. 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. Because of this concern, the prior SOA study generally presents results with claims truncated at $25,000.

CPM is closely related to the mean absolute prediction error, from which it is derived. For a given level of claim truncation, CPM will always provide the same relative ranking of risk adjuster performance as the mean absolute prediction error. However, CPM also expresses how well each risk adjuster performs on an absolute basis, whereas, the mean absolute prediction error does not. For example, a CPM of 20% means that the risk adjuster explains 20% of the variation in cost, which is generally viewed as good performance. However, since the mean absolute prediction error is not expressed on a standardized scale, it, by itself, tells us little or nothing about the performance of a model. For example, a mean absolute prediction error of $2,000 could correspond to a model that explains 1% of the variation or could correspond to a model that explains 20% of variation. It is not possible to determine which might be the case without further information.

Generalized CPM

The generalized formula for the CPM measure is:

$$CPM^x = 1 - (\sum_i | a_i - \hat{a}_i |^x) / (\sum_i | a_i - \bar{a} |^x)$$

Where:

a_i = actual claim dollars for person i
$\hat{a}_i$ = predicted claim dollars for person i
$\bar{a}$ = mean of the actual claim dollars
x = power factor (x=1 for the standard CPM measure)
i goes from 1 to n, where n is the number of people

When x is set equal to 1, CPM^x is the same as the CPM measure defined above. When x is set equal to 2, CPM^x is the same as R-squared.

Some researchers argue that the importance of a prediction error grows more rapidly than a linear function of the size of the error i.e., an error that is twice as big is more than twice as serious. Accordingly, some researchers advocate R-squared since, in essence,

it weights large errors much more heavily than small errors. The problem is that large prediction errors can end up dominating the calculation of R-squared. As a result, significant improvements in the predictive accuracy for people with small or medium size claims might have little or no impact on the R-squared measure.

With the generalized CPM measure the user can decide, through the selection of the power factor, how much extra weight, if any, to apply to the larger errors. For example, one might decide to use a power factor of 1.2. This will weight the larger prediction errors more heavily, but the resulting measure is less likely to be dominated by a few large claims, as can occur with R-squared. Similarly, if the user wanted to underweight the larger errors, for some of the same reasons that claims are truncated, the user could select a power factor of slightly less than 1, for example 0.9.

The Super Generalized CPM

The super generalized CPM is:

$$SCPM^x = 1 - (\sum_i | a_i - \hat{a}_i |^x \times w_i) / (\sum_i | a_i - \bar{a} |^x \times v_i)$$

Where:

a_i = actual claim dollars for person i
$\hat{a}_i$ = predicted claim dollars for person i
$\bar{a}$ = mean of the actual claim dollars
w_i = a set of weights for the prediction errors
v_i = a set of weights for the deviations from average
i goes from 1 to n, where n is the number of people

Note that when the weights are set as $w_i = | a_i - \hat{a}_i |^{x-1}$ and $v_i = | a_i - \bar{a} |^{x-1}$, then $SCPM^x$ is the same as CPM^x. In CPM^x, if x is other than 1, then the weights used in the numerator differ from the weights used in the denominator. An alternative approach would be to define a set of weights that are the same for both the numerator and denominator. (The researchers have not yet explored the implications of such an approach.) If $w_i = v_i$, then $SCPM^x$ will still have the desirable property that the measure equals 0 if the model predicts the average claim amount for each person.

Illustration of Sensitivity of R-Squared to Large Prediction Errors

The following example is intended to illustrate the impact of large prediction errors on R-squared and CPM. In order to make the results more evident, the example is a simplified scenario.

Example

Suppose that you have a group of 10,000 members with actual and predicted claim dollars as shown in Table 8.1. In Table 8.1, the members are put into one of four groups (low, medium, high, and very high) based on the amount of medical claims dollars. Also, suppose that a risk adjuster has predicted claims for each member as shown in the table. The last column of the table shows the prediction error for each member.

Table 8.1

Claim Size	Number of Members	Actual Claims per member (in 000's)	Actual Claims (in 000's)	Predicted Claims per member (in 000's)	Predicted Claims (in 000's)	Prediction Error per member (in 000's)
Low	8,000	.400	3,200	1.024	8,192	-.624
Medium	1,900	4.000	7,600	3.000	5,700	1.000
High	99	28.000	2,772	6.500	644	21.500
Very High	1	1,000.000	1,000	40.000	40	960.000
Total	10,000	1.457	14,572	1.458	14,576	

Table 8.2 shows the components of the absolute prediction error (which is the basis of CPM) and the square prediction error (which is the basis of R-squared).

Table 8.2

Claim Size	Absolute Value of Prediction Error per member (in 000's)	Absolute Value of Prediction Error (in 000's)	% of Total Absolute Prediction Error	Square of Prediction Error per member (in 000^2)	Square of Prediction Error (in 000^2)	% of Total Square Prediction Error
Low	.624	4,992	50.0%	.389	3,112	.3%
Medium	1.000	1,900	19.0%	1.000	1,900	.2%
High	21.500	2,129	21.3%	462.250	45,763	4.7%
Very High	960.000	960	9.6%	921,600.000	921,600	94.8%
Total		9,981	100.0%		972,375	100.0%

In this example, the total *square* prediction error (which is 972,375) is dominated by the prediction error on one claim, the one member with the $1,000,000 claim. Although this claim represents only 6.9% of the overall claim dollars, it counts for 94.8% of the overall prediction error. For the total *absolute* prediction error (which is 9,981) this one large claim accounts for only 9.6% of the overall prediction error.

As mentioned above, R-squared (which is derived from the total square prediction error) is overly sensitive to the prediction error for large claims. The corollary to this statement would be that R-squared is unduly insensitive to improvement in predictions for small or medium size claims. To illustrate this point, consider two alternative scenarios: (A) being able to decrease the prediction error by $1,000 for the one member with the very high claim, versus (B) being able to perfectly predict the claims for each of the 1,900 people with medium size claim amounts (i.e. decreasing the prediction error by $1,000 for each of these 1,900 people). It would seem that most users of risk adjusters would consider scenario B to be a much bigger improvement in predictive performance than scenario A. However, if we calculate the impact on the total square prediction error, we find that scenario A shows a bigger improvement than scenario B. In particular, the total square prediction error decreases by 1,919 in scenario A while the decrease in scenario B is only 1,900.

Impact of New Measure on Model Fitting

In calibrating the models in this study, a linear regression model was used which minimizes the mean square prediction error. Accordingly, the R-squared measure corresponds to the way the risk weights are calibrated. Some researchers might then argue that R-squared is the most appropriate measure, since it corresponds to the way the risk weights were determined. The researchers for this study believe that one should first define what is believed to be the most appropriate measure (or measures) of predictive accuracy and let that drive the way the model is calibrated, rather than vice versa.

If CPM is adopted as a new standard in measuring predictive accuracy, this might impact the way models are calibrated. In particular, calibration methods that attempt to minimize the mean absolute prediction error, rather than mean square prediction error, might lead to further improvements in model performance. It might also be surmised that methods that try to minimize the mean absolute prediction error might lead to more stable and reasonable risk weights since such methods are not impacted as much by a few large claims.

Comparison of Numerical Results: R-Squared and CPM

The following provides a detailed comparison of the numerical measures of predictive accuracy using R-squared and Cumming's Prediction Measure (CPM). This information is intended to help readers get more comfortable with this new measure by benchmarking it against an existing, commonly used measure.

Prospective Model – Offered Weights

Table 8.3 summarizes R-squared and CPM for each risk adjuster when used for a prospective application with the offered weights. A higher value indicates better predictive accuracy. The ACG method is not included in these tables since it does not come with offered weights.

Table 8.3: Summary of R-squared and CPM – Prospective Model with Offered Weights

Risk Adjuster	Type of Risk Adjuster	R-Squared with claims truncated at:			CPM 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	.127	.128	.127
DCG	Diag	.195	.180	.143	.172	.172	.169
Medicaid Rx	Rx	.116	.098	.071	.124	.123	.123
RxGroups	Rx	.206	.181	.134	.202	.200	.197
RxRisk	Rx	.175	.148	.111	.172	.168	.164
ERG	Diag+Rx	.218	.193	.146	.219	.216	.209

As can be seen in Table 8.3, CPM is similar in magnitude to R-squared. However, rankings of performance based on CPM differ slightly from rankings based on R-squared. (As mentioned above, CPM provides the same performance rankings as the mean absolute prediction error.) In general, the pharmacy-based risk adjusters rank slightly higher when using CPM than when using R-squared.

The CPM measure tends to be less sensitive to the level of claim truncation. For example, the CPM measure varies between 20.9% and 21.9% for the ERGs, depending on the level of claim truncation. Whereas, the R-squared measure varies between 14.6% and 21.8% for the ERGs.

Since R-squared is overly sensitive to large claims, many researchers truncate the claim dollars. To the extent that different studies use different levels of claim truncation, it makes the results of the studies more difficult to compare. The sensitivity of R-squared to the level of claim truncation also leads to a variety of opinions regarding what is the "right" or "optimal" level of claim truncation for analyzing predictive performance.

Prospective Model – Recalibrated Weights

Table 8.4 summarizes R-squared and Cumming's Prediction Measure (CPM) for each risk adjuster when used for a prospective application with recalibrated weights.

Table 8.4: Summary of R-squared and CPM – Prospective Model with Recalibrated Weights

Risk Adjuster	Type of Risk Adjuster	R-Squared with claims truncated at:			CPM with claims truncated at:		
		$50,000	$100,000	None	$50,000	$100,000	None
ACG	Diag	.172	.140	.099	.179	.171	.167
CDPS	Diag	.208	.186	.149	.190	.183	.178
DCG	Diag	.224	.198	.154	.208	.198	.190
Medicaid Rx	Rx	.200	.165	.119	.196	.186	.180
RxGroups	Rx	.222	.185	.132	.216	.205	.198
RxRisk	Rx	.188	.154	.111	.184	.175	.169
ERG	Diag+Rx	.230	.197	.148	.228	.218	.210

Comparison of Results with and without Recalibration

Table 8.5 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 and CPM between the prospective model with the recalibrated weights and the prospective model with the offered weights.

Table 8.5: 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:			Increase in CPM due to Recalibration with claims truncated at:		
		$50,000	$100,000	None	$50,000	$100,000	None
ACG	Diag	NA	NA	NA	NA	NA	NA
CDPS	Diag	.074	.062	.046	.063	.055	.052
DCG	Diag	.029	.018	.012	.036	.026	.021
Medicaid Rx	Rx	.084	.067	.047	.072	.063	.058
RxGroups	Rx	.015	.004	-.001	.014	.005	.000
RxRisk	Rx	.014	.005	.001	.012	.007	.005
ERG	Diag+Rx	.012	.003	.002	.009	.002	.001

The increase in performance as measured by R-squared is fairly consistent with the increase in performance as measured by CPM.

Concurrent Model – Recalibrated Weights

Table 8.6 summarizes R-squared and Cumming's Prediction Measure (CPM) for each risk adjuster when used for a concurrent application with the recalibrated weights.

Table 8.6: Summary of R-squared and CPM – Concurrent Model with Recalibrated Weights

Risk Adjuster	Type of Risk Adjuster	R-Squared with claims truncated at:			CPM with claims truncated at:		
		$50,000	$100,000	None	$50,000	$100,000	None
ACG	Diag	.429	.376	.282	.381	.369	.360
CDPS	Diag	.440	.418	.355	.343	.330	.317
DCG	Diag	.564	.547	.466	.419	.405	.385
Medicaid Rx	Rx	.372	.328	.244	.308	.291	.275
RxGroups	Rx	.420	.376	.279	.347	.327	.307
RxRisk	Rx	.339	.292	.213	.282	.268	.257
ERG	Diag+Rx	.474	.427	.347	.400	.376	.354

As can be seen in Table 8.6, CPM is similar in magnitude to R-squared. However, CPM tends to be more stable as the level of claim truncation is changed. Except for CDPS and DCGs, CPM is sometimes higher and sometimes lower than R-squared. For CDPS and DCGs, R-squared is always higher than CPM for the levels of claim truncation used in this study.

As can be seen in Table 8.6, whether based on R-squared or CPM, the diagnosis-based models outperform the pharmacy-based models when used for concurrent risk assessment.

Comparison of Prospective and Concurrent Results

Table 8.7 compares the performance of the prospective and concurrent risk adjustment models with recalibrated risk weights. Table 8.7 compares performance as measured by R-squared and CPM.

Table 8.7: Comparison of Performance of Prospective and Concurrent Risk Adjustment Models - With Recalibration of Risk Weights

Risk Adjuster	Type of Risk Adjuster	R-Squared with claims truncated at $100,000 for:		CPM with claims truncated at $100,000 for:	
		Prospective Model	Concurrent Model	Prospective Model	Concurrent Model
ACG	Diag	.140	.376	.171	.369
CDPS	Diag	.186	.418	.183	.330
DCG	Diag	.198	.547	.198	.405
Medicaid Rx	Rx	.165	.328	.186	.291
RxGroups	Rx	.185	.376	.205	.327
RxRisk	Rx	.154	.292	.175	.268
ERG	Diag+Rx	.197	.427	.218	.376

As can be seen from Table 8.7, whether based on R-squared or CPM, the concurrent models significantly outperform the prospective models.

Table 8.8 shows the increase in performance between the prospective and concurrent model. In particular, the table shows the increase in R-squared and CPM between the concurrent model and the prospective model with recalibrated weights.

Table 8.8: Increase in Performance between Prospective and Concurrent Model

Risk Adjuster	Type of Risk Adjuster	Increase in R-Squared with claims truncated at:			Increase in CPM with claims truncated at:		
		$50,000	$100,000	None	$50,000	$100,000	None
ACG	Diag	.258	.236	.183	.202	.198	.193
CDPS	Diag	.232	.232	.207	.153	.147	.139
DCG	Diag	.341	.349	.311	.211	.206	.195
Medicaid Rx	Rx	.173	.164	.126	.113	.105	.095
RxGroups	Rx	.198	.191	.147	.131	.121	.110
RxRisk	Rx	.151	.138	.102	.098	.094	.088
ERG	Diag+Rx	.245	.230	.199	.172	.158	.144

As can be seen in Table 8.8, the increase in performance as measured by CPM is slightly smaller than the increase in performance as measured by R-squared. The increase in performance as measured by CPM tends to be more stable as the level of claim truncation is changed. For example, for ACGs, the increase in CPM only varies from .193 to .202 depending on the level of claim truncation, whereas the increase in R-squared varies from .183 to .258.

Section VIII. Recommendations for Follow-up Studies

This section summarizes recommendations for follow-up studies. These are studies that may build upon the research performed here.

Recommendations for follow-up studies:

- Analyze other risk adjusters/predictive models that are common in the marketplace or actively being marketed. This might include 3M's CRGs, the Ingenix predictive models, and the Medicare selected diagnoses models.
- Make the analysis more realistic by incorporating: (a) claim lag differences between diagnosis data and pharmacy data, (b) population turnover, and (c) time lag between the risk assessment period and the payment adjustment period.
- Examine results for "real" groups of members. This might include analyzing results by employer group and benefit option (e.g., HMO vs. PPO, low deductible vs. high deductible).
- Compare the risk adjusters included in this study with predictive models based on measures of prior use.
- Analyze the possible increase in performance due to refinement of the risk weights for a given population. The refinements might include smoothing, blending, and removing non-statistically significant variables.
- Analyze results for other types of populations, such as, Medicaid and Medicare populations.
- Analyze the impact on the ERG results of using only diagnosis data or only diagnosis plus pharmacy data. (The ERGs, as presented in this study, use diagnosis codes, pharmacy data, and a limited number of surgical procedure codes.)
- Analyze results using base year, rather than prediction year, claim dollars to define non-random groups.
- Compare the consistency of pharmacy and diagnosis based models in identifying people with a particular type of medical condition. This might also include analysis of the persistency of certain chronic conditions when defined by diagnosis codes and/or pharmacy codes.
- Analyze the increase in performance that might be possible due to using alternative methods of model fitting. Specifically, the impact of using methods that try to minimize the mean absolute prediction error as opposed to methods that minimize the mean square prediction error.
- Analyze the impact on predictive performance of using more than 12 months of data in the base period.

References

(1) Ash AS, Ellis RP, Pope GC, Ayanian JZ, Bates DW, Burstin H et al. Using diagnoses to describe populations and predict costs. Health Care Finance Rev 2000; 21(3):7-28.

(2) Ash AS, Byrne-Logan S. How well do models work? Predicting health care costs. Proceedings of the Section on Statistics in Epidemiology 1998; 42-49.

(3) Bowen B. The practice of risk adjustment. Inquiry 1995; 32(1):33-40.

(4) Dunn DL, Rosenblatt A, Taira D, Latimer E, Bertko J, Stoiber T et al. A Comparative Analysis of Methods of Health Risk Assessment. 1996. Schaumburg, Illinois, Society of Actuaries. SOA Monograph M-HB96-1.

(5) Gilmer T, Kronick R, Fishman P, Gardner LB. The Medicaid Rx Model: Pharmacy-Based Risk Adjustment for Public Programs. Med Care 2001; 39(11):1188-1202.

(6) Ingber MJ. The current state of risk adjustment technology for capitation. J Ambulatory Care Manage 1998; 21(4):1-28.

(7) Knutson D, Fowles J. Implementing Health-based Risk Adjustment: Lessons from the Field. Retrieved on June 23, 2001 from Risk Adjustment Implementation and Impact Study Website, Park Nicollet Institute: http://www.riskadjustment.org.

(8) Kronick R, Dreyfus T, Lee L, Zhou Z. Diagnostic risk adjustment for Medicaid: the disability payment system. Health Care Finance Rev 1996; 17(3):7-33.

(9) Kronick R, Gilmer T, Dreyfus T, Lee L. Improving health-based payment for Medicaid Beneficiaries: CDPS. Health Care Finance Rev 2000; 21(3):29-64.

(10) Weiner JP, Starfield BH, Lieberman RN. Johns Hopkins Ambulatory Care Groups (ACGs). A case-mix system for UR, QA and capitation adjustment. HMO Pract 1992; 6(1):13-19.

(11) Joint Purchasers' Risk Assessment Project: Phase I. Final Report. 9-22-200. William M. Mercer, Inc.; Reden & Anders, Ltd.; DxCG, Inc.

(12) Final Report: A Comparison of Alternative Approaches to Risk Measurement". 1994. Park Nicollet Medical Foundation Health Research Center; The Johns Hopkins University Health Services Research Center.

(13) Summary of Key Findings "A Comparison of Alternative Approaches to Risk Measurement". 1994. Park Nicollet Medical Foundation Health Research Center; The Johns Hopkins University Health Services Research Center

May 24, 2002

Appendix A: Selected Readings

(1) Ash AS, Ellis RP, Pope GC, Ayanian JZ, Bates DW, Burstin H et al. Using diagnoses to describe populations and predict costs. Health Care Finance Rev 2000; 21(3):7-28.

(2) Bertko J, Hunt S. Case study: the Health Insurance Plan of California. Inquiry 1998; 35(2):148-153.

(3) Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. J Health Econ 1999; 18(2):153-171.

(4) Bowen B. The practice of risk adjustment. Inquiry 1995; 32(1):33-40.

(5) Dudley RA, Bowers LV, Luft HS. Reconciling quality measurement with financial risk adjustment in health plans. Jt Comm J Qual Improv 2000; 26(3):137-146.

(6) Dunn DL. Applications of health risk adjustment: what can be learned from experience to date? Inquiry 1998; 35(2):132-147.

(7) Ellis RP, Ash A. Refinements to the Diagnostic Cost Group (DCG) model. Inquiry 1995; 32(4):418-429.

(8) Ellis RP, Pope GC, Iezzoni L, Ayanian JZ, Bates DW, Burstin H et al. Diagnosis-based risk adjustment for Medicare capitation payments. Health Care Finance Rev 1996; 17(3):101-128.

(9) Ettner SL, Frank RG, McGuire TG, Newhouse JP, Notman EH. Risk adjustment of mental health and substance abuse payments. Inquiry 1998; 35(2):223-239.

(10) Ettner SL, Frank RG, McGuire TG, Hermann RC. Risk adjustment alternatives in paying for behavioral health care under Medicaid. Health Serv Res 2001; 36(4):793-811.

(11) FitzHenry F, Shultz EK. Health-risk-assessment tools used to predict costs in defined populations. J Healthc Inf Manag 2000; 14(2):31-57.

(12) Fowles JB, Weiner JP, Knutson D, Fowler E, Tucker AM, Ireland M. Taking health status into account when setting capitation rates: a comparison of risk-adjustment methods. JAMA 1996; 276(16):1316-1321.

(13) Frank RG, Glazer J, McGuire TG. Measuring adverse selection in managed health care. J Health Econ 2000; 19(6):829-854.

(14) Gilmer T, Kronick R, Fishman P, Gardner LB. The Medicaid Rx Model: Pharmacy-Based Risk Adjustment for Public Programs. Med Care 2001; 39(11):1188-1202.

(15) Goldfield N, Averill R, Eisenhandler J. Payment and provider profiling of episodes of illness of clinical illnesses involving rehabilitation. J Outcome Meas 2000; 4(3):706-720.

May 24, 2002

(16) Greenwald LM, Esposito A, Ingber MJ, Levy JM. Risk Adjustment for the Medicare program: lessons learned from research and demonstrations. Inquiry 1998; 35(2):193-209.

(17) Hendryx MS, Teague GB. Comparing alternative risk-adjustment models. J Behav Health Serv Res 2001; 28(3):247-257.

(18) Hendryx MS, Beigel A, Doucette A. Introduction: risk-adjustment issues in mental health services. J Behav Health Serv Res 2001; 28(3):225-234.

(19) Hornbrook MC, Goodman MJ. Chronic disease, functional health status, and demographics: a multi-dimensional approach to risk adjustment. Health Serv Res 1996; 31(3):283-307.

(20) Iezzoni LI. An introduction to risk adjustment. Am J Med Qual 1996; 11(1):S8-11.

(21) Ingber MJ. The current state of risk adjustment technology for capitation. J Ambulatory Care Manage 1998; 21(4):1-28.

(22) Ingber MJ. Implementation of risk adjustment for Medicare. Health Care Financ Rev 2000; 21(3):119-126.

(23) Kassirer JP, Angell M. Risk adjustment or risk avoidance? N Engl J Med 1998; 339(26):1925-1926.

(24) Knutson D. Case study: the Minneapolis Buyers Health Care Action Group. Inquiry 1998; 35(2):171-177.

(25) Knutson D, Fowles J. Implementing Health-based Risk Adjustment: Lessons from the Field. Retrieved June 23, 2001, from Risk Adjustment Implementation and Impact Study Website: http://www.riskadjustment.org

(26) Kronick R, Dreyfus T, Lee L, Zhou Z. Diagnostic risk adjustment for Medicaid: the disability payment system. Health Care Financ Rev 1996; 17(3):7-33.

(27) Kuttner R. The risk-adjustment debate. N Engl J Med 1998; 339(26):1952-1956.

(28) Lamers LM. Risk-adjusted capitation based on the Diagnostic Cost Group Model: an empirical evaluation with health survey information. Health Serv Res 1999; 33(6):1727-1744.

(29) Lee C, Rogal D. A Key to Changing Incentives in the Health Insurance Market. 3-26. 1997. Washington, D.C., Alpha Center.
Ref Type: Report

(30) Luft HS. Potential methods to reduce risk selection and its effects. Inquiry 1995; 32(1):23-32.

(31) Luft HS. Modifying managed competition to address cost and quality. Health Aff (Millwood) 1996; 15(1):23-38.

May 24, 2002

(32) Madden CW, Mackay BP, Skillman SM, Ciol M, Diehr PK. Risk adjusting capitation: applications in employed and disabled populations. Health Care Manag Sci 2000; 3(2):101-109.

(33) Maguire AM, Powe NR, Starfield B, Andrews J, Weiner JP, Anderson GF. "Carving out" conditions from global capitation rates: protecting high-cost patients, physicians, and health plans in a managed care environment. Am J Manag Care 1998; 4(6):797-806.

(34) Majeed A, Bindman AB, Weiner JP. Use of risk adjustment in setting budgets and measuring performance in primary care II: advantages, disadvantages, and practicalities. BMJ 2001; 323(7313):607-610.

(35) Meenan RT, O'Keeffe-Rosetti C, Hornbrook MC, Bachman DJ, Goodman MJ, Fishman PA et al. The sensitivity and specificity of forecasting high-cost users of medical care. Med Care 1999; 37(8):815-823.

(36) Palsbo S., Preston BJ. Implementing Risk Assessment and Risk Adjustment for People with Disabilities in State Programs: Six Case Studies. 2001. NRH Center for Health and Disability Research.

(37) Park Nicollet Medical Foundation Health Research Center, The Johns Hopkins University School of Hygiene and Public Health. Summary of Key Findings "A Comparison of Alternative Approaches to Risk Measurement". 1994.

(38) Park Nicollet Medical Foundation Health Research Center, The Johns Hopkins University Health Services Research Center. Final Report: A Comparison of Alternative Approaches to Risk Measurement". 1994.

(39) Payne SM, Cebul RD, Singer ME, Krishnaswamy J, Gharrity K. Comparison of risk-adjustment systems for the medicaid-eligible disabled population. Med Care 2000; 38(4):422-432.

(40) Pope GC, Ellis RP, Ash AS, Liu CF, Ayanian JZ, Bates DW et al. Principal inpatient diagnostic cost group model for Medicare risk adjustment. Health Care Financ Rev 2000; 21(3):93-118.

(41) Puig-Junoy J. Managing risk selection incentives in health sector reforms. Int J Health Plann Manage 1999; 14(4):287-311.

(42) Reid R., Bogdanovic B., Roos N., Black C., MacWilliam L., Mijanovich T. Do Some Physician Groups See Sicker Patients Than Others? Implications for Primary Care Policy in Manitoba. 2001. Manitoba Centre for Health Policy and Evaluation, Department of Community Health Sciences, University of Manitoba.

(43) Reid RJ, MacWilliam L, Verhulst L, Roos N, Atkinson M. Performance of the ACG case-mix system in two Canadian provinces. Med Care 2001; 39(1):86-99.

(44) Rice C, Nakahata P. Pharmacy-Based Risk Adjustment: A Guidebook for States. Center for Health Care Strategies, Inc. 2001.

May 24, 2002