Chapter 3

Risk Adjustment for Health Plan Payment

Randall P. Ellis[1], **Bruno Martins**[1] **and Sherri Rose**[2]
[1]*Department of Economics, Boston University, Boston, MA, United States,* [2]*Department of Health Care Policy, Harvard Medical School, Boston, MA, United States*

3.1 INTRODUCTION

This chapter reviews how risk adjustment can be developed and used for health plan payment, with an emphasis on practical aspects of risk adjustment model design, estimation, and implementation in healthcare insurance markets, using information at the individual level to allocate funds to competing health plans. Since our interest is in health plan payment rather than provider reimbursement, we concentrate on predictions of plan obligations for a 1-year period rather than on predicting other measures, such as the cost of hospitalizations, episodes, or spells of treatment, which are more commonly used for provider or provider network payment. We provide a brief review of the theoretical literature on risk adjustment before turning to the practical issues of specification, estimation, estimator selection, and payment implementation of risk adjustment models. We touch upon issues related to premiums, risk sharing, and market regulations in this chapter only to the extent that these issues create special considerations in the design and estimation of risk adjustment; the main discussion of these issues is elsewhere in this volume.

Risk-adjusted plan payment is only possible if there is an agent, here called the regulator,[1] which could be a government, independent agency. or employer, willing to reallocate payments to plans based on the predicted costs for each enrollee. Three other ways for a regulator to pay a health plan for their enrollees are to pay actual cost incurred by the plan (plus an administrative fee), to pay a fixed lump sum (equal say to the average cost), or to pay a competitively determined premium for each enrollee. Paying actual costs provides no incentive for plans to control costs, but does eliminate the incentive for plans to avoid unprofitable enrollees. Paying a fixed lump sum amount equal to the average cost does the opposite: maximizing cost-saving incentives, but creating strong selection incentives to avoid high-cost

Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets.
DOI: https://doi.org/10.1016/B978-0-12-811325-7.00003-8

enrollees. Premiums can be determined through competitive bidding, or by allowing health plans to charge a premium directly to enrollees based on enrollee characteristics. The disadvantages of premiums are that plans may not be perfectly competitive, and unacceptably large differences in the break-even premiums can arise. Moreover, premiums might become unaffordable for high-risk people. More than 10-fold differences in premiums can emerge based on age and gender alone, with much larger differences possible if health status or other information is used for premium setting. Elsewhere in this volume, we explore risk sharing, in which plan payments reflect combinations of actual costs, lump sum payments, and premiums. The motivation for risk adjustment is that it can correct for (some of the) predictable spending variation, while maintaining cost containment incentives.

There are many issues to consider when designing, estimating, and implementing a risk adjustment model for health plan payment. Box 3.1 organizes these issues into nine dimensions, which can be broken down into estimation and implementation issues. We organize the presentation in this chapter around these nine issues after first discussing the criteria guiding the design of risk adjustment models in the next section.[2]

BOX 3.1 Nine dimensions of risk adjustment

Risk adjustment model estimation

1. The sample on which the risk adjustment model is to be calibrated (e.g., the entire population, or specific subsets of the population).
2. The types of services for which spending is to be predicted (e.g., for the total benefit package, specific services, or specific cost elements of certain services).
3. The types of information to be used for predicting annual spending (sociodemographic, diagnostic, pharmacy, or other information).
4. The timing of the information to be used for predicting annual spending (e.g., lagged or concurrent information, or both).
5. The objective function, functional form, and statistical methodology used for selection and estimation.

Risk adjustment model implementation

6. The group of members for which risk is to be equalized (e.g., entire population, each state, or certain plan types).
7. The adjustments made for the time lag between estimation and implementation of the formula.
8. The sources of funds paid into the equalization fund to which the risk adjustment formula is applied (premiums or taxes paid by consumers, funds from the regulator, or revenues from health plans).
9. The integration of the risk adjustment with risk sharing and premiums for plan payments.

In order to illustrate key features of empirical risk adjustment models, we intermingle our discussion of concepts with empirical examples, using results from existing studies as well as new results from commercial claims data. For our new empirical results, we use a sample of US privately insured enrollees from the widely used IBM Watson/Truven MarketScan Commercial Claims and Encounter data (the "MarketScan data"). MarketScan data were used to develop and evaluate the risk adjustment formula used in the Health Insurance Marketplace for populations aged 0−64, created as part of the Affordable Care Act (ACA) of 2010 (Kautter et al., 2014). For illustrating issues related to the practical application of risk adjustment models we use an enhanced version of the hierarchical condition category (HCC) model first described in Ash et al. (2000), commonly called the DxCG-HCC model.[3]

3.2 CRITERIA GUIDING THE DESIGN OF RISK ADJUSTMENT MODELS

We discuss here criteria guiding the design of risk adjustment models, as developed and reviewed in Van de Ven and Ellis (2000), Ash et al. (2000), Kautter et al. (2014), and Van Veen et al. (2015b). We group our discussion into three categories: incentives for efficiency, fairness, and feasibility. We also discuss and expand upon the principles for model development first presented in Pope et al. (2000) and used in the United States and elsewhere.

3.2.1 Efficiency

When developing risk adjustment models, a central objective is maintaining appropriate incentives for efficient provision of care. Efficiency raises concerns about the quality of information used to set payments, and concerns about creating incentives to provide the wrong quantities or qualities of healthcare services.

3.2.1.1 Avoiding Endogenous Signals

A central concern when selecting risk adjusters is that they should not be gameable, which is to say that plans or providers cannot readily manipulate them to increase plan payments. Ideal risk adjusters are exogenous to health plan influence and readily verifiable. Age and sex are ideal risk adjusters, although unfortunately by themselves they are not highly predictive of plan obligations. Variables such as counts of visits or dollars of healthcare spending for an enrollee are much more predictive, but also more endogenous variables. Diagnoses and pharmaceutical use are also endogenous, although researchers are still documenting the extent. Endogenous variables such as prescriptions, visits, and spending can directly cause welfare losses due to treatment or quality changes; the social and other costs of changes in diagnoses made to increase payments are less clear.

Papers that document or quantify the degree of endogeneity in the United States include MEDPAC (1998), Newhouse et al. (1999), Wennberg et al. (2013), and Geruso and Layton (2015). While there is no disagreement that manipulation of risk adjustment signals does occur, there are differences of opinions about the magnitude and seriousness of the problem. Bauhoff et al. (2017) estimate that in Germany the share of diagnoses recognized for payment grew by 3%–4% over a 5-year period, which is a rate of about 0.7% per year, an amount that could be removed through the payment formula, or accommodated as an estimate of technological change. Chapter 11, Health Plan Payment in Germany and Chapter 14, Health Plan Payment in the Netherlands, discuss the presence of endogenous signals in Germany and the Netherlands, where it appears to be a growing concern.

3.2.1.2 Avoiding Noisy Signals

In addition to endogeneity, efficiency (and fairness) concerns arise if risk adjusters are noisy. Variables such as homelessness, income, race/ethnicity, and indicators of need for long-term care services are examples of risk adjusters that can be predictive, but difficult to verify. Unfortunately, few variables that predict healthcare costs are fully exogenous and readily verifiable. Diagnoses from health claims are both noisy and potentially influenced by plan effort to change coding or utilization. Fig. 3.1 documents that among the commercially insured in the United States; there is a remarkable amount of year-to-year variation in the prevalence of specific chronic conditions. Evidence on the lack of persistence of diagnoses is also evident in Abbas et al. (2012) for Germany, which now requires outpatient diagnoses to appear

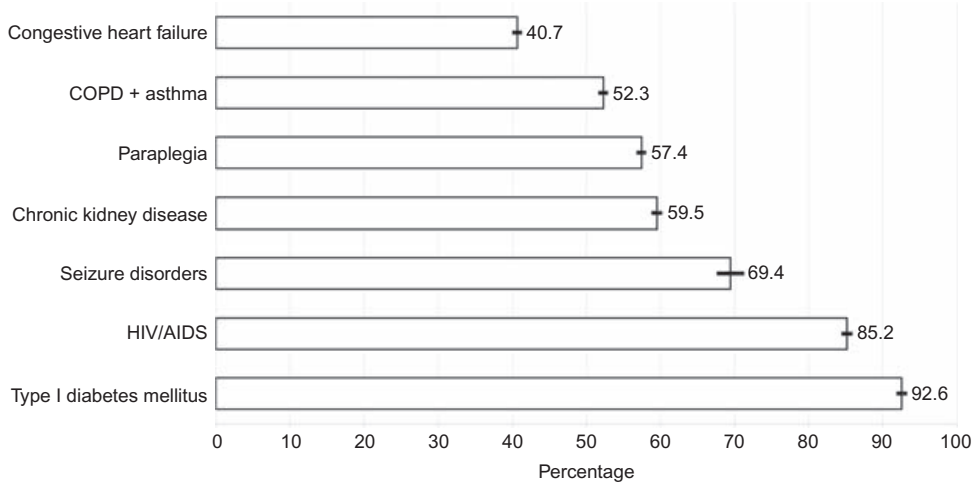

FIGURE 3.1 HCC Persistence between 2013 and 2014.
Note: Each bar shows, for individuals that had a given HCC in 2013, the percentage that had the same serious related condition in 2014. Sample corresponds to MarketScan individuals who were enrolled for 12 months in both 2013 and 2014, $N = 15{,}711{,}896$. Dark bars around the values correspond to 95% confidence intervals.

in two different quarters to affect plan payments. This evidence suggests both noisy coding and the potential for upcoding through greater plan effort.

3.2.1.3 Avoiding Incentives Not to Prevent or Cure

A related but slightly different issue for designing risk adjustment models is to avoid incentives for health plans to underspend on prevention or cure treatable conditions because this reduces future plan revenue (Pope et al., 2004). Eggleston et al. (2012) develop a two-period theoretical model of this problem, and show that achieving the first best requires a pay-for-performance type of incentive payment for prevention, and that this is difficult to implement when people can switch plans. The empirical magnitude of this problem has not been established.

3.2.1.4 Maintaining Incentives for Cost Control ("Power")

The efficiency issue that has received the greatest attention by developers of risk adjusters is to maintain incentives to control costs, which Laffont and Tirole (1993) define as the "power" of the contract to control costs. Newhouse (1996) characterizes health plan power conceptually, while more recently Geruso and McGuire (2016) develop empirical measures of power of risk-adjusted payments. Geruso and McGuire observe that because indicators for clinical conditions come from instances of healthcare utilization, a risk-adjusted payment system links costs to revenues, diluting plans' incentives to control costs. In their framework, full cost-based payments will have a power of zero, since any reductions in costs reduces revenues equally (and therefore there are no incentives to control costs). With exogenous risk adjusters like age and sex, the power of the payment system is one, which is to say that plans will face the full marginal cost of paying for each service provided. Geruso and McGuire calculate that concurrent risk adjustment has a power of 0.62 for inpatient events and 0.77 for outpatient events, versus 0.91 and 0.85, respectively, for a prospective model using the same risk adjustors. Power is one element of evaluation of health plan payment systems discussed in Chapter 5, Evaluating the Performance of Health Plan Payment Systems.

3.2.1.5 Avoiding Overpayment

Although the power of a payment system is a useful measure in terms of the *marginal* revenue generated by an incremental dollar of spending, the overall *average* revenue can also have direct effects on cost containment incentives. Even with fully capitated payments, under competition, overly generous payments can motivate providers to overprovide services, even when the calculated power is one (Ellis et al., 2016). This can be the result of either the overall payment rate being too generous or the capitated payment for a population subgroup being too high. Since consideration of payment generosity

affects cost-saving incentives in every health plan payment system, generosity is not solely a risk adjustment issue; therefore we circumvent the effects of overpayment here, and focus on risk adjustment payment schemes in which total plan payments exactly match total plan costs.

3.2.1.6 Avoiding Service-Level Selection Incentives

The central issue for risk adjustment is to avoid service-level selection incentives. Glazer and McGuire (2000) were the first to formally model how risk adjustment formulas should be modified to reduce service-level selection. Layton et al. (2017), along with many chapters in this volume, discuss how undesirable selection incentives can be reduced or quantified in the design and implementation of risk adjustment models, as well as through regulation, premium design, and risk sharing. Mitigating service-level selection incentives is perhaps the most important efficiency rationale for risk adjustment, but it is not the only efficiency concern.

3.2.2 Fairness

Although economists often focus solely on efficiency issues, a majority of health planners and consumers also care about fairness. For example, regulators might want to achieve a certain concept of equity in individuals' contributions to the health insurance system. Such objectives have implications for the design of risk adjustment. As discussed below, fairness can matter across multiple dimensions, and fairness across age and health status can conflict with fairness of payments across income, geography, or other socioeconomic variables like education and race. In the United States, fairness considerations commonly guide the choice of risk adjusters to use in payment formulas. For example, Ash et al. (2000) and Pope et al. (2004) discuss why certain variables like race and income are not appropriate risk adjusters, even if predictive, and why payments should not be lower for certain conditions such as dementia and severe developmental disability, which can lead to undertreatment. Ash et al. (2017) describe how the Massachusetts Medicaid program started using homelessness and neighborhood variables for risk adjustment in 2016 to improve the fairness of the state's Medicaid risk adjustment formula. In Europe, fairness is commonly spoken of in terms of "solidarity" across income or health when discussing risk adjustment (Chinitz et al., 1998a,b; van de Ven and Ellis, 2000; Van Kleef et al., 2009). Solidarity and fairness issues are discussed below when discussing sociodemographic variables, and in Chapter 7, Risk Adjustment in Belgium: Why and How to Introduce Socioeconomic Variables in Health Plan Payment (Belgium) and Chapter 14, Health Plan Payment in the Netherlands.

3.2.3 Feasibility

Some risk adjusters may be desirable but infeasible to implement. For example, using diagnoses from all sources may be infeasible in a health plan setting if such diagnoses are not already collected and available for use in calibrating a risk adjustment model. Data availability can be changed by regulations, and provision of data does respond to financial incentives. In the early 2000s, Germany's office-based physicians greatly improved their coding of diagnoses on office-based claims once the government announced that office-based diagnoses would be used along with inpatient diagnoses for risk adjustment. In a similar way, the prospect of using office-based diagnoses for risk adjustment in the United States for Medicare Advantage risk adjustment led to a remarkable improvement in diagnostic coding practice in the early 2000s when the change was phased in. Regulators and risk adjustment model developers should not think of imperfect data as an irremediable flaw.

Feasibility issues can also arise for other reasons. In every country, it is infeasible to obtain prior year data for new immigrants. Frequent health plan changes and the absence of unique identifiers that permit linking individuals across health plans make it infeasible in most of the United States to calibrate risk adjustment models in the private sector that span different insurers. Switching from private to Medicaid or Medicare health insurance creates similar data issues. Feasible risk adjustment in the United States must always accommodate new, partial-year enrollees for other reasons than birth and migration, at rates that vastly exceed rates of partial-year enrollment in most other countries.

Policymakers often feel that a simpler system is more feasible to implement. This view is reflected in the early efforts in the US Medicare and German systems to develop and implement risk adjustment models with only a modest number of disease categories, and simple data burdens. Early risk adjustment models have often used a "rate cell" approach (preferred by many US actuaries and currently used as a large part of the payment system in Switzerland (Chapter 16: Health Plan Payment in Switzerland) and elsewhere) in which each person is assigned to one unique rate cell, and the mean cost of people in that cell serves as a basis for the payment for that category. Such models are easy to explain, and have some implementation advantages. For example, in a rate cell system, there are no interactions between cells, and predictions for one cell can be adjusted without affecting the predictions for other cells. Rate cell payments can also be generated using aggregated rather than individual-level data, which can be a big plus. The disadvantages of rate cells are that sample size limits the number of cells for which means can be reliably estimated, and they generally have less predictive power than additive models with more variables.

More recently, and perhaps in response to growing challenges of upcoding, and worsening service-level selection, risk adjustment models in the

Netherlands, Germany, and the United States have become more complex, with separate models for different population groups and/or medical services, and increased numbers and variety of risk adjusters. In an interesting twist on the argument for simplicity, Rose (2016) has argued that complex empirical methods for estimation, such as machine learning algorithms (discussed below) confer an advantage rather than a disadvantage. If providers and plans cannot reverse engineer the payment model, they may not be in a good position to manipulate it by upcoding or other tactics.

3.2.4 Ten Principles in Pope et al. (2004)

Box 3.2 summarizes the 10 principles that guided the creation of the diagnostic classification system of the first HCC system for Medicare Advantage, and that have remained influential in the development of the Medicare Part D prescription drug (Kautter et al., 2012) and Marketplace. (Kautter et al., 2014) risk adjustment formulas. Similar principles also guided the initial development of the German diagnosis-based classification system. The advantage of specifying principles is that, once agreed upon, they can be

BOX 3.2 Principles guiding HCC model development

1. Diagnostic categories should be clinically meaningful.
2. Diagnostic categories should be predictive.
3. Diagnostic categories that will affect payments should have adequate sample sizes to permit accurate and stable estimates of expenditures.
4. Hierarchies should be used to characterize the person's illness level within each disease process, while the effects of unrelated disease processes accumulate.
5. The diagnostic classification should encourage specific coding.
6. The diagnostic classification should not reward coding proliferation.
7. Providers should not be penalized for recording additional diagnoses (monotonicity).
8. The classification system should be internally consistent (transitive) with regard to costs.
9. The diagnostic classification should assign all ICD-9-CM codes (i.e., be exhaustive).
10. Discretionary diagnostic categories should be excluded from payment models.
11. *Designers should anticipate induced changes in coding and treatment.*
12. *Designers should optimize given likely selection effects induced by payment system.*

Note: The first 10 principles are from Pope et al. (2004).

applied by researchers repeatedly without having to return to clinicians, statisticians, and policymakers as frequently for guidance.[4]

Principle 1 seems obvious but may be violated by machine learning or other algorithms that group diseases with similar costs but diverse clinical meaning. Principle 2 warns against creating categories that are clinically meaningful but not predictive. Principle 3 guides how finely to create clusters of conditions as a priori protection against overfitting ($N \geq 500$ is a common minimum cell size). Principles 4, 5, and 6 speak to designing risk adjusters to reduce sensitivity to gaming. Principles 7 and 8 reflect desirable properties for fairness and consistency. Principle 9 is primarily for bookkeeping, making it easier to identify new or unclassified diagnoses. Principle 10 recognizes that payment models can differ from predictive models (also a prominent theme with machine learning models) and can justify substantial reductions in predictive power in order to improve incentives. The final two principles, shown in italics, were not in the original Pope et al. (2004) list. We added them to reflect recent insights into risk adjustment discussed below: designers should anticipate the effects of the payment system on the risk adjusters, and try to optimize the formula against anticipated selection effects.

We now turn to a discussion of the nine dimensions of risk adjustment described in Box 3.1.

3.3 CHOICE OF ESTIMATION SAMPLE

The first decision to make in risk adjustment model development is what sample to use for model calibration. Although it would seem obvious to use a large sample from the same population as the one on which the risk adjustment model will be applied, this is often not done. One reason is feasibility, related to data availability. The US Medicare Advantage program (Chapter 19: Medicare Advantage: Regulated Competition in the Shadow of a Public Option) continues to use the traditional Medicare enrollee sample, not its own enrollee data, for calibrating its risk adjustment formula more than 30 years after first adopting risk adjustment, since the Medicare Advantage data needed for this purpose are not collected. The US Marketplace (Chapter 17: Health Plan Payment in US Marketplaces: Regulated Competition With a Weak Mandate) uses privately insured claims data from large employers for its formula for the new individual insurance market. Germany used data from only a subset of all plans to initially develop its first risk adjustment formula, although Germany now uses a national sample.

Beyond feasibility explanations, Newhouse (2017) argues that if the population on which the risk adjustment formula is to be applied for payment reflects service-level distortions, then using a sample unbiased by selection effects may be desirable. This rationale underlies the calibration of formulas

on traditional Medicare used for the Medicare Advantage enrollees. This argument is further extended in Bergquist et al. (2018), who point out that it may be not only service-level distortions, but also under- or overconsumption by various population subgroups in the estimation sample that may cause problems during estimation.

A number of empirical studies have shown that for predicting total spending, risk adjustment formulas developed on one sample are often relatively robust for prediction on different samples. Ash et al. (2000) examined correlations of risk scores generated between privately insured, Medicare, and Medicaid enrollees, while Ash and Ellis (2012) demonstrate the stability of a US formula over 6 years and seven plan types. Ellis et al. (2013a,b) found that an HCC formula calibrated using US data had predictive power nearly as strong as using 117 related condition categories, which are aggregates of HCCs, calibrated using Australian data. Rose et al. (2015) show that fit results for the US Marketplaces are similar when using the privately insured claims data versus a sample of that data selected to more accurately reflect Marketplace enrollees.

3.3.1 Sample Exclusions

It is common for risk adjustment models to be estimated on data after elimination of troublesome records. This often includes purging partial year (less than 12 month) eligibles, or, in prospective models, dropping people when the full 12 months of prior-year claims are not available. Also common is to drop extreme outliers, or alternatively to "top-code" outliers, i.e., to replace spending on individuals above a threshold (such as $250,000) with that threshold.[5] For evaluating different risk adjustment models, it is also common to focus on relatively homogeneous subgroups, such as adults, by excluding infants and children. Table 3.1 uses 2014 MarketScan data to illustrate how these exclusions affect sample means, and three measures of variability, all of which are unit-free measures and hence comparable across samples.[6] These variability measures are the coefficient of variation (CV, which is the standard deviation divided by the mean), skewness (which captures how asymmetric spending is around the mean), and kurtosis (which captures how thick the tails are). Excluding partial-year eligibles has a particularly large effect on these latter two measures, and will particularly bias risk adjustment formulas since it drops most deaths and newborns from the sample, both of which have unique characteristics and may have (very) high spending.[7] For diseases like chronic heart failure and pancreatic cancer, only including people who survive for an additional 12 calendar months in the estimation sample generates a very biased subset of these populations. Methods for incorporating and adjusting for partial-year eligibles are discussed in Section 3.3.6.

TABLE 3.1 Alternative Estimation Sample Summary Statistics on 2014 Plan Payments per Enrollee

	Number of observations	Mean spending	CV	Skewness	Kurtosis
Full sample	21,832,612	4429	1660	184.9	219,009
Removed if less than 12 months eligible in 2014	18,041,199	4322	1521	36.4	5061
As above, plus removed if less than 12 months eligible in 2013	15,710,699	4416	1507	35.8	5135
As above, plus removed if aged 0–21	10,894,520	5473	1322	29.1	4071
As above, plus removed if spending more than 1000 times mean	10,894,517	5471	1305	21.3	1177

Sample is the IBM Watson/Truven MarketScan Commercial Claims and Encounter. Variable used is plan obligations per enrollee divided by the fraction of the year eligible. All statistics generated use sample weights equal to the fraction of months enrollee was eligible in 2014. Observation counts are unweighted counts of enrollees.

3.3.2 Separate Formulas for Population Subgroups

It is relatively common to estimate separate regression models for distinct subpopulations in recognition of different patterns of disease and cost. The 2017 CMS Medicare Advantage model uses nine different formulas for different subpopulations (Chapter 19: Medicare Advantage: Regulated Competition in the Shadow of a Public Option). These formulas differ according to whether the enrollee is aged (age 65 and over) or disabled (age < 65), ineligible, fully, or partially eligible for Medicaid. In addition, three more formulas are used for institutionalized enrollees (i.e., those in a nursing home), for new enrollees with less than 9 months of prior year eligibility, and for a subset of new enrollees in chronic condition special needs plans. The US Medicare Part D risk adjustment formula uses the first eight but not the final model. The Swiss (Chapter 16: Health Plan Payment in Switzerland) have separate risk adjustment formulas within each canton (similar to a county in the United States), using age, gender, and whether people are hospitalized or not. Since they use primarily a rate cell approach

rather than a regression-based approach for risk adjustment, it is equivalent to having separate models for each geographic area canton.

Estimating separate models for population subgroups is generally a good idea if sample sizes are adequate, and there is evidence that cost patterns differ among the groups. Germany, despite having an enormous sample size, uses a single risk equalization formula for the full population, although the formula does include age-specific HCC terms that allow it to better predict certain age-related spending patterns (Chapter 11: Health Plan Payment in Germany). Estimating a single formula, but including dummy variables for population subgroups—alone or interacted with other risk adjusters—is more appropriate where sample sizes are a concern. From a modeling perspective, there is a tradeoff between obtaining greater fit by having separate models with fewer risk adjusters versus gaining from information learned across subgroups by having more complex single equation models with interactions. The Netherlands (Chapter 14: Health Plan Payment in the Netherlands) uses the latter approach extensively. Machine learning approaches, discussed below, provide an empirical basis for choosing model structure based on statistical grounds.

3.3.3 Separate Formulas for Different Health Plan Benefits

In some countries there is not one formula used for risk adjustment for a given person, but rather a family of formulas that depend on the plan the person chooses. The US Marketplace risk adjustment formula has five variants that vary according to whether the enrollee is in a platinum, gold, silver, bronze, or catastrophic plan. The Marketplace formulas were developed on the basis of one sample of enrollees, on which the effects of different degrees of benefit coverage were simulated. More concretely, Kautter et al. (2014) started with total covered spending in an estimation sample, without correcting for the existing level of plan coverage. They simulated the effects of the platinum, gold, silver, bronze, and catastrophic plan benefit levels on out-of-pocket costs, subtracted these costs from covered spending and used the resulting simulated plan obligations to estimate separate risk adjustment formulas. Empirically the risk scores from formulas estimated by Kautter et al. for different benefit plans are highly correlated, but are scaled to reflect the differences in coverage.

Adjusting payments for differences in benefit design clearly helps with predicting means correctly, but it introduces issues of fairness: how large should the subsidies be (through risk equalization) for consumers choosing more generous benefit when this generosity induces greater healthcare utilization? A significant concern, about which there is relatively little research, is how to incorporate consumer and provider behavioral response to benefit design differences across plans into the risk adjustment formula.

3.3.4 Separate Formulas for Different Types of Services

The correct dependent variable in risk adjustment modeling is plan-obligated spending, which implies calculating both the services covered and plan obligations after deducting enrollee cost-sharing payments and any payments a plan would receive from risk sharing (such as reinsurance). In the United States, Medicare Advantage plans are only required to cover specified inpatient and outpatient spending, notably not including prescription drugs (although many plans nonetheless choose to include pharmacy coverage) hence the Medicare Advantage formulas predict plan obligations only for inpatient and outpatient services covered by traditional Medicare. The Medicare program uses a separate risk adjustment formula for its prescription drug plans that cover only prescription drugs (Chapter 19: Medicare Advantage: Regulated Competition in the Shadow of a Public Option). The Netherlands has separate formulas for subsets of spending rather than subsets of the population. Their main model covers somatic health care (medical plus pharmaceutical spending, excluding certain specified categories) that encompasses about 80% of total healthcare spending under the benefits package. Separate models predict and equalize payments for short-term mental health care, long-term mental health care, and further calculations correct payments for differences in out-of-pocket payments for deductibles (Chapter 14: Health Plan Payment in the Netherlands).

Estimating separate formulas for distinct services does not create implementation problems if the formulas are combined when making plan payments, as they are in the Netherlands. But separate formulas for different services can create problems when there is a separate contract or risk adjustment equalization for these different services (which are called "carve outs" in the United States). Separate contracts may encourage inappropriate substitution between different services. For example, in the US Medicare program, risk equalization and payments for outpatient prescription drugs in the Part D program are done separately from the Medicare Advantage risk adjustment, in which some of the plans also include prescription drugs. When payments for different services come out of different bundled payments, providers may have an incentive to change care patterns and take advantage of these different payment flows. Carve outs also add budgetary complexity and encourage lobbying for favorable funding.

3.3.5 Predicting Only Covered Services

Countries vary in how fully they specify the services that must be covered by the health plans. In some systems coverage of all qualified providers and drugs is determined nationally, whereas in others considerable discretion is exercised at the plan level. An example from the United States is pharmaceutical spending where formularies, subject to some regulation, may include or

exclude a wide number of drugs. In principle, developers of risk adjustment models would also know what costs are to be included when estimating formulas. Coverage is standardized for traditional Medicare in the United States, while there is meaningful heterogeneity in what services are covered or not covered in Medicare Advantage, prescription drug plans, and the Marketplace.

Payment formulas can be adjusted when new technologies or costs are anticipated. For example, in 2016, a new hepatitis C drug in the US marketed by Gilead Sciences had a list price of $75,000 for a 12-week drug treatment, and was recommended for virtually everyone infected with hepatitis C. This had a noticeable one-time cost increase for this illness. The Medicare Part D prescription drug program (CMS, 2016a) as well as the Massachusetts Medicaid risk adjustment program (Clements et al., 2016) built these additional drug costs into their risk adjustment payment formulas in a relatively ad hoc manner without relying on regression recalibration.

In some contexts, data show that total paid and covered amounts are extremely highly correlated ($\rho = 0.998$ in our US MarketScan data, whether top-coded at $250,000 or not). In these cases, the differences in risk scores at the aggregate for a given sample are relatively small according to whether paid or total spending are used for estimating risk adjustment models. Using payments rather than total spending will matter for certain diseases or types of spending where drugs or outpatient services have higher or lower rates of coverage, and this coverage varies across health plans. In settings in which demand-side cost sharing is modest and there is little risk sharing by the regulator, the differences in relative risk scores (RRS) using total and plan-paid amounts is likely to be modest at the plan level, but differences of even a few percent may be troubling. We have not seen this issue explored empirically in settings other than the United States.

3.3.6 Accommodating Partial-Year Eligibles

For research studies, researchers often choose to focus on the cleanest sample, which usually means samples in which everyone is enrolled for all 12 months in a calendar year. For payment purposes, one still needs to make predictions for people with less than 12 months of eligibility. Using estimates based on only full-year eligibles is undesirable because partial-year enrollments are nonrandom, and have different patterns of costs, as we already illustrated in Table 3.1. Births, deaths, retirements, and changing jobs or health plans are all correlated with specific diseases and levels of health spending, and hence if partial-year enrollees are dropped, or this issue is ignored, then serious biases can result.

Ellis and Ash (1995) advocated, and many regulators adopted, a method for estimating linear risk adjustment models with annualized spending and then weighting the sample by the fraction of the year a person is eligible.

This is equivalent to using the average monthly spending on health care and then weighting by the number of months eligible. It is straightforward to show that this results in unbiased predictions of monthly spending which exactly match actual spending in every mutually exclusive cell created by the dichotomous risk adjusters (like HCCs), i.e., the formula correctly predicts actual spending for people in each HCC.

The importance of annualizing is easily seen by considering newborns. Newborns are relatively expensive on average compared to 1−10-year-olds. Suppose that on average in their first year, newborns cost $6000. Unlike most 1−10-year olds, babies are on average only eligible for coverage for about half of the year. Therefore their average monthly cost should be $1000 per month eligible. Without annualizing and weighting, a risk adjustment model will predict that babies cost only $500 per month, half of the actual value. This problem is fixed by annualizing and weighting the spending. Annualizing and weighting is particularly important in the United States where people change health plans frequently, and hence partial-year coverage is relatively common. It is also particularly important when enormous resources are spent on people in the year in which they die, which is true in the United States as well as other countries.

Using unweighted spending can be preferred when health plan eligibility data are missing or of poor quality or when supplementary plan coverage is only used rarely even when continuously available. One example is US Department of Veterans Affairs health claims data, since US veterans remain eligible for veterans' benefits continuously once eligible. Even if a veteran does not use any VA services, they are still eligible. This is true in other settings, such as with private insurance in Australia, where a supplementary benefit means that enrollees often obtain insurance from other sources.[8] With very intermittent use of the benefit, perhaps only every few years, assigning individuals to a geographic region or provider group can be problematic.

Partial-year eligibles create two problems for risk adjustment. One problem is that annual spending in the prediction year for which payments are made will be biased downward, which is addressed by predicting annualized spending, as described. A different problem arises because the base period during which diagnoses (or other risk adjusters) are observed is shortened. Chen et al. (2015) examine the bias and weaker fit from ignoring the duration of the base period and propose formulas that incorporate duration information in the prospective Medicare Advantage formula. Ericson et al. (2017) document the undercount of diagnoses in concurrent models such as the Marketplace formula and propose adjustments to improve fit and lessen bias for partial-year eligibles.

Adjustment for partial year enrollment is done differently in various countries. The United States and Switzerland use monthly eligibility to annualize spending and perform risk equalization. Germany and the Netherlands

use the number of days in the year covered for both annualizing and weighting. The choice between using monthly or daily information for annualizing and weighting could be influenced by at least two issues. In smaller sample sizes, weighting by days can introduce some very large outliers for people only eligible for a few days, and hence is less desirable than a monthly annualization.[9] The second issue is how premiums and plan revenue payments are paid. In the United States, most employers and the government pay health plans a monthly premium for each enrollee, even when an enrollee is only eligible for a fraction of the month, while Germany and the Netherlands adjust payments to health plans based on the number of days each individual is enrolled.

3.3.7 Normalizations to Create Relative Risk Scores

In the United States, risk adjustment model results are generally presented in terms of RRS rather than monetary predictions. RRS express predicted spending as a multiple of mean spending. Fig. 3.2 presents normalized spending rather than dollar amounts, which are akin to RRS. RRS are presented in most tables and figures in various government publications and software (e.g., Kautter et al., 2012, 2014). RRS always reflect a

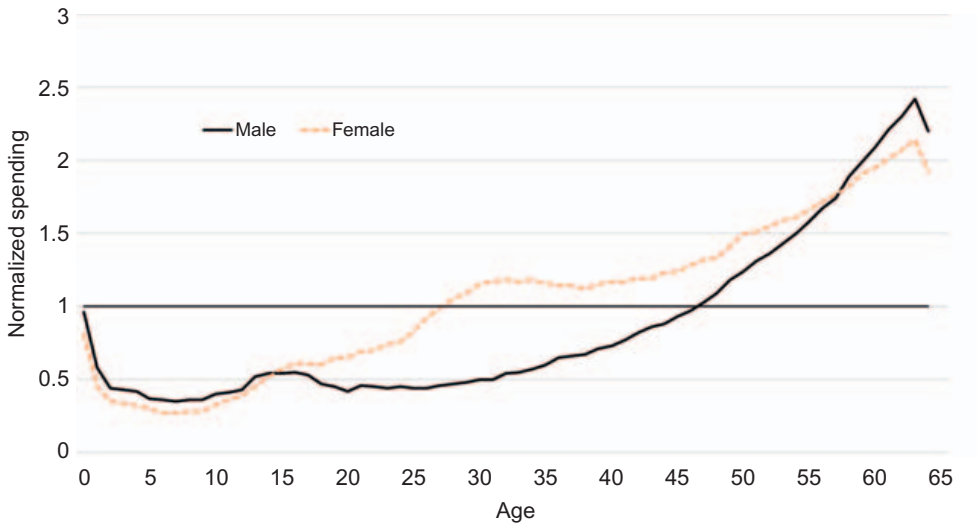

FIGURE 3.2 US normalized spending by age and sex (normalized by sample mean; $N = 21{,}832{,}612$).

Note: This figure shows normalized spending by 1-year age increments, for males and females, aged 0–64, in the 2014 US MarketScan Commercial Claims and Encounter Data using only people with no capitation payments. Normalized spending was calculated by first annualizing spending by dividing actual spending by the fraction of the year enrolled, and then calculated the weighted mean using eligibility fractions. Annualized spending was then divided by the weighted annualized average to create a normalized spending measure.

normalization to some period of time and sample, which should be specified for results to be interpreted easily.

Normalizations are particularly important to use when pooling data for estimation across different years, or multiple population subsets, where medical inflation and/or treatment intensity tends to change costs over time. To increase sample size, multiple years of claims data are often combined using medical cost deflators, such as in the United States the personal consumption expenditure medical cost index. For large samples, an alternative strategy is to normalize spending in each year by the average spending in that year before pooling.

3.4 INFORMATION USED FOR PREDICTING SPENDING (RISK ADJUSTERS)

This section discusses the types of information potentially used for risk adjustment, commonly called risk adjusters.

3.4.1 Age and Gender

The classic risk adjusters are age and gender. Fig. 3.2 illustrates the 1-year average spending per enrollee on all types of health care—inpatient, outpatient, and pharmaceutical—for a sample of 21.8 million individuals from age 0 to age 64 among the commercially insured population in the United States in 2014 by 1-year age intervals for males and females, where spending is normalized by the overall mean. Males and females show similar patterns until age 15, at which point spending starts to diverge and women have higher mean spending until around the age of 58.[10]

Fig. 3.2 reveals that the relationship between age and spending is nonlinear, and the difference between males and females is particularly noticeable during childbearing years. A similar although dampened pattern typically holds even when other risk adjusters are included. Thus, there is a strong argument for not using a simple additive sex term, but at least to use age−sex interaction terms. The HCC-CMS and HCC-HHS systems use 32 age−sex categories, with 5- or 0-year increments, approximating the curve for each sex with a step function. Even this step function approach introduces imperfect fits just before and after the break points that could be avoided by using finer age categories, including 1-year increments. As long as the overall sample size is large, then the age gender patterns can be reliably estimated with little risk of overfitting.

3.4.2 Diagnoses on Submitted Claims or Encounter Records

Both in the United States and elsewhere, diagnoses on claims or encounter records.[11] submitted by providers are the preferred set of information for risk

TABLE 3.2 Risk-Adjustment Model Results—R^2 (in Percentages) With 95% Confidence Intervals

Model	Untop-coded spending	Spending top-coded at $250,000	Log(1 + spending)
Concurrent			
Age–sex (Marketplace age groups)	1.4 (1.2, 1.6)	2.9 (2.9, 3.0)	10.68 (10.6, 10.7)
Age–sex (1-year increment)	1.5 (1.2, 1.7)	3 (3.0, 3.0)	11 (11.0, 11.1)
DxCG-HCC with age–sex (Marketplace age groups)	41.5 (35.2, 46.5)	57.9 (57.8, 58.0)	59.1 (59.1, 59.2)
Prospective			
DxCG-HCC with age–sex (Marketplace age groups)	15.3 (12.9, 17.3)	23.2 (23.1, 23.3)	29.7 (29.7, 29.7)

Each cell provides the within-sample R^2 (in percentage) for an OLS regression model that predicts total (outpatient, inpatient, and pharmacy) spending normalized by sample mean. N = 21,832,612. Age–sex variables are interactions of sex and age dummies variables using either the Marketplace age groups or 1-year age increments. DxCG-HCC refers to DxCG 394 Hierarchical Conditional Categories. 95% confidence intervals (in parentheses) are based on 500 bootstraps.

adjustment, currently in use in the United States, Germany, the Netherlands, Belgium, and Israel. Diagnoses have the advantage of being potentially verifiable in most cases by reviewing the patient's medical records. Furthermore, diagnoses are much more predictive than simply age and sex. As shown in Table 3.2, the R-squared for diagnoses-based prospective and concurrent models are 15.3% and 41.5%, versus only 1.5% for age–sex alone in US commercial data. Improvements in the root mean squared error (RMSE) and mean absolute error (MAE), two other commonly used metrics of fit (Table 3.3), are also impressive. This improvement in predictive power is even greater once the data are top-coded at $250,000, where we also see that the confidence bands are reduced to close to a zero range.

The quality of diagnoses recorded varies across providers and settings, with inpatient diagnoses generally viewed as more accurate than office-based diagnoses. Sometimes nonclinicians (e.g., home health workers or massage therapists) may report diagnoses on claims or encounters, which in the United States and in most countries are not recognized in risk adjustment models (Kautter et al., 2014; Department of Health and Human Services, 2016). We will have more to say later about how the very large number of the International Classification of Diseases (ICD) diagnoses (approximately

TABLE 3.3 Two Alternative Risk-Adjustment Model Measures of Fit

	Root mean squared error		Mean absolute errors	
Model	**Untop-coded spending**	**Spending top-coded at $250,000**	**Untop-coded spending**	**Spending top-coded at $250,000**
No risk adjustment (constant only)	14.8 (14.0, 16.1)	9.9 (9.9, 10.0)	1.26 (1.25, 1.26)	1.19 (1.18, 1.9)
Concurrent				
Age−sex (Marketplace age groups)	14.7 (13.8, 16.1)	9.8 (9.8, 9.8)	1.2 (1.19, 1.20)	1.13 (1.13, 1.13)
Age−sex (1-year increment)	14.7 (13.8, 16.1)	9.8 (9.8, 9.8)	1.2 (1.19, 1.20)	1.13 (1.13, 1.13)
DxCG-HCC with age−sex (Marketplace age groups)	11.3 (10.2, 13.0)	6.5 (6.4, 6.5)	0.71 (0.71, 0.72)	0.65 (0.65, 0.65)
Prospective				
DxCG-HCC with age−sex (Marketplace age groups)	13.6 (12.7, 15.1)	8.7 (8.7, 8.7)	1 (0.99, 1.00)	0.93 (0.93, 0.94)

Each cell provides the root mean squared error or mean absolute error for an OLS regression model that predicts total (outpatient, inpatient, and pharmacy) spending normalized by sample mean. $N = 21{,}832{,}612$. Age−sex variables are interactions of sex and age dummies variables using either the Marketplace age groups or 1-year age increment dummies. DxCG-HCC refers to DxCG 394 Hierarchical Conditional Categories. 95% confidence intervals (reported in parentheses) are based on 500 bootstraps.

68,000 legal codes in the ICD-10-CM versus 14,000 ICD-9-CM codes) are collapsed into a limited number of categories below.

It is worth mentioning that, in most years, the World Health Organization (WHO) makes changes to the ICD diagnoses. Most of these ICD changes are limited to descriptions and criteria for existing diagnosis codes, but occasionally new diagnoses are added. Less frequently, approximately once every 20 years, the WHO changes the version of its classification system more fundamentally, such as when it went from ICD-9 (1975) to ICD-10 (1994) to ICD-11 (scheduled for 2018). Many countries do not adopt the WHO ICD codes immediately or without modification. The United States (through the US National Center for Health Care Statistics together with CMS) modifies these

codes to create its own ICD-9-CM (clinical modification), which are updated annually on October 1. ICD-10-CM was only adopted in the United States in 2014, two decades after the WHO version change. These differences matter to risk adjusters since they create a necessity for each country to create and maintain risk adjustment classification systems consistent with their own coding system.[12]

3.4.3 Pharmacy Information

Pharmacy information is increasingly being used for risk adjustment, despite there being differing opinions about the desirability of doing so. On the positive side, drug use can often signal chronic conditions that are being controlled by medications, and which will be missed if only diagnoses are used for prediction. For some conditions, the pharmacy cost is an important component of plan obligations, and using pharmacy information can help predict this. Advocates of using pharmaceutical information argue that a drug prescription represents a validation of a doctor's opinion, whereas a diagnosis from a visit might only reflect a suspicion. On the negative side, using prescription drug information for risk adjustment may lead to too many prescriptions. Many drugs are given for prevention or maintenance, and basing payments on this information creates strong incentives for overuse.[13]

The Netherlands was the first to use pharmacy information for risk adjustment; it started in 2002 even before the use of diagnostic information in 2004. In 2017, the Dutch risk adjustment system used 33 pharmacy-based cost groups for risk adjustment in addition to diagnostic cost groups and diverse other measures (Chapter 14: Health Plan Payment in the Netherlands). Germany (Chapter 11: Health Plan Payment in Germany) also uses pharmacy information, although largely to validate or fill in for missing diagnoses. The United States is not currently using pharmaceutical information in its risk adjustment systems, although there was a proposal to do so for the Marketplaces (CMS, 2016c).

There are several challenges with using pharmaceutical information for prediction in risk adjustment. One challenge is the large number of different drugs prescribed. Individual drugs are identified by rich classification systems: National Drug Codes (NDC) in the United States and Anatomical Therapeutic Chemicals (ATC) in Europe. These highly detailed codes are mapped into categories of drugs, and selections of these categories are then incorporated in risk adjustment models. The US Food and Drug Administration (FDA) maintains a directory of allowed drugs that is updated daily, so keeping the list of allowed prescriptions up to date requires more effort than keeping up with the much more modest, and less frequent, diagnostic coding changes.[14] The World Health Organization updates the EU's ATC system only twice per year.

Even more challenging is that prescription practices and the plan predicted cost implications of individual drug categories can change rapidly and dramatically. The extremely popular allergy drug loratadine (better known by its brand name Claritin) went off patent in the United States in 2002, and then almost simultaneously switched from being a prescription drug to being sold over the counter (i.e., without a prescription). As a result, prescriptions for this drug, and indeed many other allergy medicines, plummeted. Visits to allergists and recordings of the diagnosis for allergies also declined. Diagnosis-based formulas predicting covered pharmacy spending overpredicted plan costs in this category until it was recalibrated, while pharmacy-based models tended to underpredict because of the disappearance of a large block of prescriptions.

The use of prescription pharmaceuticals for prediction is also complicated by the phenomenon of free samples dispensed by hospitals and clinics, unobserved pharmaceutical use in inpatient settings, and the fact that many drugs have more than one use. On this last point, some antihypertensive drugs have proven effective for preventing hair loss, while specific heart drugs have benefits in terms of sleep, acne, and weight loss. Changes in off-label uses of pharmaceuticals can change the prevalence and cost predictions of many drugs, requiring further attention. Having highlighted the challenges, one strength of pharmaceuticals is that the prescription information is generally available quickly. Moreover, some drugs are highly predictive of specific illnesses: insulin use is a very strong predictor that a person has type II diabetes. Both Germany and the Netherlands require more than one prescription of drugs in their payment formulas in order for that drug variable to be included. In the Netherlands most pharmacy-based cost groups require use of at least 181 defined daily dosages.

3.4.4 Prior-Year Spending Information

A frequently considered but rarely used risk adjuster is lagged spending. In our US MarketScan data on the commercially insured, spending in 2013 predicts spending in 2014 with a validated R-squared of 9.08%. This predictive power can be improved to 14.40% by top-coding spending used on the right-hand side at $250,000 and further improved to 21.41% by top-coding both the dependent and right-hand side variables at this level. The coefficient on the lagged spending variable in this last model is 0.49, implying that each extra dollar spent in year 1 predicts 49 cents in year 2. In terms of the Geruso and McGuire (2016) definition of power, these results imply that predictive models using lagged spending (in the form of a continuous variable) have a minimum power of 0.50 (i.e., half of spending this year is returned in payments next year. The reward to a plan is lost if a person changes plans.). While not a power of 1.0 this is still far from cost-based fee-for-service incentives where there is little incentive to reduce costs (power = 0).

Ellis and McGuire (2007) and Ellis et al. (2013a) demonstrate that one can improve prediction of year 2 spending using spending by type of service rather than total spending. Their work, using very large samples, finds that spending by type of service is even more predictive than diagnostic information (their R-squared increased from 10% to 15%). Such models would probably not be attractive to use as a payment model, in that there exist some types of spending for which a dollar spent on that service predicts more than a dollar of costs (and hence risk-adjusted payment) for the following year. Still, it is useful as a reminder that other information not desirable to use in risk adjustment will always be available for health plans to use for risk selection.

Although lagged spending is not used directly as a risk adjuster, the Dutch risk adjustment model (Chapter 14: Health Plan Payment in the Netherlands) includes dummy variables based on risk classes for people with high spending in multiple prior years, on the rationale that these people suffer from a chronic condition that may not be fully recognized by the existing diagnostic risk adjustors. Van Kleef and Van Vliet (2012) show that inclusion of these risk classes leads to substantial improvements in predictive value, even in a risk adjustment model including diagnoses- and pharmacy-based risk adjusters. Moreover, the Dutch risk adjustment model currently includes risk classes based on prior-year spending for two specific services, i.e., home care and geriatric rehabilitation care.

3.4.5 Healthcare Utilization Measures

In addition to diagnoses, pharmaceutical information, and spending, certain measures of prior-year utilization are also sometimes used as risk adjusters. The Netherlands uses flags for durable medical equipment, while Switzerland uses a dummy variable for whether or not a person has been hospitalized in the prior year. Moreover, diagnosis-based models include a reward for at least one claim associated with service with a diagnosis. It is difficult to assess the incentive effects of prior utilization on cost containment incentives, but certainly, including this as a risk adjustor reduces the power of the payment system, while improving the fit. Whether they are better or worse than much simpler cost-sharing or reinsurance programs remains to be investigated.

3.4.6 Medical Record Information

Ever since medical records became computerized there has been a desire to utilize this information for improved risk adjustment (Parkes, 2015). While the focus of this chapter is prediction of healthcare spending, the use of record information for predicting other outcomes is even more compelling. The attraction of medical record information is primarily that it is more

detailed, containing not only the diagnoses reported on claims, but also more secondary diagnoses and suspected conditions, lab test results and their interpretation, timing information, and information about who made the diagnosis. Despite the great promise of using medical record information, it has yet to be used in any risk-adjusted payment system. Medical record information is being used extensively for severity adjustment of outcomes other than spending,[15] and for reconciling and buttressing claims submissions that affect plan payments. There is an active industry in the US advising providers and plans on how to capture more diagnoses so as to increase plan revenue, but similar efforts to use this information to refine risk adjustment predictive models have not to our knowledge been developed. There are several obstacles to overcome before this can happen. First, medical records in the United States are not sufficiently standardized so that they can be easily used across different information systems or merged into a common format. Second, both privacy limitations and market competitiveness mean that many providers do not necessarily share their information with other providers, or even pharmacies and hospitals, so the medical records are often highly incomplete, both from using out-of-network providers and from whenever a patient changes their provider. Third, medical record information is inherently intermittent and, similar to diagnoses, only collected in the course of active medical treatment. Records tend to be collected when a patient is diseased, injured, in stress, being tested, or seeking preventive care. None of these is a random event, and the information collected is often very specific to that setting. None of the reviews and comparisons of risk adjusters by the Society of Actuaries in the United States or government health systems in Europe and Australia have used medical record information.[16]

3.4.7 Self-Reported Measures

Self-reported measures, which typically are collected via surveys, have long been considered good candidates for risk adjustment models. The central challenges are feasibility and bias. Feasibility relates to the high cost of surveys relative to using diagnoses from submitted claims, while bias relates to the challenges of getting adequate and representative response rates. A common type of self-reported information is perceived health status, either in its simplest form, which asks whether the respondent's health is excellent/very good/good/fair/poor, or in more elaborate forms such as the Short Form 36, which measures perceived health status along eight dimensions (Ware and Sherbourne, 1992). A different class of information measures functional health status, for which two common instruments ask about activities of daily living (ADLs) and instrumental activities of daily living (IADLs). A third class of self-reported measures relates to chronic conditions (e.g., diabetes, high blood pressure, asthma, etc.). Other self-reported measures include

information about lifestyle (smoking, drinking, and food), marital status, employment education, and whether a person can drive.

The usefulness of many of these self-reported measures for prediction has been evaluated numerous times. Much of the analysis of the Rand Health Experiment in the mid-1970s was conducted using survey information, although the modest sample size of about 10,000 person years of spending information substantially limited the statistical power for population-based prediction. Van de Ven and Ellis (2000) report fit measures (*R*-squared) for six early studies, all of which suffer from overfitting because they use very small sample sizes, with fewer than 30,000 respondents, but together question the value of using self-reported information.

Ellis, Fiebig et al. (2013b) report results using data from New South Wales, Australia, on 267,188 individuals over a 4-year panel data set, yielding a panel size of 787,000 person-years. Interestingly, the self-reported measures perform well in predicting use even 2 years before or after the survey was taken. Yet adding survey information in the form of 76 responses capturing each of the dimensions discussed above achieved an R-squared of only 10.2%, which was lower than those achieved by coarse diagnostic, pharmacy, or lagged utilization models. Survey results only added 0.8% points onto the 23.8% achieved using diagnosis, pharmacy, and lagged utilization measures. Gravelle et al. (2011) also explored the incremental information that can be acquired using surveys in addition to diagnostic information using UK data and found modest gains. Rose et al. (2016) examined the inclusion of self-reported health measures in risk adjustment formulas for accountable care organization (ACO) benchmarking and found that they decreased variation in differences between ACOs and local average FFS spending.[17] Similarly to socioeconomic variables, to which we now turn, the main value of including survey-based information is not its contribution to the overall fit of the risk adjustment model, but rather its value in improving predictions for identifiable individuals of concern.

3.4.8 Socioeconomic Variables

Demand-related variables such as race/ethnicity, income, poverty, housing, homelessness, unemployment, and language, and supply-related variables such as numbers of doctors and hospitals, provider distance and waiting time, and other measures of access are sometimes used to allocate funds geographically or to provider groups, but such information may not be available at the individual level. The UK payment system has gradually evolved from using aggregated information to using individual-level information to allocate budgets regionally and to providers such as hospitals and primary care providers. Gravelle et al. (2011) demonstrated that diagnosis-based risk adjusters largely eliminated the statistical contribution of most of the

demand- and supply-side variables for hospital budgets, while Dixon et al. (2011) found similar results for primary care trusts.

A major effort to improve risk adjustment and other payment formulas in the United States to better recognize "social risk factors" is currently mandated by Congress (US Department of HHS, 2016). Efforts are being made to incorporate these social determinants of health not only in risk adjustment, but also in hospital and other bundled payments.

A key challenge in using certain socioeconomic variables like race, language, income, or education, is that they may not be politically or socially feasible to include in a payment model: simply put, policymakers may not want to pay plans based on race, income, or language. Furthermore, if discrimination or access barriers are a problem, a subgroup (say a minority or nonnative language group) may currently receive too little health care. A regression-based model without any further adjustment will tend to perpetuate this inequity, paying less for this subgroup because it better predicts current spending. The classic risk adjustment solution is to simply omit this information from the predictive model, which makes this underpayment less visible, but does not address the inequity.

A related problem can arise when there are predictive variables that the regulator wants to exclude from a payment model for fairness reasons. For example, suppose spending is high in some region because of higher provider prices or higher intensity of treatment, and that these costs are correlated with other variables that the regulator does want to include. Simply dropping these variables can lead to an omitted-variable bias in the final payment formula. A correction for this problem is discussed in Box 3.3.

Ash et al. (2017) explore alternative ways of incorporating socioeconomic information while estimating individual-level risk adjustment models for Medicaid enrollees in Massachusetts. Using a relatively large sample ($N>800{,}000$ when pooled) they explore adding both individual-level administrative information, such as income-related Medicaid eligibility, as well as population-based measures merged on using the enrollees zip code and census block. Merging on census data at the census block level is interesting since potentially this can be done much more easily and cheaply than using survey information. Ash et al. (2017) collapse seven variables primarily related to income from the enrollee's neighborhood into a single neighborhood stress variable, and collapse two variables related to homelessness and frequent changes in mailing address into an insecure housing variable for inclusion in a regression model. Inclusion of these two new variables in the Fiscal Year 2017 payment formula for the state meaningfully improved predictive ratios for key vulnerable groups in this population although the contribution to model fit was trivial. This study is one of several in support of

BOX 3.3 Omitted-variable bias

An interesting consideration related to fairness is the distinction between risk factors for which cross-subsidization is desired (the so-called S-type factors) and risk factors for which cross-subsidization is not desired (the N-type factors; Van de Ven and Ellis, 2000). In most countries age, gender, and health status will probably be considered S-type factors, at least to a certain extent. But the regulator may decide that spending variation related to other factors, such as regional differences in supply and prices, should not be reflected in the subsidies. This has implications for risk adjustment.

When N-factors are independent of S-factors, compensation for N-factors can be avoided by simply omitting these factors from the regression model used to estimate risk-adjusted payments. Things are more complicated in the case that these two types of risk factors are correlated (Schokkaert et al., 2017). An example of such a correlation can be that sick people (S-factor) are concentrated in geographical areas with relatively high levels of supplier-induced demand (N-factor). If weights for S-factors are simply determined by a regression of observed spending on the S-factors, these weights will suffer from an omitted-variable bias. Consequently, the subsidies will (partly) reflect the spending variation due to the N-factors. Empirical illustrations by Schokkaert et al. (2004), Van Kleef et al. (2008), and Stam et al. (2010) have shown that this bias can be substantial. Different solutions have been proposed to overcome this omitted-variable bias, including Schokkaert and Van de Voorde (2004) Van Kleef et al. (2008), and Stam et al. (2010). Further discussion is provided in Chapter 7, Risk Adjustment in Belgium: Why and How to Introduce Socioeconomic Variables in Health Plan Payment, and Chapter 14, Health Plan Payment in the Netherlands.

new US initiatives to reflect social risk factors in healthcare payments by the National Quality Forum (NQF, 2014) and the National Academy of Science, Engineering and Medicine (NAS, 2016).

In Europe, sociodemographic variables are commonly used in risk adjustment models. The Dutch risk adjustment model includes risk adjusters based on household income, household size, and employment status (see Chapter 14: Health Plan Payment in the Netherlands for more details). Similar types of information are used in Belgium (see Chapter 7: Risk Adjustment in Belgium: Why and How to Introduce Socioeconomic Variables in Health Plan Payment). Though these risk adjusters do not generally lead to substantial increases in R-squared, including them in the predictive model can redistribute large amounts of money (e.g., from plans with relatively many self-employed to plans with many unemployed. See Chapter 7: Risk Adjustment in Belgium: Why and How to Introduce Socioeconomic Variables in Health Plan Payment for an extensive discussion of this point.

3.5 CHOICE OF TIMEFRAME FOR DATA USED FOR PREDICTION

The time interval over which risk adjusters are observed is called the "base period" by risk adjustment modelers, while the period for which spending is predicted is called the "prediction period" (Ash et al., 1989, 2000; Kautter, 2014). Several alternatives for choosing the base and prediction periods are possible.

3.5.1 Prospective Versus Concurrent Risk Adjusters

Two broad empirical frameworks are commonly used to characterize the information used for risk adjustment. Prospective risk adjusters come from a base period that precedes and does not overlap with the prediction period. Concurrent risk adjusters use information from a base period that coincides with the prediction period. For example, diagnoses and/or pharmaceuticals from year 1 are used to "predict" spending in year 1 in a concurrent model. Concurrent models require that the regulator must wait until the end of the year to observe all of the information used for prediction.[18]

It used to be easy to classify risk adjustment formulas as either prospective or concurrent. However, many formulas today use both types of information. Prospective models have more power (Geruso and McGuire, 2016) than concurrent models, and are less prone to endogenous signals, since diagnoses for acute conditions that are treated and resolved within 1 year matter little for prospective models. On the other hand, prospective models require more data and require a separate formula to use with newly arriving enrollees, for whom prior year information is never available. Another disadvantage of prospective models is that they have lower predictive power, leaving more risk and uncertainty for health plans. Concurrent plans suffer from greater endogeneity of diagnoses, and the data arrive for payment 1 year later, which creates its own uncertainty, administrative burdens, and planning challenges. Typically, concurrent models use provisional payments, but some plans and providers strongly resist the revenue uncertainty of retroactive payment adjustments, even though the same plans readily accept cost uncertainty.

Prospective diagnosis-based information is used for the US Medicare Advantage and Part D payment systems, and in Germany, Switzerland, the Netherlands, and Belgium. However, each system also uses concurrent information for age and sex, as well as for diverse other variables such as institutionalization and Medicaid eligibility (United States), and income (Belgium). Concurrent risk adjustment is used in some US Medicaid systems, and for Marketplace enrollees, where it is particularly attractive since turnover tends to be high in such programs, so prior year information is commonly missing for many enrollees.

3.5.2 Hybrid Risk Adjusters

In addition to prospective and concurrent risk adjustment, another possibility receiving attention is hybrid risk adjustment, which uses both concurrent and prior year information for prediction. This hybrid could be in diagnostic information, procedures, or specific types of services that are calculated separately. Dudley et al. (2003) were perhaps the first to examine such a framework, and introduce the terminology of "hybrid risk adjustment".[19] In their framework, anyone with a specified high-cost event, including pregnancies, heart attacks, and other high cost events, mostly inpatient driven, would be paid on a concurrent basis. Specifically, they identified 100 verifiable, expensive, predictive conditions that occurred among 9.3% of the population, and used a concurrent framework to pay for this subsample of the population while paying for the remaining 90.7% of the population using a prospective HCC framework. Their pioneering early work achieved an *R*-squared of 26% versus a prospective R-squared of only 8%. Further research in this direction was conducted by García-Goñi et al. (2009) to predict drug expenditures using Spanish data with similar gains in predictive power. Belgium and the Netherlands use a hybrid approach in which concurrent socioeconomic information and age and sex are combined with prospective diagnoses and utilization measures.

Any payment system that uses ex post information, such as reinsurance or outlier payments, is also a form of hybrid risk adjustment. In particular, the recent proposal by Layton and McGuire (2017) to use dollars of spending above a threshold as a risk adjuster and fixing the coefficient at the desired share (making it equivalent to reinsurance) is inherently a hybrid framework. Chapter 4, Risk Sharing, contains a more detailed discussion on this point.

3.6 CHOICE OF THE OBJECTIVE FUNCTION FOR ESTIMATING RISK ADJUSTMENT

Perhaps the most important topic for risk adjustment is the choice of the objective function to be maximized and the algorithm for maximizing it. This section reviews the key concepts relevant to objective functions, and how they are incorporated in risk adjustment model design and selection. We start by distinguishing two broad approaches to risk adjustment: traditional risk adjustment and optimal risk adjustment. While there is considerable overlap between the two approaches, one interesting theme is that traditional risk adjustment has often focused on the selection of risk adjusters for a given objective function, while optimal risk adjustment takes the risk adjusters as given and focuses on the selection of coefficients to maximize the objective function. New approaches, including machine learning techniques discussed below, try to do both simultaneously.

3.6.1 Traditional Risk Adjustment

The traditional approach to risk adjustment, as embodied in Ash et al. (2000), Pope et al. (2004), Kautter et al. (2014) and the payment systems of the Netherlands and Germany, has emphasized accuracy in matching plan obligations to predictable spending at the individual level while incorporating concerns about selection, gaming, coding accuracy, and fairness, as presented in the first 10 principles of Box 3.1.[20]A commonly stated objective is to "level the playing field" so that health plans do not gain from attracting profitable enrollees, nor lose from attracting unprofitable ones (Ash et al., 1989). Traditional risk adjustment changes health plan profit incentives by paying more for enrollees predicted to cost more and less for enrollees predicted to cost less. It has generally focused on the careful choice of risk adjusters, as well as the constraints and functional form issues. At its heart, traditional risk adjustment attempts to pay each health plan the predicted cost of each enrollee conditional on the choices of risk adjuster variables and model structure, while minimizing the unexplained variation in spending or equivalently, maximizing the model fit. Although diverse objective functions are often considered, the overwhelming favorite objective function of traditional risk adjustment is to minimize the variance of the unexplained part of spending, i.e., the sum of squared residuals between actual and predicted costs, which when normalized by the sum of squared deviations of the dependent variable to its mean is called the R-squared.

Because of its central role as a metric of risk adjustment performance, it is worth reviewing the formula and properties of the R-squared (Van Veen et al, 2015a). This metric has several attractive features. One is that because it is a unit free number, it can be compared across specifications, dependent variables, time, and samples. It also has an easy conceptual interpretation as the fraction of the total variance in the dependent variable explained by the model. We follow Ash et al. (1989, 2000) and report the *R*-squared as a percentage rather than a ratio. The *R*-squared can be calculated as

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3.1}$$

where f_i is the prediction for observation i and y_i is the actual value, and $\bar{y}$ is the sample mean of y_i. Note that the *R*-squared can be calculated using this formula for any predictive model, even when f_i is not the result of a least squares regression. Table 3.2 presents within-sample *R*-squared measures using our test sample for three alternative dependent variables and four alternative sets of right-hand side variables, which we discuss further below.

3.6.2 Optimal Risk Adjustment

Economic models of risk selection (Glazer and McGuire, 2000; Layton et al., 2017) imply that traditional risk adjustment, by focusing on explaining as much of the variance as possible, will in general not fully solve efficiency problems related to selection except under strong and implausible assumptions.[21] Glazer and McGuire (2000) show that simply maximizing the fit of a model can still lead to inefficiencies when health plans can distort premiums, plan characteristics, or the availability of specific services to attract profitable enrollees. These new models have led to an expanded set of objective functions, or welfare metrics for measuring the performance of health plan or provider payment formulas. The term "optimal" is used to characterize the maximization of a specific economic objective, rather than to signify that there is no possibility that even better risk adjustment models are not possible.

Optimal risk adjustment models start with a theory-based objective function and conceptualize risk adjustment as a tool for selecting risk adjustment weights to maximize that objective. A variety of different objective functions has been used. Glazer and McGuire (2000) use efficiency of service provision as the objective and assume health plans maximize profits through their choice of shadow prices that ration consumer access to various services. Since risk adjustment signals are imperfect, they propose overpaying (underpaying) for weak signals to correct capitation incentives to undersupply (oversupply) certain services. Building on this insight Ellis and McGuire (2007), and more recently McGuire et al. (2014) and Ellis et al. (2017b) calculate how various risk adjustment models moderate plan incentives to distort benefits and services. Minimizing incentives to distort is a conceptually attractive concept, although not a complete objective function to assume for a health plan payment system, since it reflects the health plan's private objective, not society's social objective. Einav and Finkelstein (2011), McGuire et al. (2014), and Layton et al. (2017) show how premium subsidies, risk sharing, and fairness objectives can also be incorporated into the risk adjustment calculations by specifying a social objective function to use when calculating the payment system. Insights from these papers are discussed in Chapter 4, Risk sharing and Chapter 5, Evaluating the Performance of Health Plan Payment Systems.

3.7 FUNCTIONAL FORM AND MODEL SPECIFICATION

We now turn to discussing how risk adjusters are incorporated in the prediction formulas, which includes consideration of the structure of how predictors are used, the functional form of the dependent variable, and the use of constraints and manipulations on the risk adjusters.

3.7.1 Categorical Versus Additive Models

Since the origins of risk adjustment in the 1980s, two different frameworks have been advocated: Categorical models that place each individual uniquely in a single cell, and additive models that do not classify each individual into one category but instead classify consumers along multiple dimensions. Categorical models, which reduce the estimation problem to calculating the mean for each rate cell, are used in Switzerland and Colombia, as well as in 3M's Clinically Related Groups (CRG) system in the United States.[22] An additive regression approach is more flexible than a categorical model in that a larger number of interaction terms can be incorporated in the formula without loss of power.[23] The essential difference in the modeling approach is whether predictions are additive in the explanatory factors or fundamentally mutually exclusive, as with a branching structure.

In head-to-head comparisons of models by research on large samples (i.e., with over one million observations), such as that conducted by the US Society of Actuaries (SOA) (Dunn et al., 1996; Winkelman and Mehmud, 2007; Hileman and Steele, 2016), additive models have consistently performed as well or better than other models (including categorical ones) on standard statistical measures of performance (R-squared, RMSE, and predictive ratios for policy-relevant subgroups). Cid et al. (2016) provides a summary of eight different international studies comparing various risk adjustment models, including both categorical and additive models, supporting the superior predictive power of additive models. The last two SOA studies also include machine learning models among the set of models analyzed, but in each case the attention given to machine learning was fairly cursory. We discuss further machine learning techniques below, some of which also use a categorical rather than an additive framework.

3.7.2 Transformations of the Dependent Variable

All risk adjustment performance measures are affected by transformations of the dependent variable, as discussed in Van de Ven and Ellis (2000). Such transformations are commonly done to reduce model sensitivity to skewness and kurtosis. One common transformation is to top-code the dependent variable at some level such as $250,000.[24] Hence, if Y is total spending, the transformed dependent variable Y^{TC} is the minimum of actual spending and $250,000. This has the effect of minimizing the impact of extreme outliers. It, of course, means that predicted spending does not hit the mean spending conditional on the regressors, although depending on the distributions, the resulting bias may not be large, and it may be outweighed by better precision in the estimated coefficients.

Top-coding, which retains individuals with very high levels of payments, is preferred to dropping high cost observations altogether, because extremely

high costs are often predictably associated with specific conditions. In samples of ten million or more individuals, top-coding may not be needed, since even random high-cost enrollees will be averaged out, however, for smaller samples these extreme outliers can have a dramatic effect on individual coefficients. Alternative values for top-coding ranging from $50,000 to $1 million have sometimes been used.[25] Resetting negative spending amounts to zero is also commonly done.[26]

A second, more dramatic transformation is to use natural logarithms of spending as the dependent variable. Since annual health spending is often zero, it is common to add one to spending before taking logs. If negative values of spending, Y, occur, these must also be eliminated by resetting them to one. Hence the natural log of Y, LnY, is calculated as

$$\text{Ln}Y = \text{Ln}(\max(1, Y + 1)) \tag{3.2}$$

Tables 3.2 and 3.3 present R-squared, RMSE, and MAE, respectively, for a variety of model specifications, where explanatory variables vary across rows, while dependent variables vary across columns. Across the three columns, three different dependent variables are used: untop-coded spending, $250,000 top-coded spending, and natural log of spending.[27] The R-squared shown here was calculated in the log form. For comparison across specifications, predictions from the log linear model need to be transformed back into their raw dollar level, such models invariably do worse than linear models once this is done.

Results from four model specifications are shown across rows, with three concurrent specifications, and one prospective. The first two rows use only age−sex categorical variables to predict concurrent spending, while the third row adds 394 DxCG-HCCs variables to the concurrent model. The final row in each table shows prospective model results, using the same specification as for the concurrent model. Among the two age−sex models, the first uses 28 age−sex groups, as used by the ACA Marketplace risk adjustment model,[28] while the second row uses 130 age−sex dummies, with sex interacted with 1-year age dummies. The take away from the comparison of the two age−sex-only models is that saturating the model with annual dummies, while capturing the full nonlinearity shown in Fig. 3.2, does not meaningfully improve model performance by any of the three metrics.

The first column of Table 3.2 shows the results of the model for predicting spending (with no top-coding). Using only age and sex information predicts 1.5% of the total variation but the fit can be improved simply by redefining the outcome variable. Indeed, top-coding spending at $250,000 improves the fit to almost 3% of spending variation. This improvement is explained by the large variation in spending among the top spenders of the distribution, for which their spending levels are better related to their unobserved individual characteristics rather than their age or sex. As discussed in Chapter 4, Risk Sharing, outlier policies such as reinsurance deal with the

same concerns about outliers as top-codings. Another way of removing the effect of the outliers is the logarithmic transformation (column 3), which smooths the variation in spending, specifically for larger values. Furthermore, because of skewness, this transformation also helps at the bottom tail of the distribution. Numerous studies have shown that the residuals after a log transformation do a better job of predicting the logged value (e.g., Jones, 2011). Simply using the logarithmic transformation improves the *R*-squared to over 10% in a model with age and gender only, but the gain is illusory: payments have to be made in monetary levels, not log of spending. Every loglinear model estimated to date is inferior in terms of R-squared in large samples to linear regression models when used to predict levels of spending while accommodating partial-year eligibles, i.e., for the primary purpose of risk adjustment models (Winkelman and Mehmud, 2007; Jones, 2011; Ellis et al., 2013a,b).

Transforming the dependent variable also has implications for the precision of the goodness-of-fit measures. In fact, the confidence intervals around the *R*-squared when the data are not top-coded are close to 30% of the point estimate, even with 21 million observations. These large confidence intervals arise due to the influence of outliers affecting the unexplained spending variation in the data. Top-coding these outliers—or removing them completely from the risk adjustment model—not only increases the *R*-squared of the model but also decreases the confidence interval to negligible amounts. The log transformation has the same impact on the confidence interval since it also removes the effect of outliers.

Model comparisons based on spending with no top-coding might lead to misleading results, as the estimated R-squared is sensitive to the particular draw of observations. In this sense, top-coding the dependent variable before the analysis is a more robust approach to compare different models.

3.7.3 Diagnostic Hierarchies

Even after grouping diagnoses into a manageable number of discrete categories, there are a number of strategies for introducing them into a predictive model. The simplest way is to just include them all, and decide ex ante which, if any interaction, terms enter in. The problem with this approach is that for a reasonably well-specified system with over 200 categories, there are potentially 20,000 two-way interactions terms that could be considered, with a vastly larger number of three- and higher-level interactions. Machine learning algorithms can be considered to choose among this large number of potential interactions; however, they may sacrifice accuracy for simplicity when too many variables are introduced for consideration.

The overfitting problem is particularly problematic when diagnosis categories are strongly related, which is to say that they are highly collinear: either condition A or B is needed in the model but perhaps not both. To

address this, as well as to reduce sensitivity to endogenous diagnostic coding, Ash et al. (1989) developed the concept of diagnostic hierarchies, captured in the Diagnostic Cost Group (DCG) classification system. The DCG single-hierarchy approach was further elaborated in what came to be known as the HCC approach that underlies the risk adjustment models used in the United States for Medicare Advantage, Medicare Part D, and the Marketplace, as well as in Germany. In the original DCG system 78 disease categories (or cost groups) were entered into an algorithm in which only the highest cost or most severe group overall in the sample was used for predicting individual payments. A version of the DCG approach is still used in the Netherlands. The HCC system expanded the DCG framework by considering multiple rather than only one hierarchy. The current CMS HCC system defines 30 broad body systems when imposing hierarchies, so that conditions affecting one body system do not affect risk adjusters arising from other body systems. Rather than the DCG predictions using only the single most serious condition a patient has in the year, the HCC framework uses one or more of the most serious conditions within each of 30 body systems for prediction.

Consider the following extended example to see how the hierarchical grouping works. Assume there are two diseases of interest, called A and B. For prediction, one could consider using dummy variables D_A, D_B, and $D_{A+B} = D_{A^*}D_B$. Several specifications are possible. One possibility is that A and B are simply additive, so that the first two direct effects are statistically significant, while the interaction term is not. The insignificance of the interaction term occurs frequently because spending on most diseases affecting different body systems is additive: the incremental cost of a broken arm or an allergy diagnosis is hardly affected by coexisting conditions.

Another second possibility is that conditions A and B complicate one another. Diabetes, cancer, immune disorders, heart conditions, pregnancy, and liver disorders, for instance, tend to complicate the treatment and hence the cost of other conditions. For these conditions, not only will D_A and D_B be significant but also their interaction D_{A+B} will be positive, and including this interaction term may be desirable. Indeed, the risk adjustment models used in the United States for Medicare Advantage, prescription drug spending, and the Marketplace, and the German risk adjustment formula contain a small number of interaction terms across body systems for some such situations.

A third and very common possibility is that conditions A and B are related conditions such that A represents a more serious manifestation of a given disease than B. For example, condition A might differ from condition B due to the presence of a complicating condition. Here, D_A will have a higher coefficient than D_B, but for a person with both A and B coded, then only having the more serious diagnosis A may matter. If true, then when all three terms, D_A, D_B, and D_{A+B} are included in a regression, then the D_{A+B}

dummy coefficient will be equal to the negative of the coefficient on D_B, signifying no incremental cost of B conditional on A.

Imprecise diagnostic coding in practice increases the frequency of this third possibility. Physicians choose how much effort to put into coding: even when a more serious diagnosis is present (diabetes with renal manifestations) they may only code a less specific condition (diabetes, unspecified) since that is all that matters for their reimbursement for the current visit. In such cases, the less specific condition can be uninformative in combination with the more serious code. If the two codes only appear for the same patient jointly due to imprecise coding of this form, then a regression model will estimate the coefficient on D_{A+B} to be the negative of the coefficient on D_B, just as with the complicating condition example above. For this third possibility, whereby coding is imprecise or only the more serious manifestation matters, imposing hierarchies makes use of this knowledge to specify a more parsimonious model and reduce the problem of overfitting. Instead of including three terms in the regression, D_A, D_B, and D_{A+B}, the modeler imposes the constraint that the coefficients on D_B, and D_{A+B} are equal but of opposite signs. Imposing this constraint is numerically equivalent to including only two terms D_A, and $D_{B\sim A}$, where $D_{B\sim A}$ is an indicator variable for the presence of disease B without A being present, which is what imposing a hierarchy does: only recognizes B when not accompanied by A. In effect, hierarchies embody a clinical rationale for excluding the vast majority of potential two-way interactions in the risk adjustment model. The 2017 CMS-HCC classification system includes 79 HCCs but imposes 57 hierarchical restrictions that reduce the number of regressors. Pope et al. (2004) document that adding additional interactions or omitting hierarchies has very little impact on model fit.

The ability to use a priori clinical criteria to constrain interaction terms and exclude variables from a risk adjustment formula is a major argument in favor of hierarchical classification systems. This statistical argument is true whether the system uses a single hierarchy, such as the DCG system used in the Netherlands, or multiple hierarchy systems, such as the various HCC models used in the United States and Germany. A second and equally important rationale is that hierarchies also reduce the sensitivity of formulas to gaming. One of the simplest ways of upcoding is to add all of the less serious conditions (cough, chest pain) to patients with more serious conditions (lung cancer). Additive models, without hierarchies, will tend to keep increasing predictions as more (less serious) conditions are reported.[29]

Similar issues over hierarchies arise with the combinations of diagnostic and pharmaceutical information. For example, type I diabetes can either be detected through a diagnosis code, or through prescriptions for insulin. What is to be done when both signals are encountered? Following Germany, the 2016 proposal for the ACA Marketplace is only to recognize the insulin prescription when the diagnosis has not been recorded, which is a form of

hierarchy imposed across sources of information. Other possibilities for informed variable selection also exist when adding demographic information, or considering models for specialized populations, to which we now turn.

3.7.4 Excluding Risk Adjusters

We have just argued that imposing hierarchies is equivalent to including interaction terms but constraining the coefficient on the interaction to be the negative of the coefficient of the lower-cost HCC. A related approach for traditional risk adjustment is to exclude risk adjusters when estimating the formula due to clinical or policy-motivated criteria when selecting the preferred risk adjustment model. Traditional risk adjustment often excludes eligibility or socioeconomic adjusters even when they are highly significant, in order to avoid undesirable incentives or to reduce unfairness. (See Box 3.3 for an example involving fairness.) The 2017 HHS-HCC model increased the number of HCCs from 201 in the CMS-HCC Medicare Advantage program to 264 HHS-HCCs for the Marketplace, of which 137 HCCs were excluded, leaving 127 HCCs for potential inclusion in the model. Constraints were then imposed across 26 of these remaining HCCs, thereby reducing the total number of HHS-HCCs in the model to 101 (CMS, 2016c). Although Kautter et al. (2014) provides a valuable overview of the final HHS-HCC model chosen, details of the process used for the selection of HCCs are not available. The 10 principles shown in Box 3.2 above likely played a central role. Based on the earlier work for CMS documented in Pope et al. (2004), principle 2—excluding conditions that are not predictive, principle 5—encouraging specific coding, and principle 10—excluding discretionary categories, are the three most important reasons for omitting HCCs. Principle 6—not to include coding proliferation, is another important reason why some HCCs are omitted.

3.7.5 Constrained Regression Models

An important new direction for risk adjustment estimation is reflected in a series of recent papers by Van Kleef et al. (2016), Layton et al. (2016), and Bergquist et al. (2018) who demonstrate the value of constrained regression models to simultaneously balance model fit with achievement of other goals. Van Kleef et al. (2016) extend the conceptual work of Glazer and McGuire (2002) and argue that selection incentives for specific types of services can be addressed by using constrained least squares regression techniques. If the traditional risk adjustment formula allocates too little money for people receiving home care services, e.g., then imposing constraints on the estimated coefficients can ensure more funding goes to this group, mitigating selection-related incentives. This method can reallocate funds without increasing the total budget. Van Kleef et al. (2016) use a large sample of

Dutch enrollees to show proof of concept in which underpayment for both physiotherapy and home healthcare services can be completely eliminated in constrained regressions in which the sum of squared residuals is minimized while at the same time forcing predicted payments for the group of people using these two types of services exactly match total spending on this group. Constraints will change the payments for other groups as well. Notably, as Van Kleef et al. (2016) show, a number of other previously underpaid groups have payments increased with the introduction of the constraint on home care underspending. Funding for some other groups must go down, of course, to compensate for the increase for the previously underfunded groups.

Constrained regressions can be used to address other objectives of plan payment as well. Layton et al. (2016) introduce a selection incentive metric to be minimized while estimating a regression model. In their framework, rather than estimating a model and then evaluating how well it does at reducing selection incentives, they choose a social objective function that includes both selection incentives and profit variation as objectives, and estimate models that weight both objectives. They illustrate their model using Dutch data to demonstrate how it can reduce selection incentives for 10 healthcare services. Constrained regression risk adjustment is attractive conceptually, and deserving of further research. For practical implementation, it remains to be seen whether the methodology embodied in the constraint is acceptable to policymakers, whether the models are sufficiently understandable, and whether the effects on other groups in aggregate are acceptable.

3.7.6 Quantile Regression Models

An alternative method for incorporating an "optimal risk adjustment" perspective concerns into the risk adjustment estimation is exemplified in the work of Normann Lorenz (2015, 2017). This new approach conceptualizes insurers' activities for risk selection as a contest in which insurers compete to attract enrollees. For the contest success function used in most of the contest literature, optimal transfers for a risk adjustment scheme should be determined by maximizing the Cummings Prediction Measure (CPM) via a quantile regression for the median. Depending on whether it is easier to attract healthy or repel sicker subsets of the population, other percentiles than the median should be estimated. However, quantile regressions for the median (and other percentiles) result in very biased estimates of the mean (because the median is smaller than the mean). Therefore, a constraint to ensure that mean spending is also the mean of predictions can be incorporated. With this constraint, estimates do not depend on the percentile used, so the optimal payments do not depend on whether insurers compete in attracting or repelling individuals. Empirical results show that constrained quantile regressions increase the CPM somewhat, but computation times for estimation are still an issue for complex models and very large data sets

(Lorenz et al., 2017). Whether this approach will prove attractive for policy adoption remains to be seen.

3.7.7 Machine Learning Methods

Machine learning algorithms provide automated tools to learn adaptively, based on the data, about the relationships between variables. This can be attractive since the underlying functional form of the data is generally unknown, and the algorithms can also select variables from among a large set of predictors. Incorporation of both investigator knowledge and automation may help yield improved yet interpretable prediction functions. Given the complexity involved in designing risk adjustment formulas, there is growing interest in exploring the potential of machine learning techniques, particularly as computational demands have become less onerous over time. In this section, we provide an overview of the use of machine learning for risk adjustment model selection, focusing attention on the class of nonparametric statistical models of the set of possible probability distributions of our data.

3.7.7.1 From Objective Functions to Loss Functions

Machine learning algorithms for general prediction problems have been developed across the computer science, statistics, and data science literature. The starting point is typically to define the goals for performance of an algorithm, often specified as a loss function to be minimized. One candidate loss function is to simply use the sum of squared errors commonly used for traditional risk adjustment, called the general L_2 loss function:

$$\min_{\hat{E}(Y|X)} \left\{ \sum_{i=1}^{N} \frac{1}{N} \left(y_i - \hat{y}_i \right)^2 \right\} \tag{3.3}$$

This L_2 loss function, which can be used with regression methods or a machine learning approach, is minimized by the conditional mean of our outcome, thus we minimize over candidate estimators $\hat{E}(Y|X)$ of the conditional mean $E(Y|X)$. For each algorithm (i.e., estimator that takes our covariate predictors and maps them to the real line as predicted outcome values) we can evaluate performance based on the chosen loss function and, preferably, out-of-sample validation criteria. A well-known limitation of the L_2 loss function is that it can lead to poor performance when the data deviate dramatically from the normal distribution, particularly when sample sizes are less than a million observations.

Other loss functions can be considered including a quasi-log-likelihood loss for bounded continuous outcomes, which would be an interesting approach given the bounded nature of spending. This quasi-log-likelihood loss allows for a transformed continuous outcome variable bounded within

[0,1] combined with the negative log likelihood loss function often used with binary outcomes. This approach can also be used to reduce the impact of outliers on the payment formula without either top-coding or excluding outliers. The quasi-log-likelihood loss has been used for continuous outcomes in earlier statistics literature (Wedderburn, 1974; McCullagh, 1983), and recently for effect estimation (Gruber and van der Laan, 2010), but has not been used to date for plan payment risk adjustment or machine-learning-based prediction. Transformed outcomes on the log scale can also guide the choice of loss function.

3.7.7.2 Algorithms

There are many broad classes of machine learning methods we might consider for the development of risk adjustment formulas. One of the most straightforward approaches that can be understood in the context of the regression-based OLS techniques is penalized regression, which allows for greater bias in exchange for smaller variance.[30] For linear regressions, the function to minimize can be characterized in its simplest form by:

$$\min_{\beta}\left\{\sum_{i=1}^{N}\frac{1}{N}(y_i - X\beta)^2 + \lambda R[\beta]\right\} \tag{3.4}$$

where the first term is the familiar mean squared error and the second term, $R[\beta]$, is the regularizer or penalty function, intended to capture the nature and extent of the bias accepted, or alternatively to punish the predictive model for using too many regressors or allowing coefficients to deviate too widely, which may be a priori implausible. There are many possibilities to use for regularizer function, including the sum of the absolute value of the coefficients (referred to as the lasso—least absolute shrinkage and selection operator—estimator) or the squared sum of the coefficients (a ridge estimator). Since lasso estimators put a penalty on the number of coefficients, they generate more parsimonious estimators with fewer coefficients (the functional form specification). Ridge regression will produce an estimator with coefficients shrunk toward zero, but none will be exactly zero. General elastic nets that consider combinations of the ridge and lasso penalties can also be implemented. Lasso, ridge, and general elastic net estimators have been used within ensembles for risk adjustment, discussed below.

Decision trees are another popular technique and can be described as dividing the covariate space based on homogeneity for the outcome. Trees have become widely used due to their ability to "let the data speak" and discover potentially important interactions among covariates data-adaptively. Given the sheer volume of possible interaction terms that could enter a risk adjustment formula, automating this choice with a tool such as decision trees may be desirable. To demonstrate briefly the potential advantages of tree-based methods for capturing unique interactions, consider the following

simple example. Suppose a substantial increase in spending was associated with having disease condition A, but only when age is higher than 35. A regression tree could find such an interaction that was not known a priori nor simple to include in a parametric regression without some type of data-adaptive technique to discover it.

Several papers have studied single regression trees as a primary alternative method for predicting healthcare spending. Relles et al. (2002) examined the use of a simple single regression tree for payment in inpatient rehabilitation and found that its predictive performance was very similar to other techniques. Other work, by Drozd et al. (2006), explored psychiatric payments using simple single regression trees, and their results showed an improved performance of about 20% compared to a proposed traditional nontree-based estimator. Buchner et al. (2017) implemented a regression tree approach to assess interaction terms for improving model fit. Using a sample size of 2.9 million individuals from a major German health plan, they obtain an improvement in the adjusted R-squared of from 25.43% to 25.81%, which they describe as a marginal improvement. In a similar exercise based on the Dutch risk adjustment formula of 2014, Van Veen et al. (2017) find an improvement in the adjusted R-squared of from 25.56% to 27.34%. In general, using only a single regression tree will generate a formula with high variance: averaging over many trees can improve performance. Another popular method is to create "random forests" that average over many trees (e.g., 500 or 1000) using bootstrapped samples and random subsets of covariates, to reduce variability. However, even when incorporating cross-validation, random forests may still overfit, so it is important to consider imposing constraints on the algorithm, such as on the number of terminal nodes, observations per terminal node, number of trees, or covariates allowed for each tree.

Random forests are therefore a specific type of "ensemble" algorithm, which we will define broadly as an algorithm that incorporates multiple algorithms, selecting either a single algorithm from among the collection or an average of the collection of algorithms. Random forests average over only a collection of trees, whereas a generalization of stacking algorithms (Wolpert, 1992; Breiman, 1996) called "super learning" (van der Laan et al., 2007) averages over a collection of (potentially) disparate algorithm types that may search the model space in different ways. This is accomplished by running each algorithm with K-fold cross-validation and then regressing the spending outcome on the cross-validated predicted values for each algorithm to estimate the weight vector. A key advantage of a general ensembling approach, such as a super learner, is that investigators do not need to decide beforehand which single algorithm to select; there is no penalty for implementing many in this a priori specified framework. The researcher protects against a potentially poor choice of an estimator by running multiple algorithms.

Rose (2016) developed a super learner for total annual spending in a sample of MarketScan data comparing the performance of 14 algorithm

implementations to the super learner based on a validation R-squared, considering a full set of variables, including demographic information and 74 HHS-HCCs, as well as a data-adaptively selected set of 10 variables identified by random forests for each algorithm. The collection of algorithms included OLS, penalized regressions, single regression trees, and random forests, among others. Super learner yielded a minor improvement in R^2 and the results also showed that the reduced set of 10 variables retained much of the predictive performance of the full set in most of the algorithms (e.g., OLS regression had a validation *R*-squared of 25% for the full set vs 23% for the reduced set). Further work is needed to adequately understand the policy implications of removing such a large number of variables, especially on the basis of R-squared, without considering predictive ratios and other metrics. Replication studies in other populations, including Medicare, are ongoing. Shrestha et al. (2017) present a super learner prediction function for mental health spending in MarketScan using mental health diagnosis information and comparing three sets of mental health diagnosis variables joined with demographic information: HHS-HCCs, AHRQ's clinical classification software (CCS) categories, and HHS-HCC plus CCS categories. Here, OLS regression was nontrivially outperformed by both super learning (14% better) and random forests (10% better) with respect to validation *R*-squared. This paper also finds CCS categories to be more predictive of mental health spending than HHS-HCCs. The flexibility of the super learning framework allowed these comparisons to be a priori specified and run in one global algorithm: considering many different algorithms with alternative tuning parameters and comparing different sets of variables within each algorithm. There are many other machine learning techniques; for a thorough discussion see Friedman et al. (2001).

Although the machine learning results are encouraging, machine learning techniques are not ready to replace more traditional risk adjustment models for plan payment purposes. Machine learning techniques can identify subsets of variables or interactions to include in more traditional methods, but have not yet shown their superiority in validated predictive power on large samples with millions of enrollees. We suspect that this is so for two reasons. One reason is that the greater computational burdens of machine learning techniques have until recently meant that the methods were only commonly used on samples of less than one million observations, which precludes being able to estimate additive or categorical models that allow as many risk adjusters to be used as in traditional risk adjustment. A second reason is that machine learning methods generally result in prediction functions that clinicians and policymakers find unintuitive or hard to explain. As noted above, this lack of transparency could, however, be advantageous to prevent strategic responses to the risk adjustment formula, such as by "upcoding" diagnoses or undersupplying services to unprofitable enrollees. More work is needed to understand the policy implications of deploying these techniques.

3.8 RISK ADJUSTMENT MODEL IMPLEMENTATION ISSUES

We now turn to the implementation of risk adjustment formulas, which is sometimes called risk equalization. Risk equalization involves choosing the plan enrollees among whom payments are to be reallocated, and defining precisely how available funds are used to make payments at the plan level. Since these allocations depend upon many detailed implementation decisions that tend to be country-specific, the interested reader should consult the individual country/sector chapters in Part II of this volume. Here we try to touch on some common challenges and selected solutions.

3.8.1 The Population Groups for Which Risk Is to Be Equalized

In Box 3.1 we note that in addition to choosing the sample on which to estimate the formulas, one must also define the population to whom the formula is applied. The two need not be the same. In the United States, it is often a completely separate population from the one on which the risk adjustment formula is estimated. Moreover, many systems decide to equalize payments only within certain subsets of the full population. In the US Medicaid, and US Marketplace, for instance, risk adjustment is only used to reallocate funds within each state, although for the Marketplace, risk sharing is done at a national level. In Switzerland, risk adjustment and risk sharing are done at a canton level.

The choice of region, demographic subsets, or an all-encompassing group for risk equalization is often driven by political considerations. From a risk perspective, using a national population rather than regional or demographic subsets would appear to be superior. Adjusting for cost of living differences may be necessary when doing national equalization, and hence may be a consideration in using smaller regions.

3.8.2 "Zero-Sum" Versus "Guaranteed" Risk Adjustment

A key implementation issue is how payment flows among plans are calculated. One approach is "guaranteed payment" risk adjustment, in which payments to one plan are not affected by the health status of enrollees in other health plans (Dorn et al., 2017). In this system, typically the regulator specifies the overall mean payment per standardized risk enrollee, and a health plans' revenue for an enrollee is the product of this mean payment and the person's average risk score. Adjustments are also made for the number of months eligible or geographic cost factors. This guaranteed payment approach is used in US Medicare for its Medicare Advantage program and for its part D prescription drug formulas. Box 3.4 illustrates with hypothetical numbers how a fixed budget of $100 million might be divided up among four health plans using normalized risk scores and monthly eligibility counts.

BOX 3.4 Hypothetical risk equalization with guaranteed (average) payment

Health plan	Number of eligible months	Average relative risk score (RRS)	Renormalized RRS	Risk-adjusted total revenue ($)
	A	*B*	*C* = *B*/Mean of *B*	*D* = *A***C** (Mean payment)
P1	50,000	0.900	0.874	17,475,728
P2	50,000	1.100	1.068	21,359,223
P3	30,000	1.450	1.408	16,893,204
P4	120,000	0.950	0.922	44,271,845
Totals	250,000			$ 100,000,000
Means (per month)		1.030	1.000	$ 400

A second approach, as used in the Netherlands, Germany, and the US Marketplaces, is called "zero-sum" risk adjustment in that risk equalization payments sum up to zero for a specified budget across plans.[31] Conceptually, in a zero-sum system funds are reallocated from funds with low average risks or high average revenues and given to health plans with high average risk. Zero-sum payments can be made to adjust health plan payments, as is done in the Netherlands, or designed to adjust health plan revenues, as is done in the US Marketplace. The key feature of a zero-sum payment system is that if one plan has sicker enrollees and gets more equalization funds, then payments to other plans must be decreased. The hypothetical example provided in Box 3.5 illustrating how premium revenue to four health plans from the previous example (Box 3.4) might be reallocated in a zero-sum manner if premium revenue determines the size of the total payments and payments to health plans are calculated as the net differences between their risk adjusted revenue and their premium revenue. The first five columns in the two textboxes are the same. A similar approach can be used if total plan obligations rather than premium revenue determines total payments to be allocated among the four plans.

One advantage of zero-sum payment systems is that there is no need to forecast levels of revenue or total budgets before risk equalization. Zero-sum payments also insulate the regulator from financial risk. As discussed in van de Ven and Ellis (2000) and in various country and sector chapters in this book, diverse institutional arrangements do this equalization in practice using various sources of funding.

BOX 3.5 Hypothetical risk equalization with "zero-sum" payment

Health plan	Number of eligible months	Average relative risk score (RRS)	Renormalized RRS	Risk-adjusted total revenue ($)	Average premium per month ($)	Total premium revenue ($)	Net transfers into plan ($)
	A	*B*	*C* = *B*/Mean of *B*	*D* = *A***C** (mean of *E*)	*E*	*F* = *A***E*	*G* = *D* − *F*
P1	50,000	0.900	0.874	17,475,728	400	20,000,000	−2,524,272
P2	50,000	1.100	1.068	21,359,223	400	20,000,000	1,359,223
P3	30,000	1.450	1.408	16,893,204	500	15,000,000	1,893,204
P4	120,000	0.950	0.922	44,271,845	375	45,000,000	−728,155
Totals	250,000			$100,000,000		$100,000,000	0
Means (per month)		1.030	1.000		$400		

3.8.3 Accommodating Lags Between Model Estimation and Implementation

In risk adjustment payment systems implemented to date, the payment formula has been estimated using historic data and then implemented on current experience.[32] This introduces a need to consider how adjustments can be made either to the formula or to overall payments to deal with this time lag.

In the United States, there is typically a 3−5-year lag between the data used to calibrate the risk adjustment formula and the year in which payments are calculated. In the intervening years, new diagnoses or new drugs and technologies may have occurred. New diagnostic variables are added to the CMS-HCC model approximately every 2−3 years when the payment formulas are updated. The HHS-HCC risk adjustment model, originally calibrated using 2010 data when introduced in 2014, was updated for 2016 and 2017 to use a simple average of models from 2012 to 2014 data, which enabled changes in coding and cost patterns to be incorporated (CMS, 2016c).

Further challenges arise when the risk adjustment method payment uses guaranteed payment risk equalization, which is used in the US Medicare and Part D prescription drug risk adjustment programs. In this case, healthcare cost inflation needs to be estimated and used to update mean payments, and changes in the demographic or mean risk scores of enrollees is needed. Whereas a zero-sum equalization system automatically balances spending and risk score changes over time, guaranteed payment systems must forecast levels of both the mean payment per normalized enrollee as well as changes in risk scores into the future when planning payments.

Both zero-sum and guaranteed payment risk equalization require that enrollments and potentially other demographic information at the end of the payment year are available. Hence, payments to health plans are always made or at minimum adjusted after the end of the year. This is a serious challenge when using concurrent risk adjustment formulas, since it can take a number of months for claims to arrive and be fully adjudicated. To deal with this some systems make interim payments to plans, and in other cases some portion of payments is held back (in the US funds are "sequestered" pending final reconciliation). In the US Marketplaces, the 2017 sequestration rate was 7.1% of payments for risk adjustment and 6.9% of payments for the reinsurance program (US Department of Health and Human Services, 2016). Together this means that 14% of plan revenue was withheld pending final reconciliation of risk-adjusted payments and reinsurance. In the Netherlands, risk equalization is done by continuing to make zero sum adjustments to revenues for up to 3 years after the payment period (Chapter 14: Health Plan Payment in the Netherlands).

3.8.4 The Sources of Funds Used for Equalization

In many countries diverse sources of funds finance payments to health plans. Revenues can include general taxes; designated taxes; enrollee premiums (whether calculated as fixed dollar amounts, a percent of income, from an age−sex schedule, bids from health plans); cost sharing at the time that services are received from consumers, or designated ("earmarked") budgets funded through other sources such as cigarette or alcohol taxes. A key feature for risk equalization is that funds from any of these sources can be pooled and used to reallocate funds to health plans. Funds can be captured and used either to compensate for a guaranteed payment scheme, or used for zero sum reallocation.

Along with the diversity of sources of funds used for risk equalization, a variety of institutional arrangements can be used for risk equalization. Sometimes a national government agency does redistribution (e.g., the Centers for Medicare and Medicaid Services in the United States), while other times it is an autonomous agency (Germany). Van de Ven and Ellis (2000) characterize two different organizational structures for the entity that does the equalization, but there are other possibilities, including devolving responsibilities to individual states (US Medicaid), or an association of private health plans (Chile).

Newhouse (2017) raises an important issue often overlooked, bearing on whether guaranteed payment rather than zero-sum risk equalization is appropriate. In many countries, there are options outside of the risk-adjusted pool that can be chosen by consumers. In the United States this includes traditional Medicare (with a 70% market share—see Chapter 19: Medicare Advantage: Regulated Competition in the Shadow of a Public Option), and the private insurance outside of the Marketplace (Chapter 17: Health Plan Payment in US Marketplaces: Regulated Competition With a Weak Mandate), or in Germany (Chapter 11: Health Plan Payment in Germany), the private, nonstatutory insurance plans retain 10% of the market and do not participate in the insurance risk equalization. Newhouse's analysis implies that if the payment system includes corrections for adverse selection, then either a guaranteed payment structure is needed or a zero-sum payment program will need budget adjustments for plans to break even.

3.8.5 Integrating Risk Adjustment With Risk Sharing

A key theme of this volume is that risk sharing can complement risk adjustment for reducing risk selection incentives, and reducing plan level risk. Some forms of risk sharing discussed in Chapter 4, Risk Sharing, can be implemented by modification of the risk adjustment formula. The observation to make here is that the distinction between risk adjustment and risk

sharing is blurry. Furthermore, implementation of a risk adjustment formula should at least take into account the presence of any risk-sharing program so that risk adjustment adjusts for the risks that plans are actually responsible for.

3.9 CONCLUDING THOUGHTS

This chapter has attempted to provide an overview of the huge empirical literature on the estimation, selection, use, and interpretation of risk adjustment models for health plan payment. We have tried to provide abundant references for those interested in estimating risk-adjustment models. We end by speculating on a few likely directions for future research and implementation. First, better use of timing information can be made. There are a number of new estimation approaches that use hybrid risk adjustment models, in which both concurrent (year t) as well as prospective (year $t-1$) information is used to predict and determine year t payments. More broadly, using longer prior time periods for risk adjusters, and potentially using more information about the timing during the year of new information appears promising. Second, constrained regression techniques are another promising direction. The statistical and incentive properties of these new approaches are just beginning to be understood. Third, there is enormous diversity across countries in the risk adjusters and methods used. Opportunities exist for cross-fertilization and a convergence in their approaches. Fourth, new machine learning algorithms show promise for better specifying and designing risk adjustment models. Whether these approaches can satisfy the feasibility criteria that policy decision-makers seem to desire remains an open question. Fifth, to our knowledge, none of the existing risk adjustment models has fully taken advantage of the rich new diagnostic detail included in the new ICD-10 diagnosis system (only implemented in the United States in 2014) or of the rich new information contained in electronic medical records or consumer self-reported information. Sixth, and finally, researchers need to consider how to incorporate diverse social risk factors—education, income, language barriers, homelessness, and more—into risk adjustment formulas so as to improve fairness and efficiency. Better data, methods, objectives, and payment formulas lie ahead and suggest a busy future for developers of risk adjustment models.

ACKNOWLEDGMENTS

We thank Arlene Ash and Wenjia Zhu for useful input to this chapter on early drafts, and above all Tom McGuire and Richard van Kleef for their detailed and useful comments.

ENDNOTES

1. Van de Ven and Ellis (2000) call this agent the "sponsor," emphasizing that this agent is willing to take losses on some enrollees by cross-subsidizing from the gains on others.
2. These nine dimensions parallel the dimensions of risk sharing defined in Van Barneveld et al. (2001) which are discussed in Chapter 4: Risk Sharing of this volume.
3. The DxCG-HCC predictive model, (licensed by Verscend Technologies as Version 4.2), with 394 HCCs is currently used for payment by the Massachusetts Medicaid program (which covers low-income and high-health-cost individuals) for plan payment (Ash et al., 2017), and has also been used for risk-adjusted quality and performance measures (Iezzoni, 2013; Song et al., 2011; Ash and Ellis, 2012) where more disease-specific HCCs and greater predictive power are desirable.
4. For a discussion of the rationale for each principle see Kautter et al. (2014). Principles for including or imposing hierarchies on pharmacy clusters are presented in CMS (2016b).
5. see Ash et al., 2000; Pope et al., 2004.
6. As discussed in Section 3.3.6, spending for partial-year eligibles has been annualized by dividing by the fraction of the year for which their utilization is observed.
7. In our 2014 MarketScan sample, we discovered 26 people with annualized plan obligations that exceeded 1000 times the sample mean, and hence covered costs that exceeded $369,000 per month. This including one person who was in the sample costing over $26 million in less than 12 months. Only four of these individuals were eligible for all 12 months of the year. Hence dropping partial-year eligibles eliminated 85% of these extreme outliers from the estimation sample. The last line of Table 3.1 eliminated the remaining three, with a further dramatic reduction in skewness and kurtosis, but a modest effect on the mean and CV.
8. The challenge of veterans or other secondary insurance enrollees is that they may move around without being detected, and hence it is difficult to know months of eligibility in a specific region. The modeling choices are either to assume full-year eligibility in the region in which a claim is made or to assume eligibility starts only when the first claim is made in that region. The former may be preferred. Primary insurance plans generally do a better job at tracking geographic mobility, although seasonal movements still present similar problems.
9. Consider an individual that incurs $50,000 of plan obligation in the first 5 days of the year and then dies. In terms of a daily weighting, this will be a person costing an annualized $3.65 million dollars per year with a weight of 1.36%. With monthly weighting, this will be a person costing an annualized $600,000 per year with a sample weight of 8.33%. The latter observation is much less skewed and will lead to more stable estimation results.
10. Fig. 3.2 reveals a dip in spending between 63 and 64 years old for both groups, possibly reflecting an anticipatory effect of postponing treatment until covered by Medicare, or that sicker workers are more likely to retire early, improving the pool of remaining enrollees, or the effect of deductibles which make the partial year enrollees have a lower average plan payments in the final year before exiting to Medicare (Ellis et al., 2017a).
11. The distinction between claims and encounter records is that the former is used by health plans to pay providers and charge consumers, whereas encounter records may be recorded in settings that do not use fee for service reimbursement, and hence may be devoid of the financial incentives to report the same degree and quality of information. In the United States and abroad some capitated plans do not require claims, and hence only encounter records are available.
12. Revisions to ICD-9-CM introduced by the ICD-10-CM include:

    - Relevant information for ambulatory and managed care encounters, such as whether it is an initial or follow-up encounter.
    - Expanded injury codes.
    - New combination codes for diagnosis/symptoms to reduce the number of codes needed to describe a problem fully.

- Addition of sixth and seventh digit classification.
- Classification specific to laterality (right versus left side).
- Classification refinement for increased data granularity.

Existing risk adjusters, and notably the US HCC system, although allowing mappings with the new ICD-10-CM codes, have not fully taken advantage of their greater specificity and refinements in the design of their classification and prediction systems. This is impossible to do here until data on both diagnoses and spending under the new system are available.).

13. It is not hard to find a preventative drug for a high-cost health condition that is itself inexpensive, but which is predictive of higher annual spending. Paying a plan a lot for the prescription of this drug creates incentives to overprescribe it.
14. US Food and Drug Administration, National Drug Code Directory, https://www.accessdata.fda.gov/scripts/cder/ndc/.
15. see especially Iezzoni, 2013.
16. The first author of this chapter participated in unpublished exploratory work that attempted to use simple lab test results on a moderately large sample and did not find meaningful increases in predictive measures from doing so once diagnoses were used.
17. In unpublished related work removed from the manuscript for space, the authors found that inclusion of self-reported health measures and other survey information improved validation R-squared values by 1%−3% points depending on model specification.
18. The concept of retrospective risk adjustment should be reserved for models that use a base period that follows the prediction period. For example, researchers may want to study the costs of a year that includes a heart attack, a hospitalization, or a delivery, using information from a subsequent period, such as the characteristics of the cancer, infection, or newborn that ultimately resulted. Such a retrospective analysis could also be used to reward (or punish the lack of) preventive effort.
19. (Hybrid risk adjustment is also used sometimes to refer to including diverse risk adjusters that may differ in source and not just timing.
20. Glazer and McGuire (2000) coined the term "conventional risk adjustment," which they characterize as having the goal of paying providers as close as possible to the amount the enrollee is expected to cost. Conventional risk adjustment is a statistical and data-oriented approach that is often characterized as trying to maximize the fit of the predictive model. In this chapter we use traditional risk adjustment to reflect the attention to selection incentives and coding accuracy, which lead to the imposition of constraints that intentionally sacrifice predictive power to improve incentives and fairness.
21. Sufficient assumptions so that maximizing the R-squared achieves the social optimum are that plans can discriminate at the individual level, and that there are no other plan payment features such as premiums and risk-sharing that can affect revenue (Layton et al., 2017).
22. Fuller et al. (2016) advocates for mutually exclusive categories.
23. To illustrate with one concrete example, a categorical rate cell approach, if it includes a rare condition such as HIV/AIDS, it will generally not be able to distinguish the additional costs of adding further conditions to individuals in the HIV/AIDS rate cell, while an additive approach is able to make predictions that take into account not only other common conditions, but even other rare ones among the HIV/AIDS patients.
24. Top-coding has been evaluated in research but is rarely adopted for payment models. See the two SOA reports (Winkelman and Mehmud, 2007; Hileman and Steele, 2016) for extensive analysis for the commercial setting.
25. It might seem that a correction for the bias from top-coding might be desired, such as to multiply all spending by a constant so as to maintain the same sample mean. Once it is remembered that the purpose of estimating any risk adjustment model is to come up with RRS, then this bias is immediately rectified once its predicted value, whether Y or YTC, is divided by its mean.

26. Negative values for spending can occur in the United States when a health plan reconciliation reduces the payment to a provider in the year following the original claim. Or it can occur when a claim reconciliation is incorrectly attributed to the wrong patient, or coverage for a service in the previous year is denied and the consumer pays the plan for a service previously paid for by the plan. There is no easy way to correct these negative payments just using claims data. As described in Pope et al. (2004), the US Medicare Advantage risk adjustment program leaves observed negative values unchanged in case they are correlated with specific health conditions, so that resetting spending to zero could introduce a biased payment for these conditions.
27. We also tested a model that predicts untop-coded spending, but uses the results from estimating the top-coded model. However, this model did not improve our results in any statistical measure.
28. Age groups for the DxCG model are defined as [0, 1], [2, 4], [5, 9], [10, 14], [15, 20], [21, 24], [25, 29], [30, 34], [35, 39], [40, 44], [45, 49], [50, 54], [55, 59], [60, 64].
29. Consider the following example from the Clinical Classification Software (CCS) system created by the Agency for HealthCare Quality and Research (AHRQ, 2017), which has the great advantage of being open source software. As of 2017, the CCS classification system allows different degrees of fineness, including 285 mutually exclusive diagnostic categories. But the CCS system does not propose any suggested hierarchies among CCS categories. Consider for example two single-level diagnostic categories: CCS 98 (Essential hypertension) and CCS 99 (Hypertension with complications and secondary hypertension). Here 99 is clearly a more serious manifestation of 98, but 98 will commonly be coded along with 99 on different claims. Although a modeler can include flags for both 98 and 99 and their interaction (i.e., three terms) in a model to be considered, it may be preferable to include instead only two flags: one for CCS 99 and a flag for (CCS 98 but not 99). This saves a degree of freedom, improves clinical coherence, reduces overfitting, and reduces the incentive for upcoding.
30. For a brief economist-accessible description of penalized regressions for prediction, see Kleinberg et al., 2015.
31. The budget to which the risk equalization is applied needs not be the total budget of the health plans. In the Netherlands, for instance, health plans can charge an additional premium to enrollees. These funds are not included in the zero-sum budget that is then allocated across plans. The payments are still zero-sum in the sense that if one plan has a higher risk score from coding more disease, its revenue increases by decreasing the payments to other plans.
32. In theory, the principles for estimating the payment model could be specified and the concurrent risk adjustment formula could be estimated even after the utilization and claims were observed. This has been done in some pay-for-performance systems, such as is described in Vats et al. (2013) for one health plan in Albany New York. High-quality data and speedy action would be needed, along with tolerance for delayed payments.