Summary and Discussion

This study is unique in a number of ways. First, it provides a side-by-side comparison of competing models of health risk assessment using uniform methods and the same population of enrollees. Second, the enrollee data for the study come from a diverse population and describe a variety of health plans covering a range of care management approaches and all geographic areas of the country. Finally, while many studies have focused on risk assessment for elderly persons, this research examines risk assessment for a nonelderly population, including both children and adults.

We established three main objectives for the study:

1. Compare the predictive accuracy of different risk assessment methods
2. Compare the different risk assessment methods based on other criteria, including administrative practicality, ability to resist manipulation and gaming, and incentives for efficiency
3. Explore the potential for risk adjustment using a list of high-cost conditions.

Using detailed data describing the demographic characteristics, diagnoses, medical utilization and expenditures for a large number of nonelderly individuals over a two-year period, we tested the predictive accuracy of eight different risk assessment models: a simple age-sex model, two different ACG models, and five DCG models. We also evaluated these models using other criteria including the feasibility of their implementation and the incentives they provide. In exploring the practical issues, we simulated a risk adjustment transfer process across plans using the different risk assessment methods. Finally, we developed and tested an alternative risk assessment model using a list of high-cost conditions.

In this chapter, we summarize the key findings of our investigation and discuss their implications.

A. Summary

1. Predictive Accuracy

The study findings on predictive accuracy are summarized in Table 31. At the individual level, all diagnosis-based methods outperform age and sex by a wide margin. Age and sex explains only 7% retrospectively and 33% prospectively, of the variance in individual expenditures predicted by the best diagnosis-based models. The differences among the diagnosis-based methods are less pronounced. PIPDCGs perform best for retrospective applications, while ADGs produce somewhat better prospective predictions. All diagnosis-based models predict much better retrospectively than prospectively.

All models, including age and sex, perform well for large random groups. Using groups of 2,500 enrollees, the typical prediction error was 3 to 4% of mean expenditures. We observed little difference in this measure between prospective and retrospective models.

In contrast to random groups, all models performed poorly for nonrandom groups of enrollees. In general, the models produced biased predictions for groups selected based on their prior expenditures. The methods overpredicted expenditures for persons with low expenditures in the previous year, and underpredicted expenditures for those in the opposite situation. For the high-expenditure groups, this bias increased with the size of the previous expenditures. ADGs performed best for these groups, while age and sex produced over- and underpayments markedly greater than any other model.

We also constructed nonrandom groups based on previous medical conditions. We found all models would underpredict the risk of those enrollees with an inpatient admission for heart disease or cancer in the previous year, although some models, in particular

TABLE 31
COMPARISON OF PREDICTIVE ACCURACY OF RISK ASSESSMENT METHODS

Risk Assessment Method	Prospective			Retrospective		
	Individuals	Random Groups	Nonrandom Groups*	Individuals	Random Groups	Nonrandom Groups*
Age-Sex	Poor	Very Good	X-Very Poor M-Very Poor	Poor	Very Good	X-Very Poor M-Very Poor
ACG	Good/Fair	Very Good	X-Fair M-Poor	Very Good	Very Good	X-Fair M-Poor
ADG	Good/Fair†	Very Good	X-Fair† M-Fair/Poor	Very Good	Very Good	X-Fair† M-Fair/Poor
PIPDCG	Fair	Very Good	X-Poor M-Fair	Very Good†	Very Good	X-Poor M-Fair/Poor
EDCG	Good/Fair	Very Good	X-Fair M-Fair	Very Good	Very Good	X-Fair M-Fair
ADCG	Good/Fair	Very Good	X-Fair/Poor M-Fair/Poor	Very Good	Very Good	X-Fair M-Fair
EDCGDX	Good/Fair	Very Good	X-Fair M-Good†	Very Good	Very Good	X-Fair M-Good†
ADCGDX	Good/Fair	Very Good	X-Fair/Poor M-Fair	Very Good	Very Good	X-Fair M-Good/Fair

*X = Nonrandom expenditure groups, M = Nonrandom medical condition groups.
†Best model(s) for the criteria.

EDCGDX, would come relatively close for both conditions. Again, age and sex produces the most significant bias in payments.

Our findings on predictive accuracy were quite robust to both health care management type and the population of enrollees studied. The relative performance of the models is very consistent across the 19 pools of data we analyzed. These pools describe a diverse population and include indemnity, PPO, and HMO plans.

Finally, we truncated expenditures at $25,000 for our investigation of predictive accuracy. An analysis conducted using higher thresholds for truncation ($50,000 and no truncation) showed that the relative performance of the DCG models improved, particularly when moving to no truncation.

2. General Considerations

Table 32 compares the risk assessment models based on considerations other than predictive accuracy.

The age and sex model compares favorably under all general criteria. From a practical standpoint, it is easy to administer at a relatively low cost. It is also immune to gaming and straightforward to audit. In general, age and sex provides no disincentives for efficient and high-quality care.

Some differences among the diagnosis-based methods were observed. The models requiring ambulatory diagnoses face significantly greater practical problems than those using only inpatient data. This was particularly evident in our simulation of risk transfer payments using different assessment methods, where the

TABLE 32
COMPARISON OF RISK ASSESSMENT MODEL—GENERAL CONSIDERATIONS

Risk Assessment Method	Practicality/ Administrative Cost	Ability to Restrict Manipulation	Timeliness and Predictability	Incentives for Quality and Efficiency	
				Prospective	Retrospective
Age-Sex	Very Good	Very Good	Very Good	Very Good	Very Good
ACG	Fair/Poor	Fair	Good/Fair	Good	Fair
ADG	Fair/Poor	Fair	Good/Fair	Good	Fair
PIPDCG	Good	Good	Good	Good	Fair/Poor
EDCG	Fair/Poor	Fair	Good/Fair	Good	Fair
ADCG	Fair/Poor	Fair	Good/Fair	Good	Fair
EDCGDX	Fair/Poor	Fair	Good/Fair	Good	Fair
ADCGDX	Fair/Poor	Fair	Good/Fair	Good	Fair

results were extremely sensitive to the apparent quality and completeness of the ambulatory data used. Inpatient diagnoses are recorded in a more uniform way and are also easier to audit than ambulatory data, which present significant monitoring problems.

The link between our findings of differences in risk weights across some pools (Chapter III) and the sensitivity of risk adjustment transfer amounts to data consistency and quality (Chapter V) underscores an important practical point in applying risk assessment models and formulas to risk adjustment. The risk assessment weights estimated for a plan or group of plans may not be applicable to all plans. They are likely to reflect specific data systems, patterns of medical care, benefit package design, and the assumptions used in computation. Care should be taken in applying a set of risk weights developed for one application of risk assessment to another.

All diagnosis-based models face similar issues of the timing and predictability of payments. In comparison to age and sex, these methods require significantly greater time in performing the data collection and analysis required for risk assessment.

There are some incentive issues for both ACG and DCG models. For ACGs, additional ambulatory encounters may result in a higher risk group assignment and payment. In contrast, DCGs are based on only the single, highest cost diagnosis observed. On the other hand, some DCG models distinguish between inpatient and ambulatory diagnoses when assigning risk. This distinction may provide incentives to treat patients on an inpatient rather than outpatient basis, particularly in a retrospective application. The DCG models do exclude a significant number of lower cost inpatient diagnoses from assignment to a higher DCG. Individuals with these diagnoses are assigned to the same DCG as similar individuals without an inpatient admission. As a result, in practice the incentives to admit patients with some DCG models may be small. However, it is unclear to what extent they have been completely removed.

We also noted significant differences between prospective and retrospective applications of the models in terms of the incentives for efficient care. In particular, since retrospective risk assessment more closely reflects the actual claims experience of a plan, it provides greater incentives to game costs or utilization in order to achieve a higher risk payment at the end of the year. Across all models, prospective risk assessment clearly provides greater incentives for providing efficient care.

Finally, we determined there are no significant differences between the assessment methods in terms of incentives for providing high-quality care. Given their greater predictive accuracy, retrospective applications of the models may provide fewer incentives to withhold care, particularly for higher cost cases where prospective risk payments may be insufficient.

3. High-Cost Individuals and Conditions

Using a number of criteria, we developed a fairly extensive list of high-cost conditions and tested them as an alternative to age and sex and the diagnosis-based models we investigated. Using untruncated data and a retrospective model, we found that adding the high-cost conditions to the age-sex and ADG models substantially improved their predictive accuracy. However, our list of high-cost conditions did not predict expenditures as well as either the PIPDCG or the EDCGDX models.

B. Study Limitations

In spite of the strength of the methods and data employed, this study has some limitations that should be considered when interpreting its results. First, in addition to age and sex, we considered only risk assessment models that employ diagnosis-based information available through computerized records collected commonly by many insurance companies. We did not explore alternative classes of models such as those involving medical underwriting or methods based on self-reported health status, such as that obtained using the SF-36. We did not revisit the potential of physiologic measures of health status, nor did we examine whether some measures of utilization other than inpatient admissions might be used for risk assessment without unacceptably compromising incentives for efficiency. Previous research reviewed in Chapter II suggests, however, that our essential conclusions would remain unchanged if we had.

Second, there may be limits to the generalizability of our findings. Our data explicitly excluded elderly persons and those individuals who died during the course of a year. Further, although our study sample included a diverse group of nonelderly individuals, we did not have data from Medicaid populations. A large portion of our data was also restricted to commercial carriers. However, previous research and the consistency of our

findings across the pools of data we studied suggest our findings can be generalized, particularly to nonelderly populations.

Third, our investigation focused on risk assessment of total health expenditures as part of a process of setting premiums and risk-adjusting payments between health plans. Although many of our findings can be generalized to other applications of risk assessment, we did not explicitly explore the issues of risk assessment for setting provider capitation rates, profiling physicians, and performing research on outcomes measurement.

Fourth, the list of high-cost conditions we developed can be considered an initial step. Further attention to the identification of high-cost conditions for risk assessment, and the criteria by which they are selected, could produce a better list. Nonetheless, our finding that the PIPDCG and EDCGDX models outperformed our list of high-cost conditions by a reasonable margin suggests that these models may already capture much of the predictive information that such a list could include.

Finally, we used the current version of each of the ACG and DCG models tested. As research continues on these models, it is possible that their predictive accuracy can be enhanced relative to that found in this study. However, it is unclear whether further refinements of diagnosis-based risk assessment methods alone will be sufficient to eliminate the incentives for performing risk selection, particularly in light of our results for individuals and nonrandom groups. We discuss this point further below.

C. Discussion and Conclusions

We identified several requirements for a good risk assessment method. It should:

- Predict health costs with accuracy
- Be practical and of reasonable administrative cost
- Limit the ability of health plans to financially "game" the system
- Allow timeliness and predictability in setting premiums and determining transfer payments
- Provide appropriate incentives for efficient and high-quality care.

In terms of predictive accuracy, we found that diagnosis-based methods clearly outperformed a simple demographic model based on age and sex. This finding is particularly true for individual enrollees and for the nonrandom groups we tested, for which a significant bias in payments would have resulted if risk assessment had been based solely on age and sex. Simple demographic models are certainly preferable to no risk adjustment. However, the diagnosis-based models we tested provide considerably more accurate predictions of health expenditures.

Within the diagnosis-based methods, we observed smaller differences in predictive accuracy. Some models performed better on some measures, while others performed bettter on others. Overall, the ADG, EDCGDX, and PIPDCG models had some advantages over the other models we tested.

In terms of practicality, ability to restrict manipulation and timeliness, the more complex models we evaluated—those based on diagnoses—face more significant challenges before they can become fully operational for a risk adjustment process. In particular, those models using ambulatory diagnoses raise important questions of data quality, consistency, and completeness. Given its reliance on only inpatient diagnoses, the PIPDCG model we tested may have some advantages in this regard. Inpatient data are currently collected in a more uniform way and represent a smaller number of claims. This information may also be more difficult to game and is certainly simpler to audit. Given these practical advantages and its enhanced predictive ability over demographic models, the PIPDCG model may serve a useful role as a transition approach to risk assessment until sufficient data systems are available to support those models using a wider range of clinical information. However, as discussed above, models recognizing only inpatient information may provide incentives to treat patients on an inpatient rather than outpatient basis.

Important questions remain: Are the best methods we tested sufficiently accurate to provide equitable payments to plans for the risks they enroll? Do they eliminate incentives to select risk? We found all the models we tested to perform well for large random groups. If individuals distribute themselves in a random way across health plans, then these models are sufficient. However, if they do not, then the models also need to predict well for individuals and nonrandom groups.

Again, do these models predict well enough? The answer depends, fundamentally, on the difference between the information that the methods use to predict the expenditures of individual enrollees, and the information available to the plans. Our results shed some significant light on this issue. In order to show how, some explanation is necessary.

As noted in Chapter II, an adequate risk assessment method does not need to predict or explain all the

variation in expenditures across individuals in order to prevent risk selection. It only needs to do about as well, on average, as a plan can reasonably be expected to do. Several researchers have proposed a theoretical maximum for how well a plan can do in predicting risk, based on the ability of previous enrollee claims to explain future expenditures. They suggest that a plan, given available data, should be able to predict 15 to 20% of the variation in expenditures across individuals in a future year. This then may be viewed as the target for a risk assessment method. In fact, this target may be a lower bound, particularly if a plan can obtain data such as those collected for medical underwriting or if a plan has knowledge about the expected trend in expenditures for an individual such as a recent diagnosis of AIDS. On the other hand, the costs of risk selecting behavior for insurers may allow a somewhat lower predictive accuracy to suffice. However, for our purposes, we can assume 15 to 20% (or an individual R^2 of 0.15 to 0.20) is the standard by which to judge our results.

The best diagnosis-based prospective models we tested fell short of this standard. They were able to predict only one-half to two-thirds of the variation in expenditures that plans are likely to be able to predict, using the more complete data on enrollees at their disposal. Our results are therefore consistent with previous research in indicating that prospective risk adjustment using diagnosis-based methods is unlikely to deter risk selection by insurers. Plans would still be able to, and would benefit from, identifying higher risk individuals and either avoid enrolling them, or discourage them from remaining in their plan. Furthermore, the risk adjustment process would not compensate plans with many high-risk enrollees enough to level the playing field.

The results from our retrospective analyses at first suggest that retrospective risk assessment might be the answer: applied retrospectively, the models explain 40 to 50% of the variation in expenditures, better than the plans could ever do on a prospective basis.[1] In more concrete terms, the transfers for a given year would match their claims experience for that year, individual-by-individual, more closely than the plans' own predictions made on the basis of information available in a previous year. Would this then succeed in removing incentives for risk selection, and level the playing field?

Our retrospective analyses for nonrandom groups show that the answer is no. Even though they can explain up to half of the variation in expenditures, the retrospective models systematically overpredict for select groups of enrollees and underpredict for others. In fact, they do no better than the prospective models in predicting expenditures for these groups. Even with a retrospective model, plans would benefit from attracting enrollees with relatively low prior expenditures and avoiding those with relatively high previous expenditures or certain medical conditions.

The reason why the increase in predictive accuracy achieved through retrospective risk assessment still leaves room for selection behavior lies in the nature of the information that retrospective assessment uses. As indicated above, the information that prospective methods cannot use is of two types: that which the plan or an individual knows, but the risk assessment method does not; and that which neither the risk assessment method nor the plan or individual can know ahead of time, because it corresponds to unpredictable events. An example of the first type of information would be the plan's or an individual's knowledge of test results indicating whether or not a cancer has metastatized. Such test results might not be discernible from diagnosis codes. An example of the second type of information would be knowledge that an enrollee was going to experience a heart attack the following year, or a serious auto accident.

The retrospective models capture a good deal of the second type of information, since heart attacks and serious auto accidents will typically lead to medical encounters and specific diagnoses. They also, most likely, capture some of the first type of information. But the nonrandom group results show that the retrospective models cannot do this perfectly.

In sum, the predictive power of the retrospective models looks attractive. However, much of this additional predictive ability is due to the unexpected acute problems that are partially accounted for (heart failures, acute infections, accidental trauma) by these models. These acute events are difficult to predict by health plans and not easily used for risk selection. They also might be expected to average out over a large group of enrollees. Given this, and our findings for nonrandom groups, a prospective model may do just as well at reducing incentives for risk selection. If so, the greater incentives for efficiency with a prospective model would make it the preferred approach.

Thus, our findings indicate that opportunities for profitable risk selection and inequities in payments remain even with the best risk assessment models we tested. It is unlikely that plans can avoid all the higher risks and enroll only the best ones. There are limits to risk selection, even in the current environment, without the measures typically included in proposals for insurance

market reform. However, it is equally unlikely that individuals will distribute themselves randomly across plans. As long as selection occurs, the risk adjustment process is unable to fully compensate plans for their differences in risk. The general conclusion to which previous research has already pointed thus still holds: no current risk assessment method based on diagnoses can completely remove incentives for risk selecting behavior, whether applied prospectively or retrospectively.

An increasing number of states are turning to lists of high-cost conditions as a basis for risk adjustment and a substitute for reinsurance.[2] How much promise do such lists hold? By focusing on high-cost conditions, they target those individuals who pose the more significant problems for risk selection. However, our findings indicate the PIPDCG and EDCGDX models achieve greater predictive accuracy than even a fairly extensive list of high-cost conditions screened for incentive effects. And even these DCG models, as we have seen, are inadequate to remove all incentives for risk selection. Nevertheless, such an approach may warrant further investigation, particularly if subpopulations such as those in the nonrandom groups we tested can be targeted more effectively.

Additional research may yield substantial improvements in risk measures and modeling techniques. It appears unlikely, however, that enough improvement can be achieved in the foreseeable future that risk adjustment *by itself* will remove all incentives for risk selection. Two general strategies might then be pursued:

- Market reform strategies that constrain the ability of insurers to select risks could be combined with risk adjustment. Many of these strategies have been proposed previously and include: open enrollment, guaranteed issue and renewal, standardized benefits, and no direct contact between an insurer's sales representative and applicants during the enrollment process (Luft, 1995; van de Ven et al., 1994).
- More significant changes designed to *reduce the benefits* of risk selection could also be adopted. These include community-rated prospective high-risk pooling and blended payment mechanisms that incorporate both prospective and retrospective features.[3]

It is an open question whether market regulations would be sufficient, in combination with a diagnosis-based risk adjustment process, to deter risk-selecting behavior. Community-rated prospective high-risk pooling has been implemented in some states and would probably help. Additional research is needed to determine which, among these alternatives, alone or in combination, offers the best prospects.

Risk assessment and risk adjustment will play important roles in any health care reform strategy. Our results help to illuminate the relative strengths and weaknesses of different diagnosis-based risk assessment methods, including lists of high-cost conditions. Relative to no risk adjustment, these models clearly reduce incentives for risk selection and provide more equitable payments to plans for the risks they enroll. Our pessimistic assessment of the potential for risk assessment and risk adjustment, used alone, brings into focus the need for additional measures to prevent risk selection and ensure that health plans compete on a level playing field.

END NOTES

1. The retrospective models cannot explain better than half of the variation in expenditures because: (1) diagnoses leading to a lengthy treatment, such as cancer, can be made early as well as late in the year; (2) patients with a given diagnosis can differ in their health care needs; and (3) patients with identical conditions in many cases receive different medical treatments. In other words, expenditures vary within a risk group, even for retrospective models.
2. Our primary analysis of predictive accuracy where individual expenditures were truncated at $25,000 can in many ways be thought as a reinsurance scheme with risk-adjusted payments for expenditures under $25,000 with full reimbursement for expenditures above this amount. As shown by our findings, even with this approach there is still room for identifying unprofitable patients below this threshold.
3. Community-rated prospective high-risk pooling could involve ceding individuals with expected high costs to a pool where all participating plans share in the risk. Under this approach various schemes can be applied to promote efficient care management for those in the pool (van de Ven et al., 1994). Under one option for a blended payment system, part of the payment to the plan would be a prospective capitated amount; the other part would vary with current use, just as payment under fee-for-service. The fees could be set to provide incentives for efficiency (Newhouse, 1994).