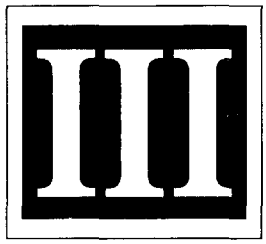


Evaluation and Comparison of Risk Assessment Methods: Predictive Accuracy

The previous research on health risk assessment has provided significant insights into the predictive ability of different methods. However, these studies have some shortcomings in this regard. First, few studies have compared competing models using the same sample of data, making it difficult to assess their relative performance. Second, as a result of an emphasis on searching for an alternative to the AAPCC, most studies have focused on risk assessment for Medicare beneficiaries. Not many have applied these methods to the under-65 population. Finally, a number of the evaluations of predictive accuracy have used risk assessment in either a prospective or retrospective design. Few have explored both approaches using the same data and methods.

This study addresses these shortcomings. In particular, we employ a national, standardized database describing use and expenditures for a large nonelderly population. Further, in assessing predictive accuracy, we apply our methods in a uniform way across all competing models to be tested. Finally, we explore both prospective and retrospective risk assessment for each model.

In this chapter, we describe the methods and data we used in evaluating predictive accuracy. We also present the findings of our investigation and discuss their implications. As described below, the analysis presented here does not address explicitly risk assessment for high-cost individuals. Instead, we present the methods and results of a separate analysis for these individuals in Chapter IV. This chapter also does not discuss in any detail the general considerations beyond predictive accuracy when comparing different methods, including practicality, administrative feasibility, and incentives for efficiency. Those topics are addressed in Chapter VI.

We begin with a discussion of the models we tested.

A. Risk Assessment Models Evaluated

As described in Chapter II, there exist a number of methods for measuring relative health risk, each employing alternative information when assessing risk. For this study, we chose two general categories of these models for investigation: (1) a simple demographic model based on enrollee age and sex and (2) diagnosis-based models that classify enrollees using the ambulatory and inpatient diagnoses recorded on insurance claim forms.

Specifically, we chose eight risk assessment models for study:

- Demographic Model
  - Age and sex
- ACG Models
  - ACGs
  - ADGs, with age and sex
- DCG Models
  - Principal inpatient diagnostic cost groups (PIPDCGs), with age and sex
  - All diagnosis diagnostic cost groups (ADCGs), with age and sex
  - Expanded diagnostic cost groups (EDCGs), with age and sex

- All diagnosis diagnostic cost groups, with high-cost coexisting conditions (ADCGDXs) and age and sex
- Expanded diagnostic cost groups, with high-cost coexisting conditions (EDCGDXs) and age and sex.

1. Age and Sex

Age and sex were measured using 28 separate groups, 14 each for males and females. For each sex, we employed ranges of five years when defining age groups, with the exception of enrollees less than five years of age, where we grouped separately those less than one year. As described below, because the study included data from a nonelderly population, age groups for enrollees 65 and over were not required. Appendix Table B-1 describes the age-sex groups used, including the distribution of enrollees across groups.[1]

2. Ambulatory Care Group Models

Ambulatory care groups (ACGs) are a diagnosis-based measure of expected resource consumption. (Starfield, et al., 1991) Each enrollee is classified into one of 52 ACGs based on all of the ambulatory diagnoses recorded for the individual over a period of time.[2] For many ACGs, gender and broad age groupings are also employed in classifying patients. We employed Version 2.0, July 1993, of the ACG model for the purposes of this study (Johns Hopkins University, 1993).[3]

The first step in assigning a person to an ACG is to identify all unique ICD9 codes (primary or subsidiary ambulatory diagnoses) recorded for that individual. Each unique ICD9 code is then assigned to one of 34 Ambulatory Diagnostic Groups (ADGs). While each diagnosis is assigned to only one ADG, a patient with multiple diagnoses can be assigned to multiple ADGs.[4]

Second, based on an enrollee's age, gender, and his or her mix of ADGs, a single ACG is assigned. Enrollees without recorded diagnoses (including those with no reported expenditures), or without a diagnosis qualifying for assignment to an ADG as described above, are assigned to ACG 52.

We used two ACG models for this study: (1) the ' 'endpoint" ACG model in which each enrollee is assigned to one of 52 ACGs and (2) an ADG model in which a patient can be assigned to zero, one, or more than one, of the 34 ADGs described above. ADGs can be considered intermediate steps, or building blocks, of the ACG model. In this way—given the aggregation involved in moving from a combination of a potentially large number of ADGs for a patient to a single, endpoint ACG—ADG assignments can be considered to be more descriptive of a patient's clinical condition. As a result, the ADG model might be expected to perform better than ACGs in terms of predictive accuracy. We tested the ADG model for this reason.

Finally, given that enrollee age and sex are not used in assigning ADGs, we added these two factors to the ADG model using the 28 age-sex groups described previously. Appendix Tables B-2 and B-3 summarize the ACG and ADG groups used in the study and the distribution of individuals across these groups.

3. DCG Models

Like the ACG models, DCGs are a patient-based measure of expected health care use. Both models use diagnoses to categorize patients into risk groups. However, ACGs and DCGs differ in the criteria used to group patients, as well as the groupings employed.

A number of variations of the DCG model currently exist; however, they all follow a similar logic. Conceptually, DCGs assume that certain diagnoses are predictably associated with higher levels of health care costs.[5] As a first step, all ICD9 diagnostic codes (primary or subsidiary diagnoses) for an individual are identified.[6] Each ICD9 code is then assigned to a single DCG diagnosis (DCGDIAG). A patient with multiple diagnoses can be assigned more than one DCGDIAG.

Next, depending on the model being employed, each DCGDIAG is mapped into a DCG. DCGs are numbered 1, 2, 3, and so on, with a higher number representing higher expected costs associated with the diagnoses included in that DCG.

Finally, an individual is assigned to a single DCG, the individual's highest numbered DCG recorded.

We employed all five models included in Version 3, May 1995, of DCGs for this study (Ellis, et al., 1995). Each of these models is based on the conceptual framework described above. The models differ, however, in the diagnostic information used to assign patients to DCGs and how that information is employed.

a. PIPDCGs

PIPDCGs are based on a person's principal diagnoses from hospital inpatient stays (one principal diagnosis per stay).[7] Diagnoses related to higher expected

costs are assigned to a higher PIPDCG. However, in doing this, not all inpatient stays are eligible for higher DCG assignments. Persons with admissions of low expected expenditures are grouped into the lowest cost DCG. In this way, these cases receive risk weights consistent with those assigned to persons without an inpatient admission. This same type of distinction is made in all of the DCG models.

b. ADCGs

ADCGs are based on all inpatient and ambulatory diagnoses with no distinction made for the source of the diagnosis. In other words, when assigning an ICD9 code to a DCGDIAG, no distinction is made between inpatient and ambulatory diagnoses. For example, an inpatient and outpatient diagnosis for viral hepatitis would both be assigned to ADCG 1.

c. EDCGs

EDCGs are based on all inpatient and ambulatory diagnoses with a distinction made between principal inpatient diagnoses and all other diagnoses. This model can be distinguished from ADCGs in that the same ICD9 diagnosis code can be assigned to a different EDCG depending on whether it is a principal inpatient or other diagnosis. For example, a principal inpatient diagnosis of viral hepatitis is assigned to EDCG 4, while the same diagnosis recorded in an ambulatory setting is assigned to EDCG 2.

d. CDXGs

In an effort to further distinguish high-cost individuals, the most recent version of DCGs also identifies 25 hierarchical CDXGs that can be added to the ADCG and EDCG models. An individual is assigned a CDXG based on any qualifying inpatient or ambulatory diagnosis for the ADCG and EDCG models. However, individuals may not be assigned a particular CDXG if that CDXG contains the same highest cost DCGDIAG that was used to assign them to their final ADCG or EDCG. (This avoids double-counting diagnoses.) Further, hierarchies are ther, hierarchies are established to avoid overlap across conditions. For example, persons with CDXG 2 (secondary and disseminated cancers) cannot be assigned CDXG 3 (high-cost cancers) or CDXG 4 (moderate cost cancers).

In contrast to the PIPDCG, ADCG, and EDCG assignments, an enrollee can be assigned to more than one CDXG. The CDXGs are not used with the PIPDCG model.

We incorporated the CDXGs into both the ADCG and EDCG models. We describe the resulting models in this report as ADCGDX and EDCGDX, respectively.

Finally, given that age and sex are not used in assigning DCGs and, that, in a non-elderly population such as that described by our data, a significant percentage of enrollees would have no inpatient admission for a year, we added these two factors to the five DCG models tested using the 28 age-sex groups described previously. Tables B-4 and B-5 in the Appendix list the DCG model groupings employed in the study and the distribution of enrollees across groups.[8] Table 1 summarizes the eight models evaluated in the study.

B. Study Data

The SOA made available a comprehensive, standardized database for use in the project. The data included information submitted from nine national and regional health care carriers and described annual health use and expenditures for a large number of individuals from plans encompassing a range of care management approaches and deductible levels. In addition, we supplemented these data by obtaining information from a network HMO plan for use in the study.

The data assembled by the SOA derive from two major data collection efforts. First, the SOA obtained the data from six carriers participating in an earlier study on risk adjustment conducted by the HIAA (HIAA, 1994). Second, in an independent effort, the SOA secured data from five carriers. It was later determined that two of the carriers had contributed data to both the HIAA study and subsequently to the SOA. We excluded the HIAA data for these carriers and used only that information collected by the SOA. The number of records included in the initial database from the ten carriers exceeded six million.

The study database covered a period of two consecutive years, 1991 and 1992, and was constructed to support analyses of risk assessment using both demographic and diagnosis-based models and both retrospective and prospective applications. For each insured individual and year, a record in the data file contains demographic information including a unique ID, age, sex, zip code, and insurance plan type (indemnity, HMO, PPO, level of deductible, etc.). Both inpatient and ambulatory clinical information are also recorded. For each admission, carriers were asked to supply the principal diagnosis code (ICD9), principal procedure code (ICD9), days of stay, hospital charges, and the

TABLE 1
SUMMARY OF ASSESSMENT MODELS TESTED

Model	Number of Groups	Diagnoses Used	Notes
Age-Sex	28 Age-Sex	None	
ACG	52 ACGs	Ambulatory	Mutually exclusive groups
ADG	34 ADGs 28 Age-Sex	Ambulatory	Individuals can have multiple ADGs
PIPDCG	12 PIPDCGs 28 Age-Sex	Inpatient	Mutually exclusive groups
ADCG	12 ADCGs 28 Age-Sex	Inpatient Ambulatory	Mutually exclusive groups No distinction between inpatient and ambulatory diagnoses
EDCG	12 EDCGs 28 Age-Sex	Inpatient Ambulatory	Mutually exclusive groups Distinction between principal inpatient and other diagnoses
ADCGDX	12 ADCGs 28 Age-Sex 25 CDXGs	Inpatient Ambulatory	ADCGs are mutually exclusive groups CDCXGs added Individual can have more than one CDXG No distinction between inpatient and ambulatory diagnoses
EDCGDX	12 EDCGs 28 Age-Sex 25 CDXGs	Inpatient Ambulatory	EDCGs are mutually exclusive groups CDXGs added Individual can have more than one CDXG Distinction between principal inpatient and other diagnoses

assigned DRG. For each ambulatory encounter, the principal outpatient diagnosis code is recorded.[9]

The expense data for an individual were aggregated by type of service. For each major category of service, plans were asked to submit total billed charges prior to any contractual limitations or discounts. For hospital services, total inpatient charges and hospital outpatient charges for the year were recorded. For physician and professional services, total surgical charges and total medical charges, aggregated using ranges of common procedural technology (CPT) codes, are available. (Radiology and pathology charges are the exception and are grouped under an "All Other" category.) For prescription drugs (out-of-hospital prescription drugs) the total charge amount is included. Charges not included in the above categories are recorded in an "All Other" category. Finally, the amount of total charges for the insured over the calendar year (equaling the total of the charges described above) is included. Total annual charges served as the dependent variable in all our analyses. Table 2 describes each of the variables included in the study data file.

For all individuals, charges are before deductibles and copayments. Any costs incurred for services covered where a claim was not submitted due to a deductible or any other out-of-pocket costs are not measured.

Finally, carriers were asked to supply data for insured individuals in a plan for the entire year (1991 or 1992) and include all enrollees, whether they incurred claims or not. When family coverage is applicable, separate records were to be submitted for each dependent.[10]

1. Constructing Pools of Data for Study

Carriers submitted data describing individuals from a range of health plan types. For the purpose of the analyses described in this chapter, we segmented the study data into *pools* based on carrier and plan type, as measured by care management approach and level of deductible. As shown in Table 2, a number of health care management approaches were delineated, including indemnity with utilization review, indemnity without utilization review, gatekeeper and non-gatekeeper PPOs, and different HMO models. Further, three levels of deductibles were defined: less than $500, $500 to $2,000, and greater than $2,000.[11]

In initially constructing pools, we made three observations. First, a significant number of pools were available in the study data. Including all ten carriers, more than 40 pools were identified, more than enough to support our analyses. Second, although many of the pools comprised data from a large number of individuals (for example, greater than 15,000 enrollees for each year), an equal number were small in size (for example, less than 3,000 enrollees). Such small pools would likely

TABLE 2
DESCRIPTION OF VARIABLES IN STUDY DATABASE

	Demographic Data
ID#	Identifies each unique individual or insured.
Year	The calendar year during which services were performed.
Age	Age of the individual for the year under study.
Sex	Sex of the individual.
Zip Code	5-digit zip code of the individual.
Plan Type-1	A 1-character code to identify the type of health plan (see below).
Plan Type-2	A 1-character code to identify the type of underwriting and deductible amounts (see below).
Admissions	The total number of inpatient admissions for the year.
	Charges
Surgical	Professional surgical charges for the following CPT-4 procedure codes: 10000 to 16999, 92970 to 92990, and 93501 to 93562.
Medical	Professional medical charges for all other CPT-4 procedure codes except for codes 70000 to 89999. Codes 70000 to 89999 are radiology and pathology charges, which are included in the All Other category.
Inpatient	Inpatient hospital facility charges associated with an admission to a hospital during the calendar year including any preadmission testing. This equals the sum across all hospital admissions reported for this individual.
Outpatient	Outpatient hospital facility charges. Charges for preadmission testing when a patient is not subsequently admitted to the hospital would be outpatient charges.
Drug	Prescription drug charges (out-of-hospital prescription drugs).
All Other	All other charges not included above.
Total	Total charges equal to the sum of all charges shown above.
	For Each Hospital Admission
Diagnosis	5-digit alpha-numeric primary diagnosis code (ICD-9)
DRG	Diagnostic related grouping
Hospital Charges	Total inpatient hospital charges for the admission.
Surgical Procedure	4-digit alpha-numeric ICD-9 surgical procedure code for the primary surgical procedure performed during the admission, if any.
Days	Length of stay for the admission.
	For Each Outpatient/Medical Visit
Diagnosis	5-digit alpha-numeric primary diagnosis code (ICD-9)
	Available Values of Plan Type Code #1
A.	Indemnity
B.	Indemnity with utilization management
C.	Nongatekeeper PPO
D.	Gatekeeper PPO
E.	Exclusive provider organization (EPO)
F.	Point-of-service HMO
G.	Individual practice association (IPA) model HMO
H.	Network model HMO
I.	Group model HMO
J.	Staff model HMO
K.	Mixed model HMO
	Available Values of Plan Type Code #2
1.	Deductibles of $500 or less
2.	Deductibles of over $500 and under $2,000
3.	Deductibles of $2,000 and over

provide less precise results, particularly for the diagnosis-based risk assessment models (ACGs and DCGs) where only a small where only a small percentage of all enrollees are expected to be assigned to some groups. Third, the majority of the smaller pools were for plan types with deductible levels greater than $500.

Based on these observations, we excluded from further analysis all pools with less than 3,000 enrollees for a year and pools with deductible levels greater than $500. The decision to exclude higher deductible pools is further supported by the greater possibility of unreported expenses for individuals in these plans. As noted above, any expenses incurred for services covered where a claim was not submitted due to a deductible or any other out-of-pocket costs are not measured.

2. *Preparation of Study Data*

The preparation of the study data for analysis involved five major steps:

- Assessment of data validity
- Editing of data to exclude invalid or unusable records
- Adjustment of expenditures for differences across geographic areas
- Adjustment of expenditures for differences across years
- Creation of working files for the analysis.

We discuss each separately below. Figure 1 provides a description of the data preparation.

a. Assessment of Data Validity

We conducted a number of analyses to determine data validity. These involved simple checks such as valid entries for key variables (charges, ICD9 codes, age, sex, and so on) and the consistency of entries for related variables (the consistency of the sum of the component charges with total charges, admissions or ICD9 codings for individuals with no reported expenditures, and the like). We also performed more sophisticated data checks including the consistency of ICD9 codings with the age and sex of the individual; the distribution of expenditures for a pool (that is, mean, standard deviation, percentiles and extreme values); and the frequency of ICD9 codings for individuals with different ranges of total expenditures.

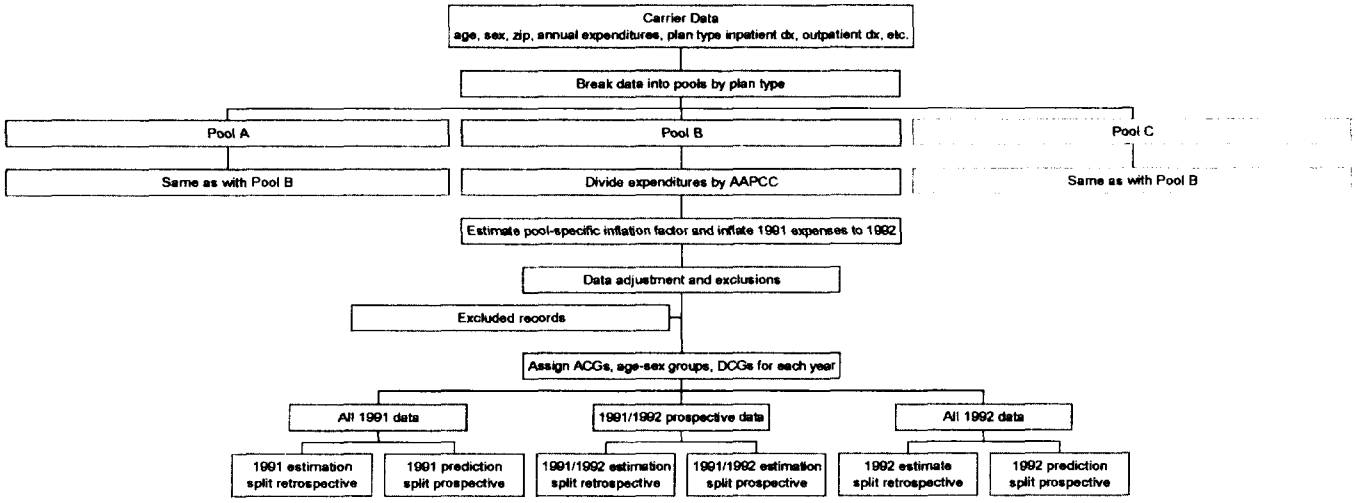
b. Editing of Study Data

Based on the validation analyses, we identified a number of potential problems with the data available for the study. Where appropriate, we contacted the carriers submitting the data to verify each problem and identify possible solutions. In many cases, the problem was resolved, with the carrier in some instances submitting further data. In other cases, a solution was not possible given either the time constraints of the project, the lack of alternative data, or the resources required of the carrier to remedy the problem.

We identified significant data validity problems for three carriers, none of which could be remedied for this study. These problems included: (1) insufficient or no diagnostic detail (ICD9 codings) for the large majority of individuals, even those with relatively high expenditures (one carrier); (2) expenses and data reported for subscribers rather than enrollees, each record thus describing use and expenses for potentially multiple individuals (one carrier); and (3) extreme and inconsistent expenditures for a large number of records (one carrier). We excluded the data from these carriers from the study.

For four additional carriers, we identified less significant problems that were resolved in a manner sufficient to allow the data to be used in the analysis. These problems included: (1) extreme expenditures for a large number of individuals (one carrier) and (2) data missing for those enrollees without claims (three carriers). In the first instance, we determined that the data responsible

FIGURE 1
DESCRIPTION OF DATA PREPARATION



for the extreme values were from a single pool. We excluded that pool from the analysis.

In the second instance, with the assistance of the actuarial advisors to the project, we developed an algorithm to estimate the number of nonclaimants, pool-by-pool. To do this, we examined the ratios of nonclaimants to claimants by age and sex group using data from other carriers with similar health plan types and complete records. Using these ratios and the number of claimants in each age and sex group, we imputed the number of nonclaimants for each of the problem pools. For example, we found that for indemnity pools, on average, 53% of male enrollees ages 25 to 34 were claimants in a year. We applied this assumption to indemnity pools with incomplete data to compute the number of nonclaimants. This imputation is described in greater detail in Appendix B.

In addition to the data problems identified at the carrier and pool level, we also noted invalid or insufficient data for a small number of individual records within the pools included for study. These problems included mostly invalid or missing age and sex, negative expenses, significant expenses with no diagnostic information, and significant diagnostic information without expenses. As a whole, these records represented only a small percentage (less than 1%) of all observations and were excluded from the study.

Finally, the available data for most pools included a small number of enrollees with age greater than 65. Given the potential biases introduced with the data for these individuals (for example, competing Medicare coverage, a select sample of the elderly, and so on), there was a consensus that these records should be excluded.

c. Adjusting Expenditures for Differences across Geographic Areas

In addition to health risk, medical costs can vary among individuals for a number of reasons including input prices facing providers, provider fees facing insurers (net of input price differences), regional practice patterns, and the level of health care management. Conducting the analyses pool by pool and reporting the results separately for each plan type will address the issue of differences in care management to some extent. However, differences in input prices, fees, and practice approaches across geographic areas within a pool should be controlled for in order to assess appropriately the predictive accuracy of different risk adjustment methods. This is not to say that health plans should be rewarded for paying higher procedure fees or encouraging more aggressive care. However, it is important to control for these differences when using the type of data and methods employed in this study.

To adjust expenditures for systematic differences across geographic areas, we employed the AAPCC used by the HCFA to adjust Medicare HMO premiums. As described in Chapter II, the AAPCC represents the relative health expenditures for Medicare enrollees in a county and captures both systematic differences in prices across areas and differences in volume or service intensity per beneficiary.

The AAPCC index we used is the ratio of Medicare's 1995 AAPCC rate (Part A plus Part B) for a county divided by the national rate. Since the AAPCC is based on a five-year rolling average, and a two-year lag exists between payment year and data reporting year, the 1995 AAPCC represents average relative Medicare beneficiary expenditures (non-HMO beneficiaries) for the years 1989 through 1993. We divided all expenditure variables by this index to place them on a common geographic basis.[12]

d. Adjusting Expenditures for Differences across Years

In addition to differences across geographic areas, health expenditures can vary over time due to differences in prices and the volume and intensity of services provided.

To adjust for differences in expenditures across years for each pool, we used ordinary least squares regression with total expenditures as the dependent variable and as the independent variables, the dummy variables for the 28 age and sex groupings described above and a dummy variable describing the year each observation represents (1991 or 1992). The estimated coefficient for the year variable provided an estimate of the systematic difference in medical expenses between the two years for the particular pool. We used this coefficient to adjust, pool by pool, all 1991 expenditures to 1992 levels. These inflation factors varied from 6 to 18% across all pools and were on average approximately 11%.[13]

e. Creation of Working Files for the Analysis

We chose a split-half design for estimating each model and using the estimates to make predictions. Specifically, we estimated a model's parameters using half of the data available for a pool, randomly selected. We then applied these estimates to the second half of the data to assess predictive accuracy. A split-half design avoids the potential problem of overfitting, where the experience of a group is used in predicting its own experience.[14] As a result, we created two data files for

each analysis: an estimation sample and a prediction sample.

For each pool, we created a maximum of eight working files. Two files included all individuals for each year: (a) all individuals for 1991 and (b) all individuals for 1992. We used these files for summarizing the study data and for selected analyses for which the split-half sample approach was not required.

For the analysis of retrospective risk assessment (using the risk information for a year to predict expenditures for that same year), for each pool and year, we constructed two files, one to be used for estimating the models (the estimation half) and the second to be used for assessing predictive accuracy (the prediction half):
(c) 1991 Estimation sample, retrospective
(d) 1991 Prediction sample, retrospective
(e) 1992 Estimation sample, retrospective
(f) 1992 Prediction sample, retrospective.
These files supported the split-half design used for model estimation and prediction. Individuals were assigned randomly to the estimation and prediction halves for each year.

Finally, we constructed two pools to be used for the analysis of prospective risk assessment, one each for the estimation and prediction halves:
(g) 1991 and 1992 Estimation sample, prospective
(h) 1991 and 1992 Prediction sample, prospective.
The prospective data files included risk information for 1991 and expenditure data for 1992.

Not all carriers submitted data for both years; some submitted data for 1992 only. As a result, for the pools from these carriers, only retrospective 1992 applications of each model could be analyzed. Further, for those carriers with two years of data, not all enrollees are represented in a pool for both years. Some individuals were enrolled in a health plan in 1991 but not in 1992 (the leavers), while others were not enrolled in 1991 but were enrolled in 1992 (the joiners). As a result, only individuals in the same pool for both years were eligible for the prospective analyses.[15]

Finally, for those pools from carriers failing to submit data for nonclaimants, we were able to estimate, with some level of confidence, the mix of claimants and nonclaimants, by age and sex, in any one year. (All three carriers with this problem submitted data for both years.) This allowed us to perform the retrospective analyses for these pools for 1991 and 1992. However, although we could have estimated with some accuracy the number of individuals, by age and sex, who were in these pools for two years and were nonclaimants in both years, we did not have sufficient information to impute data for those who were in a pool for two years and were a nonclaimant in one year and a claimant for the second, or those who were both a nonclaimant and a leaver or joiner. As a result, we were unable to conduct prospective analyses for the pools from these three carriers.

Table 3 describes the final pools and years of data used for the study. As shown, the final study data included 19 pools from seven carriers.[16] All pools are for plans with deductibles less than $500. Of the 19 pools, eight, seven, and four pools represent indemnity, PPO, and HMO plans, respectively.

Table 3 also indicates the availability of data for the retrospective and prospective analyses. As shown, we were able to conduct the 1991 and 1992 retrospective analyses for 11 and 19 pools, respectively. Given the limitations of the data described above, we were able to conduct only prospective analyses for six pools.

Finally, Table 3 shows all of the pools had at least 9,500 observations for any one year, with the total number of observations available for analysis being approximately 1.1 million and 3.0 million for the 1991 and 1992 retrospective analyses, respectively, and 321,500 for the prospective analysis. Average pool sizes for the retrospective and prospective analyses were 136,600 and 53,600.

C. Study Methods

1. General Analytical Design

The general approach we used in testing predictive accuracy involved four major steps. First, using the methods described above and the information from health insurance claims, we assigned all individuals to risk groups for each of the eight models. In other words, individuals were assigned to an age-sex group, an ACG, an ADCG, and so on, depending on the demographic and clinical information recorded for them.

Second, using these risk group assignments and observed annual total health care costs (inpatient and outpatient) for the individuals in a risk group, the models' parameters were estimated. These estimates describe the expected costs, or risk, associated with a particular group or condition. The average costs by age-sex group or the incremental costs associated with a particular ADG are examples of such estimates.

Third, we used these estimates to predict the health care costs for each enrollee under a model. For example, for age and sex, if the expected annual cost for

TABLE 3
SUMMARY OF DATA USED IN STUDY

Pool ID	Number of Observations: Retrospective 1991	Number of Observations: Retrospective 1992	Number of Observations: Prospective	Notes
Indemnity				
IA	182,295	153,525	87,556	
IB	146,419	150,182	120,392	
IC	41,039	22,944	18,517	
ID	55,994	47,935		Imputed nonclaimants
IE	232,124	255,086		Imputed nonclaimants
IF		254,298		Imputed nonclaimants
IG	39,523	9,645		Imputed nonclaimants
IH	171,614	171,763		Imputed nonclaimants
PPO				
PA	120,141	169,512	66,015	
PB		271,690		
PC		204,350		
PD		153,101		
PE		460,392		
PF	116,929	194,888		Imputed nonclaimants
PG		194,879		Imputed nonclaimants
HMO				
HA	19,096	20,097	14,219	
HB		120,964		
HC		203,120		
HD	19,644	18,497	14,855	
Totals				
Indemnity Pools	869,008	1,065,378	226,465	
PPO Pools	237,070	1,648,812	66,015	
HMO Pools	38,740	362,678	29,074	
All Pools	1,144,818	3,076,868	321,554	

females, ages 15 to 19, is estimated to be $1,000, then that amount becomes the predicted value for all individuals in that age-sex group. Alternatively, if the expected cost for individuals in ACG 6 is determined to be $1,200, then that becomes the predicted cost for all enrollees assigned to that ACG, and so on.

Finally, model by model, we compared the predicted cost for each individual with the actual costs observed. The difference between predicted and actual costs, and how it varies across individuals, determines predictive accuracy.

As described previously, the data used in the study were obtained from a number of health insurance carriers and represented a range of health plan types—in terms of both health care management style and level of deductible. For the purpose of the analyses described in this chapter, the data were subdivided into separate *pools* based on carrier and plan type. We estimated model parameters and assessed predictive accuracy separately, pool by pool. We then summarized the findings across pools to derive general conclusions.

a. Assigning Individuals to Risk Groups

Assigning individuals to risk groups was straightforward. Based on an individual's age and sex, they were assigned to one of the age-sex groups described above. Based on their ambulatory diagnoses, as represented by ICD-9 codes, individuals were assigned to a particular ACG or zero, one or more ADG(s). Using inpatient only (for PIPDCGs) and both inpatient and ambulatory ICD-9 diagnostic codes, individuals were assigned to a DCG, depending on the DCG model employed. DCG high-cost coexisting conditions were also assigned.[17]

We employed a commercially available software product to assign patients to ACGs and ADGs (Johns Hopkins University, 1993). Using algorithms made available to us by DCG researchers, we constructed

computer programs to assign individuals to DCGs (Ellis, et al., 1995).

Table 4 provides examples of the assignment of risk groupings for five individuals from the study database. The upper portion of the table presents the information used in grouping each individual, while the lower portion describes the actual assignments made based on these data.

b. Estimating Model Parameters

We estimated the parameters separately for each risk assessment model using ordinary least squares regression. The dependent variable in a model was the total annual recorded health expenditures for an individual. The independent variables are a vector of dichotomous, or dummy variables, each describing the risk status of the individual under a particular model. For example, for age-sex, 28 dummy variables were employed in the regression (one for each age-sex group). Each dummy variable was given a value of 1 if the individual was in the age-sex group represented by the variable, and zero otherwise.[18]

For each risk assessment method, we estimated a model of the form:

$$Y_i = X_i * \beta + \epsilon_i \quad i = 1, \ldots, N \qquad (1)$$

where

i indexes the n individuals that belong to the sample of data used in estimating the model

β is a P by 1 matrix of the parameters (risk weights) to be estimated

X is a n by P matrix of the independent variables included in the model (the risk group dummy variables)

Y_i are the actual expenditures for an individual

ϵ is a 1 by n vector of disturbances/random error terms.

While Y_i is the same for a pool and year for each model tested, X_i differs across models. This is because the P independent variables in each model differ. For example, as shown in Table 1, for age and sex, 28 independent variables are used. For ADGs, 62 variables are employed (28 age and sex plus 34 ADGs). For ADCGDX, P=65 (28 age and sex, 12 ADCGs, and 25 CDXGs), and so on.

As described above, we conducted all analyses pool by pool. Specifically, we estimated a separate set of parameters for each risk assessment model for each pool of data.

c. Predicting Expenditures

Once each model was estimated using the "estimation" half of the data, its parameters could then be used to predict the annual health expenditures for individuals in the "prediction" half. The predicted expenditures for an individual are a function of the model parameter

TABLE 4
EXAMPLE OF ASSIGNMENT OF INDIVIDUALS TO RISK GROUPS

	Individual A	Individual B	Individual C	Individual D	Individual E
Age	24	39	30	32	56
Sex	F	M	F	M	F
ICD9 Codes for Ambulatory Diagnoses	V72.30 296.20 296.30 386.30 464.00 799.00	None	628.90 695.30	461.90 722.10 722.20	250.00 528.90 348.90 695.40 365.10 710.00 446.50 716.90 447.60 725.00 465.90 729.00 781.90
ICD9 Codes for Inpatient Diagnoses	None	None	656.31	722.10	307.81
Risk Groups					
Age-Sex	F2024	M3539	F3034	M3034	F5559
ACG	27	52	28	39	49
ADG(s)	1,2,23,24,31	None	20,28	8,16	2,10,11,18,20,26,27,28
PIPDCG	2	2	2	5	4
ADCG	5	1	1	1	5
EDCG	4	1	1	4	5
ADCGDX (CDXG)	None	None	None	None	6
EDCGDX (CDXG)	None	None	None	None	6

estimates and the individual's risk status and can be described generally as:

$$Y_i^{\wedge} \; X_i * b^{\wedge} \qquad (2)$$

where

$Y^{\wedge}$ is predicted health expenditures
$b^{\wedge}$ is a vector of the parameter estimates
X is a matrix of the dummy variables describing the risk status of individual i under an assessment model.

For each application (retrospective or prospective) and model, a separate prediction was made for each individual. Examples of such predictions for the same five individuals in Table 4 are shown in Table 5.[19]

d. Measuring Predictive Accuracy

There are at present no standard methods for comparing the predictive accuracy of different risk assessment methods. In many ways, the objectives of, and context for, risk adjustment dictate the comparisons to be made. For example, the ability to provide for equitable transfers between plans is related to how well a model predicts expenditures for a large group. The measure of a method's ability to minimize plans' incentives to selectively enroll or disenroll individuals or specific subgroups of individuals has to do more with how close predictions for individuals and these subgroups are to their actual amounts. In some applications, both individual and group predictive accuracy matter. The question is how to measure each of these constructs.

As discussed in Chapter II, there is some debate concerning the appropriate measure of the goodness of prediction for a model. Some argue that a risk adjustment formula only needs to predict well for large groups (Robinson, et al., 1993; Lubitz, 1987), while others maintain that this argument ignores the behavioral incentives of a plan to avoid high risks and attract low risks within a risk category (Newhouse, 1994). As a result, many studies explore the ability of risk assessment methods to predict costs at both the individual and group level. We followed this approach.

We compared actual expenditures with those predicted by a model using a number of measures and assessed predictive accuracy at three different levels, for (1) individuals, (2) random groups, and (3) nonrandom groups chosen for their expected low or high risk.

Individual Predictive Accuracy. We employed the following measures of individual predictive accuracy:

a. Individual adjusted R^2 (adjusted coefficient of multiple determination)
b. Mean absolute prediction error
c. Standard deviation of the absolute prediction error
d. Percentage of absolute prediction errors within $500
e. Percentage of absolute prediction errors within $1,000
f. Percentage of absolute prediction errors greater than $5,000
g. Percentage of absolute prediction errors greater than $10,000.

The adjusted R^2 can be described as the percentage of the variation in annual total health expenditures explained by a model.[20] The prediction error is the actual

TABLE 5
EXAMPLE OF PREDICTION OF EXPENDITURES FOR INDIVIDUALS UNDER DIFFERENT RISK ASSESSMENT METHODS—RETROSPECTIVE APPLICATION

	Individual A	Individual B	Individual C	Individual D	Individual E
Actual Expenditures	$743	$0	$5,509	$9,247	$9,046
Predicted Expenditures					
Age-Sex	1,217	734	1,673	683	1,812
ACG	2,628	0	1,541	1,695	6,412
ADG	3,063	27	2,723	1,465	8,605
PIPDCG	933	144	1,344	11,284	11,288
ADCG	5,570	167	994	174	5,721
EDCG	3,762	9	810	3,320	8,651
ADCGDX	5,177	204	1,025	203	5,906
EDCGDX	3,621	27	815	3,172	7,481

expenditures for an individual minus those predicted, while the absolute prediction error is the absolute value of that amount.

Each of these measures is designed to focus on a different aspect of individual predictive accuracy. The individual R^2 is a standard, summary measure of the ability of a model to predict individual values for the dependent variable. However, R^2 presents some limitations. First, 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. Second, although R^2 provides a meaningful measure in terms of a model's ability to predict the *variation* in expenditures across individuals, it does not provide any measure of the typical size of the errors or their distribution. The same R^2 can reflect either a large number of moderate sized errors, or many small and a few very large errors.[21]

We address the first limitation of the R^2 by using the mean absolute prediction error. The remaining measures address the second limitation. The standard deviation describes how dispersed the prediction errors are around the mean, while the percentage absolute errors by different ranges of dollar amounts provide a picture of the distribution of errors. In particular, measures (d) and (e), above, summarize the extent to which the errors are clustered around zero, while measures (f) and (g) summarize the distribution at its extremes, or "tails."

Group Level Predictive Accuracy—Random Groups. We assessed a model's ability to predict expenditures at the group level for both random and nonrandom groups. The purpose of random groups is to represent how a model would perform for large groups of enrollees based on the actual mix of enrollee risk represented by a pool in our study data. We chose to study nonrandom groups to simulate how well the models perform for those individuals of high or low risk who might choose health plans in a systematic way or might be targets for selective enrollment or disenrollment behavior by the plans themselves.

We assessed predictive accuracy for random groups by selecting a number of large groups of individuals from the sample of all individuals included in the prediction half of a pool of data. In particular, we randomly selected 100 groups of 2,500 individuals and, group by group, compared the actual expenditures for all individuals in the group with those predicted for these same individuals by a particular model.[22] We then summarized the results across the 100 groups to obtain measures of predictive accuracy.

In constructing random groups, we sampled individuals with replacement. By "with replacement," we refer to selecting an individual for a group from the larger sample and then returning that individual to the larger sample to be made available for selection into further groups. This approach can be compared to selection ' 'without replacement" where the selected groups are mutually exclusive—an individual can be in only one group.[23]

We employed six different measures of predictive accuracy for random groups:

a. Mean absolute prediction error
b. Standard deviation of the absolute prediction error
c. Mean absolute percentage prediction error
d. Predictive ratio
e. Percentage of absolute prediction errors within 5%
f. Percentage of absolute prediction errors within 10%.

The mean absolute prediction error is the average of the absolute values of the prediction errors across the 100 groups, while the standard deviation is a measure of the dispersion of those errors around the mean. The percentage error is the prediction error reported as a percentage of the predicted value. The predictive ratio is equal to total predicted expenses divided by total actual.

Group-Level Predictive Accuracy—Nonrandom Groups. Our objective in analyzing nonrandom groups was to identify those individuals who may present potential selection and equity problems for any risk adjustment process and to see how well the different models predict their expenditures. In doing this, we attempted to identify groups of individuals for whom a risk assessment method might consistently produce either over- or underpredictions.

We employed two criteria in selecting nonrandom groups for study: (1) those individuals with previously high or low expenditures relative to the mean and (2) those individuals with previous high-cost conditions that can be expected to persist into the future.[24] As described in our discussion of data below, we had two consecutive years of data for a large number of individuals in our sample. We used the expenditures and diagnostic information from the first year of data for these individuals to identify nonrandom groups. We then assessed each model's ability to predict their expenditures for the second year.[25]

We used the following criteria to assign individuals to three previous expenditure groups:

- *"Low" expenditure group.* Those individuals with first-year expenditures less than one-third of the mean expenditures for all individuals in a pool.
- *"HI1" expenditure group.* Those with first-year expenditures between three and six times the mean.
- *"HI2" expenditure group.* Those with first-year expenditures more than six times the mean.

We also identified three high-cost conditions for study[26]:

- Those patients hospitalized in the first year with higher cost cancers (leukemia, multiple melanoma, and cancer of the bone, breast, prostate, respiratory and digestive system, trachea, bronchus, and lungs)
- Those patients hospitalized in the first year with higher cost heart conditions (aortic valve disorders, mitral valve disorders, and acute ischemic heart disease)
- Those patients hospitalized in the first year with all conditions other than cancer and heart disease.

For each nonrandom group, we compared the mean predicted expenditures for the individuals in the group with those predicted by a model. We summarized this comparison using the predictive ratio, or the ratio of predicted to actual expenditures. In each case, the results for individuals not identified for the three expenditure groups (those with "medium" expenditures) and those without the specified clinical conditions were not summarized.

2. Retrospective versus Prospective Application of Risk Assessment Models

As described above, risk assessment can be applied in a prospective or retrospective fashion. Prospectively, the risk experience of one year would be used to compute expected expenditures for a future year. Retrospectively, the relationship between risk experience and expected expenditures for the same year would be assessed.

The two years of data available for the study allowed us to use both a prospective and retrospective approach when applying each risk assessment model. For a prospective application, we used the risk assessment information (age-sex, ACGs, ADCGs, and so on) for an individual from the first year of data to predict his or her second-year expenditures. This involved first estimating the model parameters by using second-year expenditures as the dependent variable and variables describing the first-year risk status as independent variables. This model was then applied to assess predictive accuracy under a prospective design.

For a retrospective application, we used the risk assessment information for an individual in a year to predict expenditures for that same year. This involved first estimating the model parameters using expenditures for a year as the dependent variable and variables describing risk status for that year as the independent variables. This model was then applied to assess predictive accuracy. Since two years of data were available for most pools, where possible, we applied the retrospective model twice, once for each year.

In addition to predictive accuracy, there are a number of important issues regarding the choice of retrospective versus prospective risk assessment, including the practical issues in administering such systems and the incentives for gaming and the provision of efficient medical care. We discuss these issues in Chapter VI.

3. Truncation of Expenses for High-Cost Individuals

As expected, we observed the distribution of health expenditures across individuals in our study data to be highly skewed. In particular, a significant portion of all enrollees have no reported expenses for a particular year; many others have relatively low expenditures; and a small number have extremely high expenditures.

This skewness can produce biased and imprecise parameter estimates, thus compromising a model's ability to provide accurate predictions. After discussion with the advisors to the project, we chose to truncate total expenditures at $25,000 for the purposes of the analyses described in this chapter. If an enrollee had a total expenditure amount exceeding $25,000 for a year, this amount was set to $25,000. In other words, for all individuals, we only included the first $25,000 of expenses for these analyses.[27]

Any risk adjustment approach will likely require some method to deal with higher cost cases. One such mechanism could involve a dollar limit for an individual, after which all expenses will be reinsured in some way. Truncating at $25,000 simulates such a reinsurance approach.

We discuss alternative methods for the risk assessment of high-cost individuals, including different forms of reinsurance, in Chapter IV. In addition, to assess the sensitivity of our findings to the decision to truncate expenditures at $25,000, we also repeated selected

analyses using $50,000 as the truncation amount and with no truncation. The results of these analyses are presented later in this chapter.

D. Results

1. A Statistical Description of the Study Data

Tables 6 through 10 summarize the characteristics of the data used in the study. These tables provide an overall summary, by plan type, for the year 1992. Detailed summaries, pool-by-pool, for both 1991 and 1992 are included in Appendix C.

Table 6 shows the study data included enrollees from the entire range of age groups and were evenly distributed by gender. Table 7 shows the data to be national in scope, covering a wide range of geographic areas, and to be, in general, geographically representative of the nonelderly U.S. population. Table 7 also shows that approximately 95% of all enrollees were without a hospital admission during the year. This percentage has important implications for the risk assessment methods (DCGs) that employ inpatient diagnoses in grouping individuals.

As expected, a significant portion of enrollees in any pool had either no, or relatively modest expenditures, while only a small percentage had expenditures greater than $25,000 (Tables 8 and 9). (Note that since we truncated the data at $25,000 for most analyses, most statistics on expenditures are reported using the truncated data.) Figure 2 describes graphically the skewness of the distribution of enrollees in the study, by level of expenditures.[28]

Some results in Tables 8 and 9 are worth noting. First, approximately one-third of all enrollees had no expenditures in any year, a significant proportion of the sample. Further, more than three-fourths of all enrollees have expenses less than $1,000. Third, for most pools, less than 1% of the enrollees had expenditures greater than $25,000. Finally, although these higher cost enrollees are small in number, they represent a significant percentage of total expenditures—on average, more than 25% across all pools. This result underlines the importance of higher cost individuals for any risk adjustment process.

Finally, Table 10 shows that inpatient expenditures comprise the largest portion of total reported expenditures, closely followed by professional (mostly physician) medical and surgical expenses, outpatient facility

TABLE 6
AGE AND GENDER DISTRIBUTION OF ENROLLEES BY PLAN TYPE, 1992

			Age Groups							
Plan Type	N	% Female	0	1–4	5–14	15–24	25–34	35–44	45–54	55–64
Indemnity	1,065,378	52.3	2.7	6.4	14.4	13.0	19.5	21.0	13.6	9.4
PPO	1,648,812	49.2	1.9	6.5	15.5	12.7	21.4	22.0	13.2	7.1
HMO	362,678	51.3	1.4	6.9	17.3	12.0	20.3	21.3	13.1	7.4
All Pools	3,076,868	50.5	2.1	6.5	15.3	12.7	20.6	21.5	13.4	8.0

TABLE 7
REGIONAL DISTRIBUTION OF ENROLLEES BY PLAN TYPE, 1992

	Percentage Enrollees by Region*										% w/no
Plan Type	ENC	ESC	MdAt	Mtn	NE	Pac	SAtl	WNC	WSC	Other	Admissions
Indemnity	42.1	5.0	3.2	4.8	.4	5.6	13.4	13.8	11.1	.4	94.4
PPO	16.3	2.6	13.8	5.0	2.2	12.8	16.1	8.7	22.6	.04	95.0
HMO	5.7	.1	13.5	.2	.1	30.1	2.5	.2	47.7	.06	95.4
All Pools	20.4	2.8	11.5	4.3	1.5	13.7	13.6	8.7	12.7	.1	95.1
US Pop <65 1990	17.2	6.2	15.2	5.5	5.2	15.7	16.9	7.2	11.0	—	—

*Key: ENC = East North Central
ESC = East South Central
MdAt = Mid-Atlantic
Mtn = Mountain
NE = New England
Pac = Pacific
SAtl = South Atlantic
WNC = West North Central
WSC = West South Central

TABLE 8
TOTAL EXPENDITURE DISTRIBUTION BY PLAN TYPE, 1992

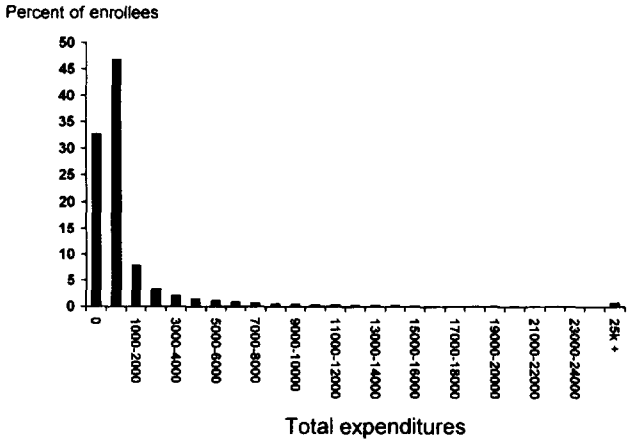
Plan Type	% 0	Q1	Median	Q3	95%	99%	Max	Mean	STD	%>25K	Mean w/o Truncation
Indemnity	34	56	279	1,132	8,478	22,861	25,000	1,617	3,869	1.46	2,328
PPO	33	0	148	665	5,385	17,075	25,000	1,063	2,975	0.73	1,246
HMO	37	10	97	492	4,926	16,770	25,000	947	2,913	0.53	1,127
HMO w/o HC*	24	22	202	779	5,774	18,970	25,000	1,179	3,147	0.68	1,430
All Pools	34	13	170	743	6,000	18,313	25,000	1,168	3,165	0.9	1,466

*Measures for this table were computed separately with and without HC pool due to significant differences between that pool and other HMO pools.

TABLE 9
FREQUENCY DISTRIBUTION OF EXPENDITURES BY PLAN TYPE, 1992

	Percentage Distribution of Expenditures Truncated at $25,000								Percentage Distribution of Expenditures >$25K				
Plan Type	$0	$1–100	$100–500	$500–2000	$2K–5K	$5K–10K	$10K–15K	$15K–25K	$25K–50K	$50K–100K	$100K–250K	$250K–500K	$500K+
Indemnity	33.6	10.1	22.5	19.0	7.3	3.8	1.5	2.0	67.0	24.5	7.4	0.7	0.14
PPO	32.8	12.0	25.5	18.3	6.2	3.1	1.1	1.2	68.3	23.2	7.0	1.2	0.17
HMO	36.9	15.6	22.9	14.4	5.18	2.8	1.1	1.2	62.0	25.6	10.3	1.8	0.43
All Pools	35.6	12.1	24.5	17.9	6.3	3.2	1.2	1.4	67.5	23.6	7.4	1.2	0.19

FIGURE 2
PERCENTAGE OF ENROLLEES BY $1000 EXPENDITURE GROUPS



costs, and expenses in the "other" category. Reported drug expenditures are relatively smaller for all pools.

The actuarial consultants to the study who reviewed these data considered the distributions to be consistent, in general, with those they had observed in practice.

2. *Predictive Accuracy for Individuals*

Table 11 presents a summary of the predictive accuracy of each model at the individual level. All summary tables presented in this section represent the weighted average results across all relevant pools, with the weights being the number of observations included in the prediction sample for each pool. Detailed versions of these tables for each pool and year, are included in Appendix C.

In Table 11, each row summarizes the findings for a particular model across all pools, while each column represents a measure of accuracy. Both retrospective and prospective results are reported. Since retrospective findings for 1991 and 1992 were very similar, only those for 1992 are reported here.

Retrospective. For the retrospective approach, the diagnosis-based models clearly outperform age and sex at the individual level—predictive accuracy improves with the addition of health status information. Among the diagnosis-based models, the differences are less marked. In terms of mean absolute error, the ACG and ADG models show slightly better results, while the PIPDCG model performs best in terms of the standard deviation of the absolute errors. The ACG and ADG models consistently show a larger percentage of absolute errors less than $500—a result partly due to the fact that the ACG model uses a separate ACG for those without claims, and the ADG model specifies no ADGs for these individuals. By definition, ACGs and ADGs

TABLE 10
AVERAGE EXPENDITURE BY SERVICE GROUP, BY PLAN TYPE (TOTAL EXPENDITURES TRUNCATED AT $25,000), 1992

	Medical		Surgical		Inpatient		Outpatient		Drug		Other	
Plan Type	Avg.	% of Total	Avg.	% of Total	Avg.	% of Total	Avg.	% of Total	Avg.	% of Total	Avg.	% of Total
Indemnity	$191	11.8%	$261	16.1%	$431	26.7%	$298	18.4%	$111	6.8%	$325	20.1%
PPO	221	20.8	158	14.9	268	25.2	180	16.9	57	5.4	179	16.8
HMO	127	13.4	254	26.8	292	30.8	136	14.3	106	11.2	32	3.4
All Pools	206	17.6	194	16.6	306	26.2	198	17.0	74	6.3	190	16.2

TABLE 11
SUMMARY OF PREDICTIVE ACCURACY FOR INDIVIDUAL RESULTS RETROSPECTIVE AND PROSPECTIVE ANALYSIS OF ALL POOLS

Risk Assessment Method	Mean		Standard Deviation Absolute Error	% Absolute Error				Adjusted R^2
	Predicted	Absolute Error		Within $500	Within $1,000	Over $5,000	Over $10,000	
Retrospective Analysis								
Age-sex	1,133	1,465	2,713	29	58	4.5	2.1	0.032
ACGs	1,134	1,048	3,434	65	75	4.1	1.6	0.286
ADGs	1,133	1,036	2,291	64	75	4.0	1.5	0.357
PIPDCGs	1,133	1,114	2,079	46	73	4.1	1.4	0.428
EDCG	1,140	1,140	2,280	53	74	4.3	1.6	0.343
ADCG	1,140	1,213	2,432	50	72	4.7	1.8	0.252
EDCGDX	1,131	1,100	2,223	53	74	4.2	1.5	0.372
ADCGDX	1,132	1,175	2,356	52	72	4.5	1.7	0.293
Prospective Analysis								
Age-sex	1,351	1,673	2,907	22	54	5.1	2.5	0.039
ACGs	1,343	1,570	2,859	19	67	5.0	2.3	0.091
ADGs	1,343	1,536	2,832	35	61	5.0	2.3	0.112
PIPDCGs	1,349	1,658	2,873	23	49	5.0	2.3	0.061
EDCG	1,346	1,603	2,844	30	53	5.0	2.4	0.091
ADCG	1,345	1,613	2,851	29	54	5.0	2.3	0.084
EDCGDX	1,346	1,599	2,867	33	53	5.3	2.4	0.096
ADCGDX	1,345	1,606	2,869	32	55	5.2	2.4	0.093

thus predict expenditures with perfect or near-perfect accuracy for nonclaimants in a retrospective model.[29]

The differences between percentage prediction errors over $5,000 and $10,000 in the models are small with the ADG and PIPDCG models having a slight advantage. This result for PIPDCGs contrasts with its finding for errors less than $500, where it falls short of the other models.

The adjusted R^2 values are consistently higher for PIPDCGs (0.428), followed by the EDCGDX (0.372), ADG (0.357), and EDCG (0.343) models, which show similar results. The high relative performance on this and other measures for PIPDCGs is likely a result of its use of only inpatient diagnoses in grouping patients, clearly distinguishing these retrospectively higher cost individuals from other enrollees.[30]

Prospective. The prospective results in Table 11 again show that the diagnosis-based models outperform age and sex at the individual level. However, the difference between these two types of models is not as large as observed for the retrospective design. In fact, while the retrospective and prospective results for age and sex are quite similar, those for the diagnosis-based models show a marked decrease in predictive ability when moving from a retrospective to a prospective application. This is not surprising given that age and sex for an individual

is almost identical in the prospective and retrospective models (same sex and a year older). For the diagnosis-based models, the risk data for an individual for a year (diagnoses and so on) would be expected to be more closely correlated with their expenditures in that year (retrospective) than in the following year (prospective). The decrease in predictive accuracy for these models reflects this relationship.

There are some differences in predictive accuracy among the prospective diagnosis-based models, but they are not large. On most measures, including the mean and standard deviation of the absolute prediction error and the percentage of errors greater than $5,000 and $10,000, most of the diagnosis-based models perform equally well. Some differences are again found for the absolute errors within $500 where ADGs, EDCGDXs, and ADCGDXs show somewhat better results.

For the adjusted R^2, the ADG model shows consistently somewhat higher values across all pools studied (0.112). With the exception of PIPDCGs, the other diagnosis-based models have comparable values, ranging from 0.084 for ADCGs to 0.096 for the EDCGDX model. The R^2 values for PIPDCGs are somewhat lower (0.061) and can be contrasted with those for this model under the retrospective design. While the prospective PIPDCG model continues to show a link between hospitalizations of different types and expected costs, this link is considerably weaker than under the retrospective application. The prospective performance of PIPDCGs versus the other diagnosis-based models also demonstrates the value of including ambulatory diagnoses when predicting future year's expenditures, perhaps due to their ability to capture other chronic conditions that may persist over time.

Figures 3, 4, and 5 graphically summarize selected individual retrospective and prospective results.

Comparisons Across Pools and Plan Types. Table 12 summarizes the adjusted R^2 values across pools and plan types. As shown, the relative performance of the models is quite consistent, suggesting the findings are robust to both health care management type and the population of enrollees, as measured by carrier. Given the potential for different levels of coverage/benefits across pools, the findings suggest the relative performance of the models may also be robust along this dimension.

Predicted versus Actual Expenditures. Finally, Figures 6 and 7 present the relationship between predicted and actual expenditures for selected models. Each plot point in these figures describes the mean predicted and mean actual expenditures for individuals grouped by increments of $1,000 of actual expenditures (25 groups, $0 through $25,000). A 45 degree line is added to represent the equivalence of predicted and actual amounts.

FIGURE 3
SUMMARY OF PREDICTIVE ACCURACY, INDIVIDUAL RESULTS, ADJUSTED R^2, ALL POOLS, 1992

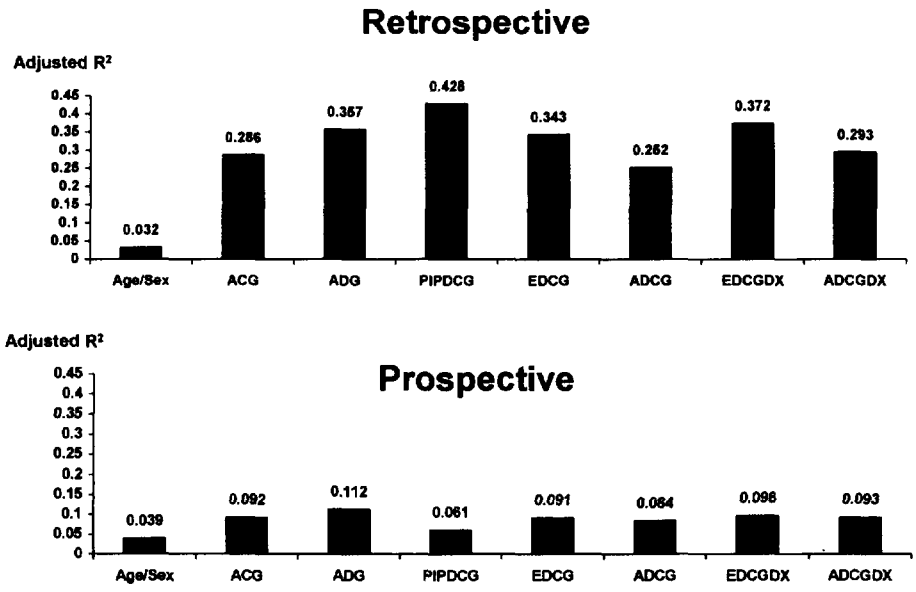


FIGURE 4
SUMMARY OF PREDICTIVE ACCURACY, INDIVIDUAL RESULTS,
PERCENTAGE OF ABSOLUTE ERRORS WITHIN $500, ALL POOLS 1992

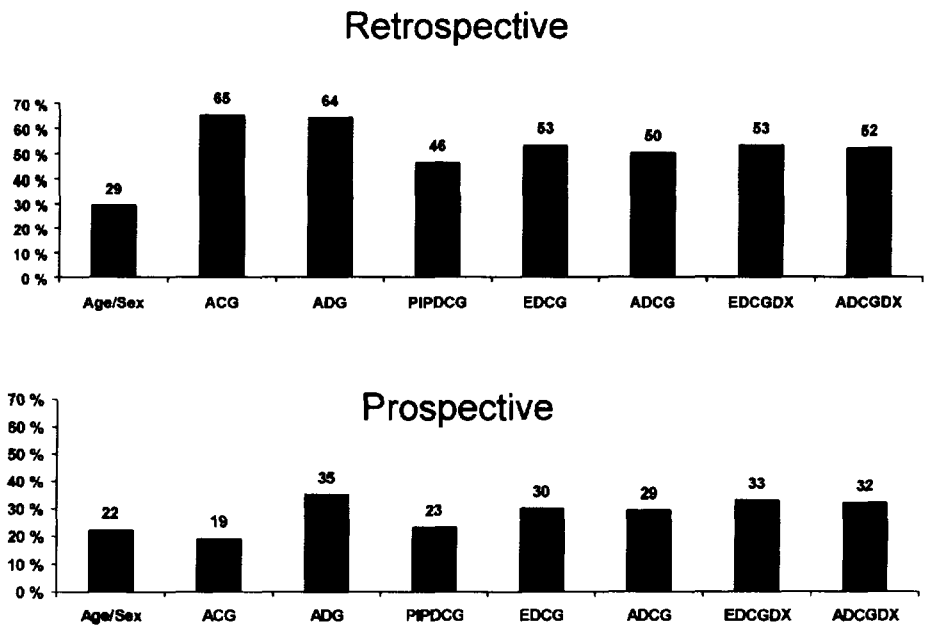


FIGURE 5
SUMMARY OF PREDICTIVE ACCURACY, INDIVIDUAL RESULTS,
PERCENTAGE OF ABSOLUTE ERRORS GREATER THAN $5,000, ALL POOLS, 1992

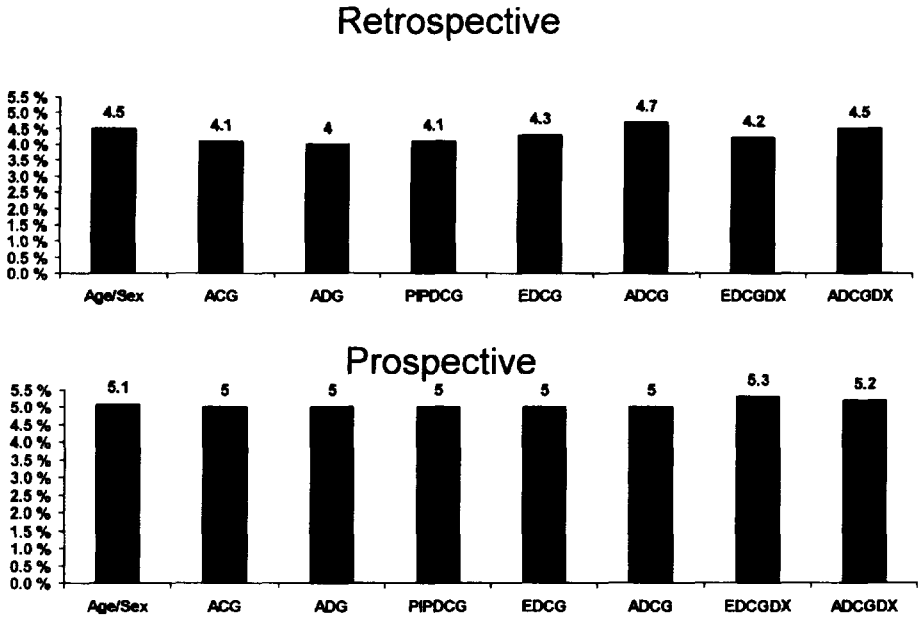


TABLE 12
ADJUSTED R^2 FOR SELECTED POOLS, BY METHOD, INDIVIDUAL LEVEL ANALYSES

Pool	Age-Sex	ACGs	ADGs	PIPDCGs	EDCG	ADCG	EDCGDX	ADCGDX
				Retrospective 1992				
IA–	.037	.274	.357	.492	.369	.265	.393	.303
IB–	.038	.299	.356	.568	.377	.271	.403	.306
IC–	.036	.314	.352	.500	.371	.279	*	*
IE–	.031	.327	.394	.292	.313	.259	.345	.296
IH–	.037	.187	.257	.562	.404	.246	.413	.256
PA–	.039	.266	.331	.459	.327	.230	.351	.265
PC–	.029	.260	.329	.496	.364	.251	.392	.297
PE–	.028	.268	.338	.486	.357	.254	.389	.300
PF–	.036	.301	.373	.311	.291	.227	.325	.271
HA–	.035	.253	.338	.485	.336	.221	*	*
HB–	.029	.284	.351	.452	.356	.260	.371	.308
HC–	.029	.381	.473	.377	.337	.276	.362	.306
HD–	.046	.281	.390	.527	.322	.205	*	*
Indemnity	.034	.291	.360	.408	.346	.260	.372	.293
PPO	.031	.271	.341	.444	.341	.244	.372	.290
HMO	.030	.337	.421	.416	.343	.264	.365	.307
All Pools	.032	.286	.357	.428	.343	.252	.372	.293
				Prospective				
IA–	.034	.090	.107	.060	.093	.084	.096	.090
IB–	.043	.095	.120	.069	.102	.095	.105	.102
IC–	.035	.090	.118	.059	.097	.095	*	*
PA–	.039	.088	.106	.051	.074	.069	.078	.081
HA–	.039	.078	.103	.059	.064	.077	*	*
HD–	.034	.086	.103	.054	.088	.061	*	*
All Pools	.039	.091	.112	.061	.091	.084	.096	.093

* Insufficient number of enrollees to test the EDCGDX and ADCGDX models.

Points below the line are underpredictions, while points above the line represent overpredictions. Figure 6 presents the retrospective results, while Figure 7 includes the prospective findings.

For all models shown, the predicted values exceed the actual amounts for lower levels of actual expenditures, while the reverse is true for higher levels. Those with relatively low or high actual expenditures also tend to be low or high relative to those values predicted by a model.

A second observation from Figures 6 and 7 concerns the relative range of predicted versus actual expenditures. In general, the range of predicted values is much smaller than that of the actual amounts. This is particularly true for prospective models where the maximum predicted value for these groups is $3,199 (EDCGDX). For retrospective models, the maximum predicted values for these groups is $13,250 (PIPDCG). The actual values range up to $25,000.

The results summarized in Figures 6 and 7 indicate that the models do not predict well for the lower and higher cost cases. Further, the range of predicted values is only a fraction of that for actual amounts. In some ways, this finding is to be expected, given both the skewness in the distribution of expenditures and that a model is predicting average expenditures for each risk group. Actual expenditures will distribute themselves around this mean. If we're looking at individuals at the low end of the distribution of actual expenditures, these persons are also likely to be lower than average in their risk group. The reverse is true for higher cost individuals. Figures 6 and 7 show a strong relationship between lower/higher cost individuals in a risk group and lower/higher cost individuals overall. The figures also show that the variation in individual expenditures left unexplained by these risk assessment models may present a greater problem for higher cost cases. If this unexplained variation is unpredictable and these high-cost individuals distribute themselves randomly across plans, then this result is not a problem for risk adjustment. However, if it can be predicted by plans or individuals,

FIGURE 6
MEAN PREDICTED VERSUS ACTUAL 1992 EXPENDITURES, GROUPED BY 1992 ACTUAL EXPENDITURES, RETROSPECTIVE MODELS

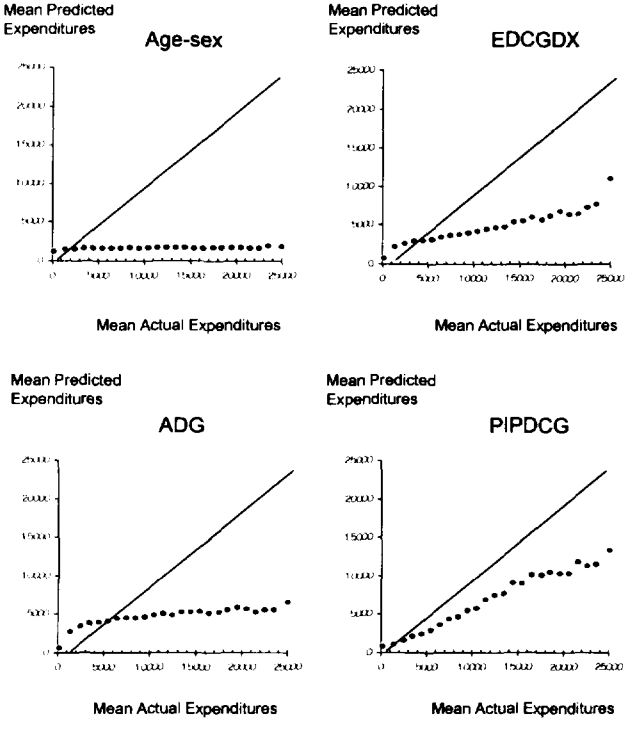
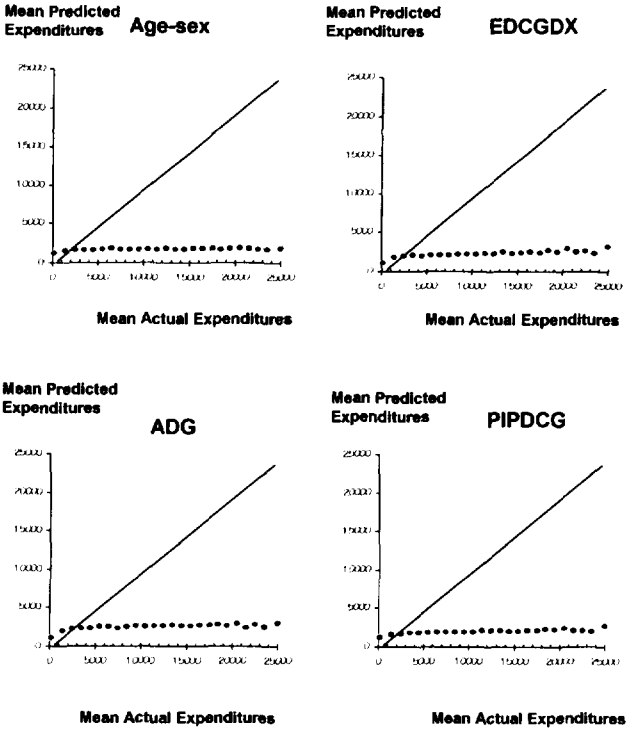


FIGURE 7
MEAN PREDICTED VERSUS ACTUAL 1992 EXPENDITURES, GROUPED BY 1992 ACTUAL EXPENDITURES, PROSPECTIVE MODELS



selection is possible. We explore this issue further below for nonrandom groups and later in this report.

3. *Predictive Accuracy for Random Groups*

Table 13 summarizes the findings on predictive accuracy for large random groups. Since we selected for each pool 100 groups of 2500 enrollees for this analysis, each of the measures of accuracy for a pool describes the results across the 100 groups. For example, for each of the 100 groups chosen, mean actual and predicted expenditures per individual were computed. The difference between these two amounts is the prediction error for the group. Table 13 summarizes these prediction errors using different measures across all pools.

All the models, including age and sex, perform well for large random groups. There is some decrease in predictive accuracy when moving from a retrospective to prospective design, but the decrease is not large. For the retrospective applications, the mean absolute percentage prediction errors range from 3.5% for PIPDCG to 4.6% for age-sex. For the prospective applications, these values range from 4.2% for both ACGs and ADGs to 4.6% for PIPDCGs.

The findings for random groups illustrate the effects of pooling the risk of a large number of individuals when performing risk assessment. By the law of large numbers, individual prediction errors tend to cancel each other out when combined into large groups. Even simple models, such as age and sex, can predict expenditures for the group which are, on average, quite close to the actual amounts.

Table 14 summarizes the mean absolute percentage prediction errors for random groups across pools and plan types. As with the individual results, the relative performance of the models is quite consistent, again suggesting the findings are robust with respect to health care management type, the population of enrollees, and

TABLE 13
SUMMARY OF PREDICTIVE ACCURACY—GROUP RESULTS RETROSPECTIVE AND PROSPECTIVE ANALYSIS OF ALL POOLS

Risk Assessment Method	Mean		Standard Deviation	Predictive	% Absolute Error	
	Absolute Error	Absolute % Error	Absolute Error	Ratio	Within 5%	Within 10%
Retrospective Analysis						
Age-sex	51	4.6	38	1.00	60	92
ACGs	42	3.8	34	1.01	70	95
ADGs	41	3.6	32	1.00	73	97
PIPDCG	39	3.5	29	1.00	75	98
EDCG	41	3.6	30	1.00	71	98
ADCG	43	3.8	32	1.01	70	97
EDCGDX	41	3.7	30	1.00	72	97
ADCGDX	43	3.8	32	1.00	66	97
Prospective Analysis						
Age-sex	58	4.3	40	1.02	63	94
ACGs	58	4.2	39	1.02	64	95
ADGs	56	4.2	39	1.02	63	95
PIPDCG	61	4.6	41	1.02	60	93
EDCG	59	4.4	39	1.02	63	94
ADCG	58	4.3	39	1.02	63	95
EDCGDX	58	4.3	39	1.02	63	95
ADCGDX	57	4.3	39	1.02	64	94

TABLE 14
MEAN ABSOLUTE PERCENTAGE ERROR FOR SELECTED POOLS, BY METHOD, GROUP LEVEL ANALYSES

Pool	Age-Sex	ACGs	ADGs	PIPDCGs	EDCG	ADCG	EDCGDX	ADCGDX
Retrospective Analysis 1992								
IA–	3.7	3.6	3.5	2.8	3.4	3.8	3.5	3.8
IB–	4.4	3.3	3.4	2.8	3.4	3.5	3.3	3.5
IC–	3.9	3.3	3.0	3.6	3.3	3.5	*	*
IE–	4.5	3.7	3.6	3.6	3.2	3.5	3.6	3.8
IH–	4.2	4.0	3.9	2.7	3.4	3.9	3.5	4.0
PA–	4.4	3.5	3.2	3.4	3.4	3.4	3.4	3.3
PC–	4.3	3.4	3.2	3.2	3.6	3.5	3.4	3.4
PE–	4.8	3.7	3.4	3.4	3.7	3.8	3.7	3.8
PF–	4.7	3.9	3.9	4.0	3.8	3.9	4.1	4.0
PG–	5.0	3.7	3.7	3.8	3.5	3.8	3.7	4.2
HA–	3.9	4.5	3.1	2.7	3.7	4.3	*	*
HB–	4.8	4.2	4.1	3.1	3.7	4.1	3.7	4.0
HC–	4.7	3.7	3.7	4.4	4.2	4.6	4.2	4.5
HD–	3.5	3.8	2.8	2.2	2.8	2.9	*	*
Indemnity	4.3	3.7	3.6	3.3	3.4	3.6	3.5	3.7
PPO	4.7	3.8	3.6	3.5	3.7	3.8	3.7	3.8
HMO	4.6	3.9	3.8	3.8	3.9	4.3	4.0	4.3
All Pools	4.6	3.8	3.6	3.5	3.6	3.8	3.7	3.8
Prospective Analysis								
IA–	3.9	4.0	4.2	4.2	4.1	4.1	4.0	4.1
IB–	4.3	4.0	4.2	4.5	4.2	4.2	4.2	4.2
IC–	3.8	3.2	3.1	3.7	3.4	3.4	*	*
PA–	4.4	5.0	4.4	5.2	4.9	4.7	4.9	4.7
HA–	7.5	7.0	6.9	7.0	7.6	7.4	*	*
HD–	3.2	3.0	2.9	3.2	3.1	3.0	*	*
All Pools	4.3	4.2	4.2	4.6	4.4	4.3	4.3	4.3

* Insufficient number of enrollees to test the EDCGDX and ADCGDX models.

plan benefits. Figure 8 summarizes graphically the random group findings across all pools.

4. *Predictive Accuracy for Nonrandom Groups*

If individuals of different risk were in fact distributed randomly across different health plans, as simulated by the random group analysis described above, risk adjustment across plans takes on lesser importance. To fully compare the different models also requires testing them for nonrandom groups.

As described above, we chose two types of nonrandom groups: (1) those with relatively high or low expenditures in 1991 and (2) those with selected high-cost conditions in 1991. We then compared the ability of the different models to predict the expenditures for these groups for 1992. In doing this, we used both the retrospective and prospective applications of each model. Since the methods we employed for nonrandom groups required two years of complete data, data from only six pools were used for these analyses (pools IA, IB, IC, PA, HA, and HD).

Figures 9 through 11 summarize the findings for nonrandom groups. Figure 9 presents the results for the nonrandom groups selected based on previous expenditures. The horizontal axis of each plot notes the eight models tested, while the vertical axis measures the predictive ratio, or the ratio of predicted to actual expenditures. For each model, a set of three points is included, one for each of the three expenditure groups (LOW, HI1, and HI2). (For example, for the age-sex model, retrospective application, the predictive ratios were 1.70, 0.58, and 0.37 for the LOW, HI1 and HI2 groups.) A horizontal reference line is placed at a predictive ratio of 1.00, indicating the equivalence of predicted and actual amounts. (As a further reference, these predictive ratios for random groups were, on average, close to 1.00.) Points above the reference line indicate, that, on average, overpayments for a group would result, while those below the line indicate underpayments.

In general, the models would overpay for those groups of individuals with previously lower expenditures and underpay for those with previously higher amounts. The figure also shows systematic differences across the models. The greatest inequities occur for age and sex, where expenditures are overpredicted for the low group by 70% and underpredicted for the HI1 and HI2 groups by 40 and 60%, respectively. Overall, the ADG model performs best, overpredicting expenditures

FIGURE 8
SUMMARY OF PREDICTIVE ACCURACY, RANDOM GROUP RESULTS

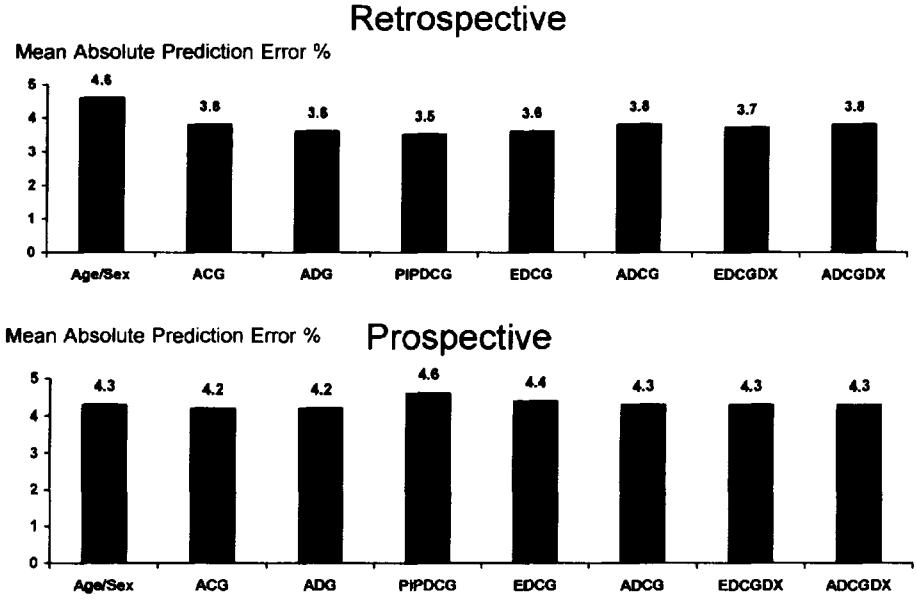
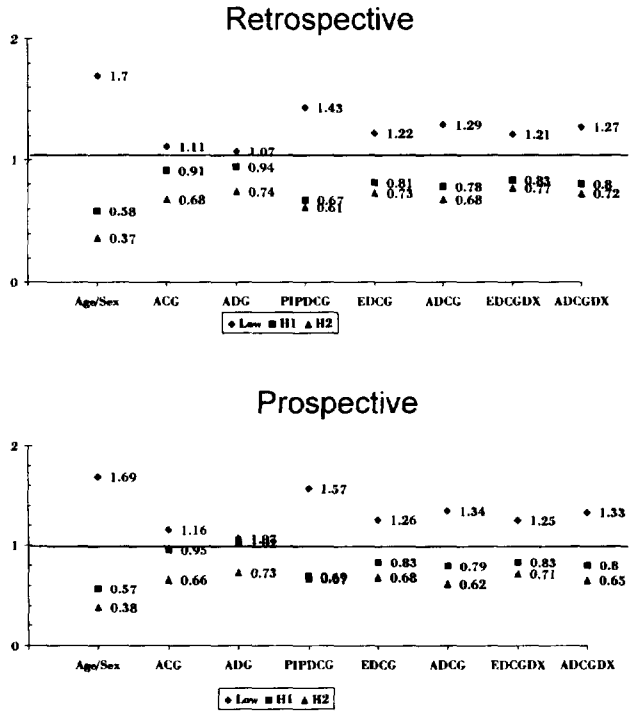


FIGURE 9
RATIO OF PREDICTED TO ACTUAL 1992 EXPENDITURES FOR NONRANDOM GROUPS OF ENROLLEES, GROUPINGS BASED ON RELATIVELY LOW OR HIGH EXPENDITURES IN 1991



for the low group by only 7%, coming close, on average, to predicting actual expenditures for the HI1 group, and underpredicting for the HI2 group by 26%. The EDCG and EDCGDX models performed comparably to ADGs for the highest expenditure group.

Another interesting finding from Figure 9 is the comparable performance of the retrospective and prospective models. For most models, the two applications produce almost identical results. Where differences exist, the retrospective model typically performs better, but the differences are small. This finding is somewhat surprising given the differences between the predictive ability of the retrospective and prospective applications at the individual level.[31]

Figure 10 presents a more refined picture of the relationship between previous year's expenditures and the relationship between predicted and actual expenditures for 1992. For illustrative purposes, only the prospective results for the age-sex, ADG, PIPDCG, and EDCGDX models are presented. The results are similar for retrospective models. As shown, all models overpredict 1992 expenditures for individuals with less than $1,000 in 1991 expenditures. The age-sex model underpredicts all other expenditures. The other models also tend to underpredict expenditures for 1991 expenditure groups greater than $1,000. Of the four models, the ADG model performs the best, with predictive ratios close to one for all groups with 1991 expenditures less than $10,000.

Figures 9 and 10 show there is a clear relationship between the prior and current years' expenditures, even after the application of the risk assessment models tested. This finding suggests a health plan could use prior expenditures to predict likely "winners" and "losers" under these methods, particularly for individuals with extremely high or low prior claims.

Figure 11 presents the results for the nonrandom groups based on a 1991 inpatient diagnosis of heart disease, cancer, or all other conditions. As shown, all the models underpredict expenses for the groups of individuals with a prior admission for heart disease or cancer. The models come closer to predicting expenditures for other inpatient diagnoses.

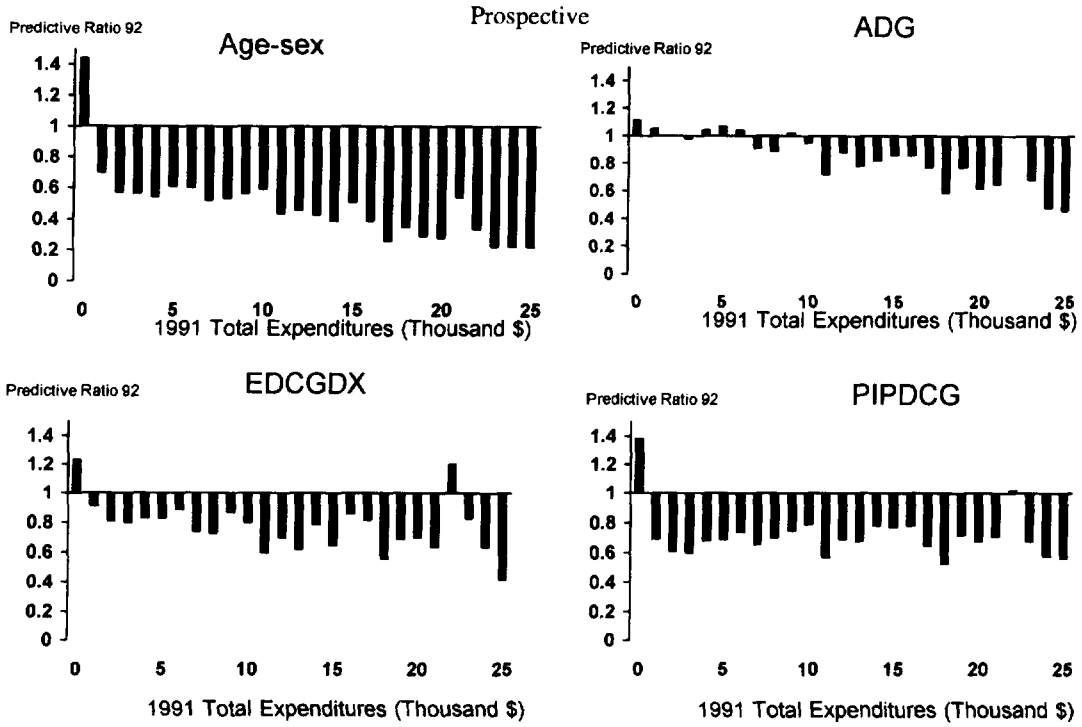
Figure 11 also shows some systematic differences between models. In general, the EDCGDX model performs best with the predictive ratios for the three groups closest to 1.0 under both the retrospective and prospective designs. Age-sex consistently underpredicts expenses for all three groups by significant amounts. As before, these results suggest plans can use prior information about individuals to predict likely "losers" even after risk adjustment using these methods.

Finally, as found for the nonrandom groups based on extreme expenditures, the retrospective and prospective models produce similar results, with the exception of the "other" group where the predictive ratios are closest to 1.0 for prospective models. While the retrospective models clearly produce superior predictions for individuals, they produce a bias similar to that found for prospective models when predicting expenditures for nonrandom groups.

5. *Sensitivity Analyses*

We tested the sensitivity of our findings to the data and assumptions we employed. In particular, we conducted four separate investigations:

FIGURE 10
RATIO OF PREDICTED TO ACTUAL 1992 EXPENDITURES FOR NONRANDOM GROUPS OF ENROLLEES, BASED ON $1,000 GROUPINGS OF 1991 EXPENDITURES, SELECTED METHODS



- The sensitivity of predictive accuracy for individuals and random groups to the decision to truncate expenditure data at $25,000
- The sensitivity of predictive accuracy for individuals and random groups to the use of inpatient diagnoses in the ACG and ADG models
- The differences in predictive accuracy for leavers versus joiners versus stayers
- The impact of group size on the predictive accuracy results for random groups.

We describe briefly below the methods used in these analyses. To perform all of these sensitivity analyses, we used three pools of data (pools IA, PA, and IB). We summarize here only the key findings across these three pools. More detailed results are presented in Appendix C.

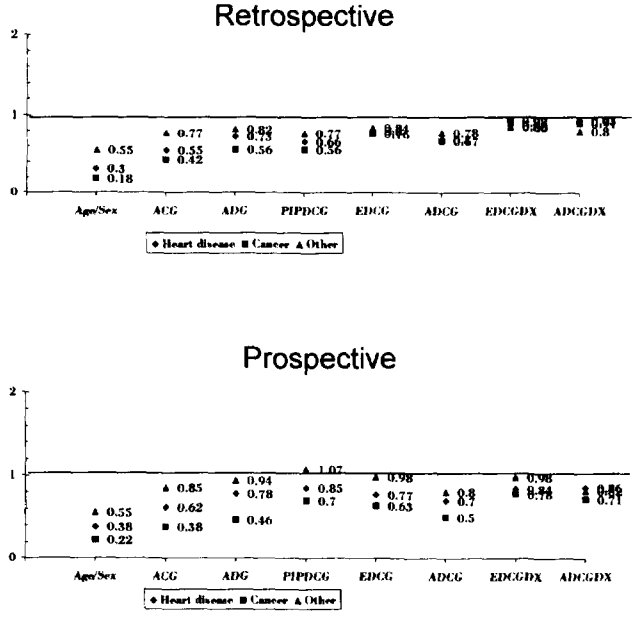
a. Truncation of Expenditures

As described previously, for the analyses presented in this chapter, we truncated expenditures at $25,000. We explored the sensitivity of our findings to this decision by repeating analyses for individuals and random groups using higher thresholds: a truncation amount of $50,000 and no truncation.

Table 15 summarizes these results. In general, the R^2 values decrease with higher thresholds for truncation. The exception is the retrospective EDCGDX model, which actually shows an increase in R^2 when moving from a $25,000 to $50,000 threshold.

Given that R^2 values can be sensitive to the variance in the expenditures across individuals, and that variance will increase with higher thresholds for truncation, a more valid test is the relative performance of a model within a threshold. This comparison shows that for both the retrospective and prospective designs, the relative performance of the five DCG models improves with the increase in the truncation threshold. This finding is particularly true for the EDCGDX and ADCGDX models, which were designed to capture the effects of

FIGURE 11
RATIO OF PREDICTED TO ACTUAL 1992 EXPENDITURES FOR NONRANDOM GROUPS OF ENROLLEES, GROUPINGS BASED ON INPATIENT DIAGNOSIS OF CANCER, HEART DISEASE, OR OTHER IN 1991



high-cost coexisting conditions on expenditures. In terms of R^2, these two models consistently outperform ACGs and ADGs with no truncation.

Finally, predictive accuracy for large random groups, as measured by the mean absolute percentage error, decreases with higher levels of truncation. However, the relative findings for the models do not differ systematically across the levels.

b. Use of Inpatient Diagnoses in ACG and ADG Models

As described above, although the ACG and ADG models used in this study were developed using diagnoses from ambulatory encounters, inpatient diagnoses could also be employed when assigning individuals to these groups. For this study, we used a strict definition of ACGs and ADGs and included only ambulatory diagnoses when assigning patients to risk groups.

We explored the sensitivity of our findings to this assumption by repeating the ACG and ADG analyses for individuals and random groups using both inpatient and ambulatory diagnoses. While we would expect these diagnoses to overlap for most episodes of care, there may be some inpatient events that are not preceded or followed by related outpatient services and therefore are not captured in the ambulatory information we employed. An emergency hospital admission is one example of such an event. Including inpatient diagnoses may enhance predictive accuracy.

Table 16 summarizes the results of this analysis. As shown, at the individual level, there is a small increase in predictive accuracy for the retrospective models when inpatient diagnoses are added. The adjusted R^2 values increase by 9 and 7% for the ACG and ADG models, respectively. In contrast, there is no measurable change in the results for the two prospective models. Finally, the predictive accuracy for large random groups is essentially the same with and without inpatient diagnoses.

The finding of increased accuracy for the retrospective ACG and ADG models when adding inpatient diagnoses—but no change for the prospective models—is interesting. It highlights both the nature of inpatient events not also recorded in ambulatory data and the type of variation in expenditures captured by retrospective versus prospective models. As noted above, inpatient episodes likely to be missed in ambulatory data would include unexpected, possibly emergency, hospital admissions. In general, such events often involve problems that are acute: they will produce higher expenditures in the year in which they occur but may have less of an impact on future claims. Given its design, some of the additional variation in expenditures explained by a retrospective model is related to such unexpected, acute events. In contrast, prospective models capture variation that is more related to chronic events and conditions that are expected to persist to a greater extent over time. Consistent with this premise, the addition of inpatient diagnoses increases somewhat the predictive accuracy for the retrospective models, but has little impact on the prospective findings. We discuss in greater detail the differences between retrospective and prospective models in Chapter VII.

We conclude that the use of only ambulatory diagnoses when assigning ACGs and ADGs for our analyses had little impact on our overall findings on predictive accuracy. There was a small increase in individual accuracy for the retrospective models with the addition of inpatient information, however, the prospective results were unchanged. The findings for random groups were also unaffected.

TABLE 15
PREDICTIVE ACCURACY FOR THREE POOLS AT DIFFERENT TRUNCATION LEVELS

Method	Individual Adjusted R^2			Mean Absolute Percentage Error (Random Group Analysis)		
	Truncated at $25K	Truncated at $50K	No Truncation	Truncated at $25K	Truncated at $50K	No Truncation
Retrospective Analysis 1992						
Age-Sex	0.038	0.030	0.017	4.2	4.7	6.7
ACGs	0.279	0.224	0.131	3.5	4.1	6.2
ADGs	0.347	0.289	0.181	3.4	4.0	6.1
PIPDCGs	0.504	0.472	0.334	3.0	3.6	5.6
EDCG	0.356	0.354	0.268	3.4	3.9	5.9
ADCG	0.254	0.242	0.197	3.6	4.0	6.1
EDCGDX	0.381	0.394	0.316	3.4	3.9	5.8
ADCGDX	0.290	0.282	0.232	3.5	4.0	5.8
Prospective Analysis						
Age-Sex	0.039	0.030	0.015	4.2	5.1	7.2
ACGs	0.091	0.071	0.040	4.4	4.9	6.8
ADGs	0.111	0.087	0.048	4.3	4.8	6.6
PIPDCGs	0.060	0.060	0.039	4.7	5.1	6.9
EDCG	0.089	0.081	0.054	4.4	4.8	6.8
ADCG	0.082	0.075	0.051	4.4	4.8	6.8
EDCGDX	0.092	0.084	0.061	4.4	4.9	6.7
ADCGDX	0.091	0.078	0.060	4.3	4.8	6.8

TABLE 16
SENSITIVITY ANALYSIS OF PREDICTIVE ACCURACY FOR THREE POOLS FOR ACG AND ADG MODELS, WITH AND WITHOUT INPATIENT DIAGNOSES, 1992

	Individual Adjusted R^2		Mean Absolute Percent Error (Random Group Analysis)	
	Without Inpatient Diagnosis	With Inpatient Diagnosis	Without Inpatient Diagnosis	With Inpatient Diagnosis
Retrospective Analysis				
ACGs	.279	.305	3.5	3.5
ADGs	.347	.371	3.4	3.3
Prospective Analysis				
ACGs	.091	.092	4.4	4.4
ADGs	.111	.111	4.3	4.3

TABLE 17
TOTAL EXPENDITURES FOR LEAVERS, JOINERS, AND STAYERS

	Mean	Standard Deviation	*N*
1991 Expenditures			
Leavers	$1,274	3,407	175,572
Stayers	1,211*	3,249	273,314
1992 Expenditures			
Joiners	$1,180	3,189	199,909
Stayers	1,334*	3,466	273,314

* Difference between mean expenditures for stayers and other group is statistically significant at the 0.01 level.

c. Leavers versus Joiners versus Stayers

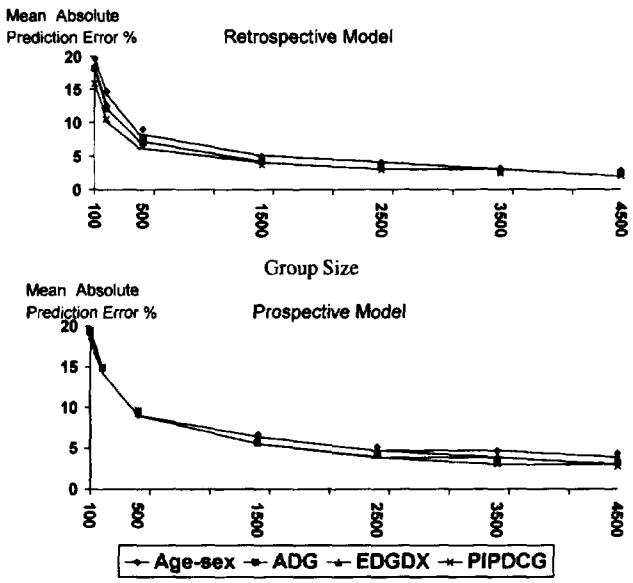
As described previously, the prospective analyses could only be performed using data for individuals in the same pool for 1991 and 1992—potentially a select sample of enrollees. To test the sensitivity of the study findings to this issue, we evaluated the retrospective results separately for the leavers (those in the pool in 1991 but not in 1992), joiners (those in the pool in 1992 but not in 1991), and stayers (those in the pool both years).[32]

Table 17 summarizes statistically the expenditures of the leavers, joiners, and stayers. The leavers and stayers are compared using 1991 expenditures, while the joiners and stayers are compared using 1992 data. As shown, for those pools analyzed (two indemnity plans and one PPO plan), a significant number of individuals either left or joined the plans described for the two

TABLE 18
RATIO OF PREDICTED TO ACTUAL EXPENDITURES FOR LEAVERS, JOINERS, AND STAYERS, BY RISK ASSESSMENT METHOD
RETROSPECTIVE MODELS

Method	Year	Stayers	Leavers	Joiners
Age-Sex	1991	1.01	0.99	
	1992	0.99		1.02
ACG	1991	1.02	0.96	
	1992	1.01		0.99
ADG	1991	1.01	0.97	
	1992	1.01		0.99
PIPDCG	1991	1.01	0.99	
	1992	0.98		1.05
EDCG	1991	1.01	0.99	
	1992	0.99		1.01
ADCG	1991	1.01	0.99	
	1992	0.99		1.02
EDCGDX	1991	1.01	0.99	
	1992	0.99		1.01
ADCGDX	1991	1.01	0.99	
	1992	0.99		1.02

FIGURE 12
EFFECT OF GROUP SIZE ON PREDICTIVE ACCURACY



years. Approximately 40% of those in the pools in 1991 were not in the pools in 1992 (leavers). A similar percentage of individuals in the pools in 1992 were not in the pools in 1991 (joiners). The mean expenditures for the leavers is similar to that of the stayers—the two amounts differing by only 5%. The difference in mean expenditures for the stayers versus the joiners is somewhat larger (13%).

We investigated the impact on the prospective findings of using the select group of individuals who remained in a pool for both years by evaluating whether the retrospective models tested produced systematic over- or underpredictions for these individuals. To do this, we computed the predictive ratios for stayers, leavers, and joiners under each model. As before, a predictive ratio greater than 1.0 represents an overprediction, while a ratio less than 1.0 is an underprediction.

The results of the analysis are summarized in Table 18. As shown, the ratios for each of the groups are similar and are all close to 1.0. There are also no systematic differences in these ratios across models.

We conclude that the study findings and conclusions for prospective models are not sensitivity to the select group of enrollees used in the analysis.

d. Effect of Group Size

We explored the sensitivity of the results for random groups to the group size chosen. To do this, we repeated the analysis using group sizes of 100, 200, 500, 1,500, 3,500, and 4,500 in addition to the group size of 2,500 we used in the analysis described previously. Figure 12 shows the results.

As expected, predictive accuracy improves with group size. However, in general, the relative performance of the models was unaffected. The additional increase in predictive accuracy is significant when moving from group sizes of 100 to 1500. However, improvements in accuracy level off after that point.

6. *The Stability of Risk Weights for Assessment Selected Methods across Pools*

We conducted the analyses of predictive accuracy described in this chapter pool-by-pool. In this way, we allowed the risk weights for each method to be computed uniquely pool-by-pool and applied them to a pool's data in making predictions. However, in

implementing a risk adjustment transfer process such as that for a purchasing alliance, the same risk weights would likely be applied to all plans participating. Further, if standard sets of risk weights can be developed, they might be used in different risk adjustment applications.

We explored the stability of risk weights across pools for three different models: age and sex, ACGs, and EDCGs. In doing this, as before, we computed the risk weights pool-by-pool (These risk weights are comparable to the estimated parameters described in equation 2, presented previously in this chapter. However, for these purposes, rather than using the split-half method employed for the investigation of predictive accuracy, we used all of the data available for a pool. In this way, we increased the likely precision of the estimates of risk weights.) For each method, we compare the weights from both the retrospective and prospective models. For the retrospective models, we use data from 1992. We present these analyses for seven pools retrospectively and five pools prospectively.

To facilitate comparison across pools, we standardized the risk weights for each method by dividing each weight by an "adjusted" mean expenditure per enrollee for the pool. This adjusted mean expenditure was computed by assuming the same distribution of enrollees by age and sex across all pools, but allowing the pools to maintain their level of expenditures per enrollee for each age and sex group. An alternative approach would have been to simply divide each risk weight by the mean expenditure per enrollee for the pool. However, this mean expenditure is likely to be sensitive to the mix of enrollees in the pool. For example, a pool comprised of older individuals would be expected to have a higher mean expenditure than one with a younger population. Our approach adjusts for these differences by imposing the same age and sex distribution across all pools across all pools. A standardized risk weight greater/lessor than 1.00 represents an expected cost greater/lessor than this adjusted mean amount.[33]

Tables 19 through 21 present the results. With a few exceptions, the findings for age and sex (Table 19) show consistent results across the pools for each age-sex group. As expected, in general, the retrospective and prospective weights are similar (since age only differs by one year in the prospective model). The risk weights for age and sex range from about 0.30 to 2.30, a multiple of approximately 7.5.

In general, for most pools consistencies were found in the weights within a given ACG (Table 20). Some differences are observed for particular ACGs. These differences could exist for a number of reasons including different medical practice, inconsistency in diagnositic coding, or data from a small number of enrollees used in computing a risk weight.

Differences in ACG weights are also observed for some pools. In particular, the weights for pool HD are consistently lower for most ACGs. The weights are also lower for many ACGs for pool HA. As discussed in greater detail in Chapter V, we observed a significantly greater intensity of diagnostic coding for ambulatory encounters for pool HD. Pool HA also has a somewhat greater intensity of ambulatory encounters. These differences in coding may occur for a number of reasons including better data systems or a different mix of services. However, for ACGs, these differences lead to a potentially greater number of ADG assignments per enrollee and a greater number of assignments to what are typically higher cost ACGs. As a result, lower expenditures per enrollee in most ACGs would be expected for these pools relative to others. This finding underscores the importance of data consistency in applying diagnosis-based assessment models across pools, as in a risk adjustment transfer process. We discuss this issue further in Chapters V and VI.

A second difference observable for the ACG risk weights are the differences between the weights for the retrospective and prospective models. In general, the range of weights is more compressed in the prospective model—about 0.3 to 5.0—than the retrospective model—about 0.2 to 8.5 (excluding ACG 52). This compression is most evident for the higher cost ACGs (40 through 50) and reflects regression toward the mean for these more expensive cases. However, both the retrospective and prospective ACG models show a considerably wider range of weights than that observed for age and sex. This result was also observed for EDCGs.

The results for EDCGs (Table 21) also show some consistencies for many DCGs.[34] As before, some differences are observed. In particular, the results for pool HD again deviate from those of other pools. As with ACGs, EDCGs employ ambulatory diagnoses in assigning patients to risk groups. The differences in ambulatory coding for pool HD impact its assignment of EDCGs and the computation of its risk weights.

For the retrospective models, the EDCG risk weights increase with the higher DCGs. This is to be expected given that the model was developed so that higher cost diagnoses would be assigned to a higher DCG. The range in weights for the retrospective models is about

TABLE 19
COMPARISON OF RISK WEIGHTS ACROSS SELECTED POOLS: AGE AND SEX MODEL (WEIGHTS STANDARDIZED TO ADJUSTED MEAN EXPENDITURE FOR POOL)

	Retrospective							Prospective						
	IA	IB	PA	PC	HA	HB	HD	IA	IB	PA	PC	HA	HB	HD
Female														
00–00	0.781	0.729	0.770	0.834	1.254	0.730	0.663	0.518	0.553	0.511		0.813		0.565
01–04	0.391	0.417	0.389	0.452	0.517	0.394	0.446	0.367	0.417	0.359		0.412		0.372
05–09	0.320	0.314	0.315	0.333	0.353	0.308	0.414	0.277	0.344	0.319		0.283		0.407
10–14	0.396	0.447	0.450	0.416	0.396	0.371	0.426	0.515	0.555	0.558		0.498		0.483
15–19	0.745	0.754	0.808	0.710	0.773	0.745	0.804	0.729	0.816	0.729		0.827		0.895
20–24	1.122	0.927	1.204	1.181	1.084	1.092	1.221	1.186	1.031	1.051		0.995		1.137
25–29	1.628	1.432	1.556	1.486	1.424	1.684	1.771	1.602	1.541	1.468		1.373		1.861
30–34	1.464	1.247	1.581	1.622	1.326	1.634	1.635	1.377	1.383	1.586		1.289		1.468
35–39	1.310	1.103	1.458	1.345	1.264	1.477	1.403	1.209	1.182	0.956		1.124		1.372
40–44	1.221	1.202	1.273	1.250	1.234	1.357	1.244	1.261	1.343	0.954		1.277		1.159
45–49	1.427	1.383	1.389	1.451	1.360	1.412	1.395	1.441	1.561	0.940		1.385		1.491
50–54	1.603	1.592	1.550	1.642	1.521	1.543	1.763	1.543	1.855	0.843		1.574		1.808
55–59	1.862	1.834	1.691	1.756	1.935	1.671	1.777	1.776	1.972	1.816		2.062		1.748
60–64	2.120	2.004	1.904	1.668	2.154	1.844	1.880	2.260	2.213	1.956		2.374		2.010
Male														
00–00	1.059	1.049	0.890	1.097	1.595	0.921	1.323	0.730	0.712	0.648		0.876		0.604
01–04	0.542	0.516	0.523	0.573	0.564	0.452	0.523	0.507	0.507	0.471		0.442		0.438
05–09	0.380	0.371	0.387	0.403	0.388	0.365	0.368	0.365	0.385	0.388		0.360		0.353
10–14	0.472	0.452	0.470	0.464	0.535	0.431	0.359	0.555	0.552	0.571		0.597		0.407
15–19	0.669	0.622	0.695	0.571	0.676	0.506	0.415	0.645	0.640	0.544		0.719		0.411
20–24	0.449	0.438	0.453	0.496	0.427	0.503	0.492	0.450	0.479	0.456		0.449		0.610
25–29	0.534	0.525	0.543	0.599	0.562	0.539	0.557	0.526	0.590	0.601		0.516		0.504
30–34	0.546	0.583	0.610	0.662	0.637	0.645	0.677	0.603	0.643	0.606		0.580		0.665
35–39	0.646	0.605	0.694	0.712	0.682	0.761	0.625	0.630	0.688	0.708		0.727		0.684
40–44	0.807	0.777	0.811	0.800	0.841	0.830	0.797	0.855	0.942	0.794		0.774		0.791
45–49	1.058	1.067	0.987	0.995	1.067	1.065	0.950	1.076	1.243	1.087		1.276		1.048
50–54	1.406	1.400	1.245	1.411	1.327	1.299	1.184	1.376	1.626	1.321		1.416		1.191
55–59	1.734	1.892	1.696	1.697	1.756	1.599	1.632	1.920	2.154	1.895		1.752		2.021
60–64	2.419	2.038	2.121	1.910	2.318	1.978	1.906	2.448	2.449	2.222		2.567		1.543
No. of Enrollees	153,525	150,812	169,512	204,350	20,097	120,964	18,497	87,556	120,392	66,015		14,219		14,855

TABLE 20
COMPARISON OF RISK WEIGHTS ACROSS SELECTED POOLS: ACG MODEL
(WEIGHTS STANDARDIZED TO ADJUSTED MEAN EXPENDITURE FOR POOL)

ACG	IA	IB	PA	PC	HA	HB	HD
			Retrospective				
1	0.270	0.274	0.227	0.472	0.658	0.289	0.318
2	0.253	0.177	0.181	0.261	0.204	0.196	0.159
3	0.375	0.346	0.324	0.354	0.314	0.368	0.153
4	0.936	0.754	0.802	0.941	0.852	0.810	0.368
5	0.607	0.449	0.459	0.593	0.406	0.532	0.205
6	0.678	0.552	0.668	0.612	0.608	0.718	0.361
7	0.744	0.500	0.535	0.631	0.321	0.634	0.190
8	1.765	1.456	1.360	1.957	1.416	1.368	0.822
9	0.723	0.500	0.534	0.577	0.446	0.563	0.269
10	0.633	0.507	0.469	0.741	0.598	0.552	0.276
11	0.571	0.344	0.460	0.329	0.244	0.325	0.078
12	0.944	0.563	0.591	0.650	0.650	0.776	0.171
13	0.633	0.472	0.514	0.628	0.273	0.451	0.184
14	1.356	1.135	1.283	1.609	0.663	1.322	0.512
15	2.256	2.300	2.223	1.982	1.133	2.039	2.361
16	0.469	0.292	0.325	0.340	0.246	0.341	0.084
17	3.730	3.144	3.440	4.320	3.019	3.363	3.688
18	1.353	1.080	1.119	1.137	1.092	1.168	0.624
19	0.637	0.619	0.602	0.754	0.832	0.634	0.591
20	0.540	0.536	0.476	0.555	0.510	0.556	0.387
21	0.868	0.763	0.722	0.811	0.651	0.829	0.386
22	0.998	0.812	0.883	0.826	0.928	1.006	0.552
23	0.872	0.903	0.828	0.776	0.787	0.657	0.441
24	0.774	0.438	0.562	0.391	0.472	0.434	0.198
25	0.848	0.696	0.614	0.744	0.581	0.611	0.310
26	1.460	1.468	1.455	1.596	1.009	1.793	0.912
27	2.176	2.511	2.228	2.548	1.020	1.434	0.726
28	1.458	1.230	1.148	1.494	1.157	1.453	0.986
29	1.293	1.237	1.088	1.331	1.533	1.117	0.919
30	1.058	1.171	0.858	1.109	0.784	1.174	0.696
31	1.074	1.092	1.027	1.139	1.112	1.071	0.785
32	2.258	1.839	1.822	1.990	1.464	2.027	1.296
33	2.159	1.701	1.740	1.793	1.547	2.374	1.225
34	1.220	0.890	1.066	0.832	0.805	0.618	0.436
35	1.848	1.621	1.434	1.740	1.328	1.748	1.025
36	3.088	2.817	2.610	2.830	2.668	2.758	1.838
37	3.080	3.031	2.752	2.896	2.393	2.925	1.690
38	1.202	1.153	1.033	1.129	0.881	1.294	0.521
39	1.958	1.694	1.593	1.644	1.319	1.769	0.680
40	2.840	2.220	2.460	2.744	1.923	2.599	1.496
41	2.175	1.707	1.704	2.064	1.509	1.990	0.740
42	2.026	1.808	1.472	2.190	1.442	2.195	0.891
43	3.324	3.014	2.843	3.132	2.468	3.337	1.875
44	3.779	3.388	3.119	3.870	2.838	3.772	1.717
45	4.200	3.734	2.883	3.373	3.570	4.130	2.659
46	3.306	3.576	2.842	3.773	2.766	4.040	1.948
47	5.612	4.975	4.763	5.054	5.023	6.302	3.189
48	4.895	4.670	4.437	5.327	3.734	5.410	3.463
49	5.542	5.206	4.688	6.219	4.816	5.849	3.665
50	9.688	8.094	7.389	7.728	8.361	9.992	7.371
51	1.370	0.702	1.348	0.276	0.623	0.203	0.131
52	0.072	0.056	0.072	0.000	0.059	0.000	0.003
No. of Enrollees	153,525	150,812	169,512	204,350	20,097	120,964	18,497

TABLE 20—*Continued*
COMPARISON OF RISK WEIGHTS ACROSS SELECTED POOLS: ACG MODEL
(WEIGHTS STANDARDIZED TO ADJUSTED MEAN EXPENDITURE FOR POOL)

ACG	IA	IB	PA	PC	HA	HB	HD
			Prospective				
1	0.331	0.314	0.270		0.351		0.157
2	0.305	0.242	0.318		0.295		0.392
3	0.742	0.673	0.631		0.641		0.413
4	0.749	0.674	0.668		0.516		0.554
5	0.669	0.659	0.632		0.616		0.605
6	0.949	0.717	0.845		0.431		0.638
7	0.667	0.700	0.608		0.592		0.547
8	1.847	2.108	1.290		1.604		1.533
9	1.156	1.163	1.003		1.048		0.791
10	1.079	0.832	1.433		0.148		1.015
11	0.544	0.546	0.692		0.448		0.292
12	1.096	0.950	0.933		0.827		0.340
13	1.069	1.162	0.618		0.912		0.770
14	1.073	1.350	1.208		1.105		0.656
15	1.062	1.321	1.974		0.526		0.312
16	0.781	0.672	0.624		0.460		0.336
17	2.150	1.889	1.947		1.798		0.797
18	1.097	0.993	0.818		0.865		0.626
19	0.540	0.506	0.529		0.572		0.394
20	0.470	0.497	0.420		0.446		0.395
21	0.932	0.845	0.766		0.849		0.620
22	0.813	0.894	0.948		0.489		0.929
23	1.206	1.307	1.361		0.900		0.773
24	0.746	0.960	1.060		0.973		0.504
25	1.277	1.272	1.102		1.881		0.521
26	1.357	1.606	1.350		1.013		0.632
27	2.141	1.130	2.467		1.640		0.271
28	1.007	1.071	0.818		0.942		0.761
29	0.591	0.756	0.568		0.818		0.524
30	0.654	0.565	0.564		0.551		0.505
31	0.592	0.742	0.571		0.704		0.539
32	1.575	1.606	1.381		1.441		1.003
33	1.840	1.159	2.044		2.866		1.432
34	1.658	1.056	1.045		0.910		0.498
35	1.804	1.594	1.629		1.169		1.016
36	2.197	2.260	1.949		1.742		1.691
37	2.329	2.526	2.138		2.419		1.360
38	0.875	0.940	0.906		0.745		0.556
39	1.012	1.203	0.994		0.746		0.713
40	2.026	1.959	1.815		1.292		1.582
41	1.694	1.792	1.543		1.454		1.357
42	1.331	1.447	1.055		1.204		0.858
43	2.220	2.199	1.974		1.703		1.533
44	2.734	2.893	2.222		2.497		1.751
45	2.504	2.525	1.626		2.174		0.938
46	2.353	2.774	1.921		1.290		1.242
47	3.244	2.972	1.939		4.156		1.935
48	3.261	3.402	2.827		2.845		2.583
49	3.764	3.710	3.148		3.030		2.563
50	5.357	6.729	4.840		4.677		4.202
51	0.852	1.103	1.006		0.461		0.197
52	0.464	0.545	0.448		0.441		0.336
No. of Enrollees	87,556	120,392	66,015		14,219		14,855

TABLE 21
COMPARISON OF RISK WEIGHTS ACROSS SELECTED POOLS: EDCG MODEL*

	Pool						
EDCG	IA	IB	PA	PC	HA	HB	HD
				Retrospective			
1	0.213	0.189	0.158	0.144	0.266		0.145
2	0.442	0.391	0.299	0.431	0.318		0.060
4	2.651	2.498	2.077	2.502	2.270		0.963
5	6.801	6.276	6.642	7.230	4.914		6.840
6	6.736	6.127	5.829	6.557	6.086		3.684
7	11.054	11.195	11.218	12.747	12.000		7.495
8	7.666	7.546	7.114	8.716	7.502		6.316
10	9.583	9.075	8.775	11.824	7.854		6.158
12	10.167	9.257	9.709	14.843	11.222		9.111
14	11.950	10.775	9.171	13.576	11.189		11.554
17	16.630	14.367	15.118	18.241	16.912		17.595
23	17.626	15.410	17.342	21.649	17.047		17.618
No. of Enrollees	153,525	150,812	169,512	204,350	20,097		18,497
				Prospective			
1	0.678	0.820	0.713		0.442		0.580
2	0.110	0.125	0.163		0.102		0.050
4	0.942	1.149	0.836		0.727		0.290
5	1.621	1.982	1.185		2.038		1.794
6	2.041	2.879	1.602		1.606		0.909
7	1.989	3.663	2.258		2.381		1.255
8	2.608	2.927	2.548		2.063		0.926
10	2.147	3.553	1.650		1.871		2.383
12	2.936	5.167	2.236		7.572		6.197
14	4.999	6.180	4.013		2.035		5.336
17	10.003	6.837	7.457				10.446
23	12.574	11.783	11.035				15.448
No. of Enrollees	87,556	120,392	66,015		14,219		14,855

*Since age and sex were also included in the EDCG model we tested, risk weights are reported for a single age-sex group—males 45–49 years of age.

0.2 to 17.5. In comparison, the prospective risk weights increase more gradually and exhibit a few unexpected changes in rank order. The range in weights for these models is about 0.1 to 12.0.

The findings in Tables 19 to 21 show some consistencies in risk weights across different sources of data. This is particularly true for the age and sex model. However, we did observe some differences across pools for the diagnosis-based methods tested—the most significant ones due to differences in the intensity of coding of ambulatory diagnoses. There may ultimately exist the potential for the development of standard risk weights for risk assessment models. However, given the likely differences in the recording of medical encounters and differences in medical practice across plans and geographic areas, care should be taken in applying a set of risk weights to different plans and populations.

E. Summary

In this section, we summarize the relative findings on predictive accuracy for the methods compared. A more general discussion of the implications of these findings is reserved for Chapter VII. Table 22 summarizes the results for predictive accuracy:

1. Individuals

- All diagnosis-based methods outperform age and sex at the individual level, by a wide margin. Age and

sex explains less than 10% retrospectively and 33% prospectively of the variance in individual expenditures predictable by the best diagnosis-based models.

- The differences among the diagnosis-based models are less pronounced.
- PIPDCG performs best retrospectively, explaining 43% of the variation in individual expenditures.
- Prospectively, ADGs produces somewhat better predictions than the other methods. However, even ADGs leave unexplained almost 90% of the variation in individual expenditures.
- Within the ACG models, the additional level of clinical detail available in ADGs allows it to consistently outpredict ACGs in all applications.
- Within the DCG models:
  - — Although PIPDCGs perform best retrospectively, they fall short, prospectively, of the models that all also include ambulatory diagnoses.
  - — EDCGs distinguish inpatient from outpatient diagnoses, while ADCGs do not. This additional information in the EDCG model increases predictive accuracy by 36% retrospectively and 8% prospectively relative to ADCGs.
  - — The addition of high-cost coexisting conditions to ADCGs and EDCGs improve their predictive accuracy by 10 to 15% retrospectively and 5 to 10% prospectively. The addition of these conditions causes a greater marginal improvement for the ADCG model.
- All diagnosis-based models predict much better retrospectively than prospectively.
- The findings on predictive accuracy were robust with respect to both health care management type and the population of enrollees. Consistent findings were observed across pools, including indemnity, PPO and HMO plans.

2. Random Groups

- All models, including age and sex perform well for large random groups of enrollees. For random groups of 2,500 enrollees, the typical prediction error was 3 to 4% of mean expenditures.
- Predictive accuracy increases with group size, particularly for groups under 2,000, after which the increase in accuracy levels off. Even for groups of 500, the models produce an average error of less than 10%.
- There is little difference in predictive accuracy for random groups between retrospective and prospective models.

3. Nonrandom Groups

- All models performed poorly for nonrandom groups. In general, the models produced biased predictions for groups selected based on their prior expenditures, The methods overpredict expenditures for those individuals with relatively low expenditures in the previous year and underpredict expenditures for those with higher expenditures in the previous year. This bias increases with the size of previous expenditures. ADGs performed best for these groups, while age and sex produced over- or underpayments markedly greater than any other model.
- For nonrandom groups constructed based on previous high-cost medical conditions, all models produced

TABLE 22
COMPARISON OF PREDICTIVE ACCURACY OF RISK ASSESSMENT METHODS

Risk Assessment Method	Prospective						Retrospective					
			Nonrandom Groups Predictive Ratio						Nonrandom Groups Predictive Ratio			
			Previous Expenditures		Medical Conditions				Previous Expenditures		Medical Conditions	
	Individual R^2	Mean Absolute Percentage Error	Low	H2	Heart	Cancer	Individual R^2	Mean Absolute Percentage Error	Low	H2	Heart	Cancer
Age-Sex	0.039	4.3	1.69	0.38	0.38	0.22	0.032	4.6	1.70	0.37	0.30	0.18
ACG	0.091	4.2	1.16	0.66	0.62	0.38	0.286	3.8	1.11	0.68	0.55	0.42
ADG	0.112	4.2	1.07	0.73	0.78	0.46	0.357	3.6	1.07	0.74	0.73	0.56
PIPDCG	0.061	4.6	1.57	0.67	0.85	0.70	0.428	3.5	1.43	0.61	0.66	0.56
EDCG	0.091	4.4	1.26	0.68	0.77	0.63	0.343	3.6	1.22	0.73	0.80	0.76
ADCG	0.084	4.3	1.34	0.62	0.70	0.50	0.252	3.8	1.29	0.68	0.70	0.67
EDCGDX	0.096	4.3	1.25	0.71	0.84	0.78	0.372	3.7	1.21	0.77	0.89	0.92
ADCGDX	0.093	4.3	1.33	0.65	0.86	0.71	0.293	3.8	1.27	0.72	0.93	0.80

underpayments. The models would, on average, underpredict the risk of individuals with an inpatient admission for heart disease or cancer in the previous year. The EDCGDX model performed best for both conditions. Again, age and sex produces a significant bias in payments.

- Although retrospective models provide better predictions for individuals than prospective models, retrospective and prospective models produced similar findings for nonrandom groups.

4. *Other Findings*

- The results for predictive accuracy were robust with respect to health care management type and the population of enrollees studied. The relative performance of the models is consistent across pools of data describing indemnity, PPO, and HMO plans and including a demographically diverse population and potentially different benefit packages.
- We observed some consistency in the estimated risk weights across pools, particularly for the age and sex model. However, because these risk weights reflect differences in diagnostic coding and medical practice across plan types and geographical areas, care should be taken in applying a set of risk weights to different plans and populations.
- We truncated expenditures at $25,000 for our primary investigation of predictive accuracy. The results using a truncation threshold of $50,000 and no truncation showed an increase in the relative performance of the DCG models with higher thresholds.

END NOTES

1. We explored the use of different age-sex groupings for the analysis, including wider age ranges and the interaction of enrollee age in years with sex. These alternative groupings had little impact on the study results.
2. Although the current ACG model was developed using diagnoses from ambulatory encounters, inpatient diagnoses could also be included when assigning ACGs. For this study, we employed a strict interpretation of ACGs and used only ambulatory diagnoses for assigning patients to these groups. We discuss the implications of this assumption later in this chapter.
3. As with many of the risk assessment methods, both ACGs and DCGs are continually undergoing revision and refinement. We used the most current version of each method available. Our description of these methods and the study findings are based on these versions. At the time of this report, research on ACGs is focusing, among other things, on incorporating inpatient diagnostic information into the ACG model and applying the model to the Medicare population. For DCGs, research is now being undertaken to refine the groupings of diagnoses within each DCG and to further identify coexisting high-cost conditions for the ADCGDX and EDCGDX models. The application of DCGs to nonelderly populations is also being explored. (Note that ACGs were initially developed using data from a nonelderly population, while DCGs were developed using Medicare data.) The individuals undertaking research on these models should be contacted about their current status.
4. The criteria used for assigning ICD9 codes to ADGs were developed by a team of clinicians and were selected based on their status as determinants of anticipated health care resource use. The criteria included: likelihood of the persistence or recurrence of the problem; likelihood of continued services and treatments and their expected cost; likelihood of the need for specialist's services; likelihood of decreased life expectancy; and likelihood of short- or long-term patient disability.
5. The version of DCGs used in this study was developed based on the expected relationship between this year's diagnoses and future health care needs, or what we refer to in this report as a "prospective" assessment model. In contrast, ACGs were initially developed using a retrospective, or concurrent, design—this year's diagnoses and this year's use and costs. As discussed below, we test each of these models using both prospective and retrospective applications.
6. As with ACGs, diagnoses for a period of one year are typically used in assigning persons to a DCG.
7. Of the five DCG models we evaluated, the PIPDCG model is the most similar to the earlier versions of DCGs which also employed only inpatient diagnoses in assigning DCGs (Ash et al., 1989).
8. Table B-4 in Appendix B also shows, for the PIPDCG model, that 35% of the individuals with one or more inpatient admissions were assigned to the same lowest cost DCG as that for those individuals without a hospital admission. This percentage reflects the inclusion of those admissions with lower expected costs in the lowest cost DCG.
9. One limitation of the study database is the omission of diagnoses other than the principal for inpatient and ambulatory encounters. As described above, both the ACG and DCG models also employ subsidiary diagnoses in assigning individuals to risk groups. It is expected that if these diagnoses were available, the predictive accuracy of both the ACG and DCG models would have improved. At this time, it is uncertain what the size of this improvement would have been, although it would likely have been small. The relative impact on

the findings for the models also would be expected to be small. Finally, the use of secondary and other diagnoses in assessing risk raises a number of practical concerns such as those discussed in Chapters V and VI.

10. Plans were asked to exclude data for those individuals enrolled for less than the full year. Included in this group are persons who died during the year. Some of these individuals might be expected to experience significant expenditures in the time period immediately preceding death. However, given the low percentage of these persons in a population of nonelderly enrollees, it is unlikely their exclusion significantly influenced the study findings. This issue is of greater concern for risk assessment for the Medicare population.
11. In addition to management approach and level of deductible, carriers also noted on each record whether the insurance policy was based on group or individual underwriting, and if individual underwriting, whether it was first year or renewal. Tabulations showed employer group underwriting to represent an overwhelming majority of the individuals included in the data from any carrier. Given the small numbers of enrollees with individual underwriting in the sample, we did not analyze these records separately. Instead, we combined all individuals within a plan type when constructing a pool of data for analysis.
12. There was some discussion among the project advisors regarding the use of the AAPCC to adjust expenditures. Given that the AAPCC is based on consistent data covering all geographic areas and that good alternatives are unavailable, there was some consensus that it should be used for the study. In response to concerns about the AAPCC, we scrutinized our findings for areas with high HMO penetration (for example, southern Florida, Minnesota, and southern California). Exploratory analyses suggested that the results for these areas, at least, are not very sensitive to the use of the AAPCC.
13. The adjustments for differences across geographic areas and across years were made in keeping with the objective of evaluating alternative risk assessment methods rather than estimating weights for a particular pool or region.
14. We used the split-half design for all analyses and models described here, including both retrospective and prospective applications.
15. In some ways, the individuals in the same pool for both years can be considered a somewhat select sample. To test the sensitivity of the study findings to this issue, we conducted selected analyses separately for the leavers, joiners, and "stayers." These analyses are described below.
16. For confidentiality reasons, pools are not identified by carrier in this report. Instead, we use a two-digit code to identify pools. The first digit represents plan type (I = Indemnity, P = PPO, H = HMO). The second digit identifies unique pools within plan type.
17. As described in Section III-B, only the principal diagnosis for each ambulatory encounter or inpatient stay was recorded for the study data. As a result, no subsidiary diagnoses were used in assigning ACGs, ADGs, and DCGs.
18. The use of multiple regression to estimate the model parameters can be compared with simply taking, risk group by risk group, the mean annual expenditures of the individuals in a group. In all cases, both methods provide the same result. Further, for the two models involving a reasonable number of mutually exclusive groupings—age-sex and ACGs—computing mean expenditures by group would have been straightforward computationally. However, for the remaining models, where the groupings may not be mutually exclusive (ADGs and the coexisting conditions in the DCG models) and where age-sex groupings are applied in addition to the other risk assessment criteria (for example, PIPDCGs, with age and sex), using multiple regression greatly simplified computation of the models' parameters.
19. For actuaries most familiar with the age-sex model, these predictions are similar to applying average costs for each age-sex classification to a group of individuals in order to achieve the predicted expenditures.
20. More technically, R^2, or the coefficient of multiple determination, is equal to the 1 − (SSR/TSS), where SSR is the sum of the squares of the regression residuals and TSS is the total sum of the squares, or sum over i of $(X_i - \text{mean}(X))^2$. The adjusted R^2 is an alternative to R^2 that is adjusted for the number of parameters in the model and is equal to $1 - [(n^*(1-R^2))/(n-p)]$, where n is the number of observations used to estimate the model and p is the number of model parameters. In the analyses reported here, n is always very large relative to p, so that the difference between the R^2 and adjusted R^2 is quite small.
21. A third limitation one might bring up is the sensitivity of R^2 to the variance of the sample of individuals. But in the present context this is unimportant, as the data on which the comparisons are made are the same for each risk-adjustment method (that is, pools of data).
22. The choice of a group size of 2,500 was based on the observed relationship between group size and predictive accuracy. As expected, we found the predictive accuracy for groups under all risk assessment methods to increase with the size of the group. However, beyond a group size of 2,000 to 3,000 individuals, this increase in accuracy levels off. Groups of 2,500 are thus fairly representative of large groups in general. We explore the effect of group size on predictive accuracy later in this report.
23. There was considerable discussion among the project advisors about whether selection with or without replacement

was more appropriate. From a statistical perspective, it was agreed that the analysis would be simpler if groups were selected with replacement. (Without replacement, the sample gets smaller and smaller and is different for each successive group selected.) From a practical perspective, there was less agreement. Some argued that, within each market, there is only one of each individual and this individual ends up in only one plan. For example, an extremely high-cost individual will show up only once, in the pool of a single carrier, and should not be present multiple times for an analysis. Others argued that the experiment described by the random group analysis is attempting to simulate the distribution of likely group prediction errors that would occur in practice with a similar population. Since we don't have the entire population, we are attempting to infer the results for the larger universe by performing repeated experiments on the sample data (drawing repeated groups of a certain size). For each of these experiments (groups) to be comparable, sampling with replacement should be done. In other words, they should all be starting from the same point, being drawn from the same sample of data. The consensus (not necessarily unanimous) was that we should use group selection with replacement as the primary method for the project. Some exploratory analyses suggested that the general conclusions of the study are not sensitive to this choice.

24. An alternative approach in selecting nonrandom groups would be to include those individuals in high-risk pools of "last resort," for example, individuals who cannot access health insurance on an individual basis because of failure to pass medical underwriting screens. Unfortunately, the study data did not support this type of analysis.
25. It would have been preferable to employ more than two years of data for identifying nonrandom groups and evaluating each models' performance. For example, three or four years of data would be more useful in identifying those individuals who are consistently of high or low risk, including those with expensive, chronic conditions. However, the approach employed in this study, using the first year of data to identify nonrandom groups and the second year to assess predictive accuracy, should provide insights into the models' expected performance for these types of individuals.
26. The cancer and heart high-cost condition groups were defined by using inpatient ICD9 diagnosis codes and aggregating the high-cost condition groups we developed as described in Chapter IV. The ICD9 codes included in each group are presented in Appendix Table B-6 of this report.
27. We also explored the use of the natural logarithm of expenditures as a dependent variable. Although this approach addresses the skewness of the distribution of expenditures, we encountered difficulties in evaluating its results on the untransformed scale (actual dollars). Further, such an approach may have less intuitive appeal and present conceptual problems for some when applied in a risk adjustment process. Another alternative for addressing the distributional properties of health expenditures is a multipart model such as that proposed by Robinson et al., 1991. We did not investigate this alternative.
28. All monetary amounts in this report are expressed in 1992 dollars.
29. As a comparison, a retrospective model including only a dummy variable indicating whether an individual had any claims or not (expenditures or not) produced an adjusted R^2 of approximately 0.08.
30. For comparison purposes, a retrospective model including only a dummy variable indicating whether an individual had an inpatient admission or not during the year produced an adjusted R^2 of approximately 0.50. The reason this model performs better than PIPDCGs, which not only recognizes an inpatient stay but also distinguishes between stays of different expected costs, is that the PIPDCG model classifies lower cost admissions into the lowest cost DCG—the same DCG as those persons without an inpatient stay. Prospectively, this model produced an R^2 of 0.01.
31. One possible explanation for the enhanced relative performance of the prospective models on this measure may derive from their inherent design, where the risk information for the first year (1991) is used to predict expenditures for the following year (1992). The prospective model captures the relationship between these two sets of information. Since we are using expenditures for 1991 to create the three nonrandom expenditure groups and the risk information for 1991 used to predict expenditures is likely to be correlated with these amounts, the prospective models may have some advantage in the analytical design employed. Additional explanation is provided by a further analysis which showed that the variation in the predictive ratios for individuals within each of these three groups was greater for the prospective models—a result consistent with the findings on individual predictive accuracy.
32. Since leavers and joiners are only in a pool for one year, we could not conduct prospective analyses for these individuals. Instead, we infer the likely impacts on the prospective results from differences in the retrospective findings.
33. The objective of this analysis is not to estimate a standard set of risk weights to be used with each method. Instead, we present this comparison to explore the potential for standardizing estimates across different sources of data and to highlight the issues in estimating risk weights. Given the nature of the data and methods

used in this study, including the potential for unreported expenditures, adjustments for geographic areas, and the truncation of expenditures at $25,000 for the investigation of predictive accuracy, care should be taken in applying any of the risk weights we report in this chapter.

34. Since age and sex were also included in the EDCG model we tested, the risk weights reported in Table 21 are for a single age and sex group—males 45-49 years of age. This group was chosen due to the similarity of its mean expenditures to that for all enrollees.