September 27, 2017

Risk Adjustment for Health Plan Payment

Randall P. Ellis, Bruno Martins and Sherri Rose[1]

Chapter 3 in Thomas G. McGuire and Richard van Kleef (eds.) *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice.* Elsevier Press.

Summary

Risk adjustment for health plan payment is a bundled payment strategy in which payments are based on mathematical formulas that predict health plan obligation for spending on each enrollee. The risk adjustment formula quantifies the relationship between spending and explanatory variables called risk adjusters. Selection of risk adjusters involves tradeoffs between cost saving and selection incentives. Exogenous variables like age and sex that are independent of spending choices and utilization maximize cost savings incentives but leave strong incentives for selection, while endogenous, more gameable variables that reflect utilization and taste choices may better reduce selection incentives but weaken cost control. Estimation requires selecting a population, the types of spending to be adjusted, an objective function and choosing a statistical approach for estimation, for which machine learning may be appropriate. Implementation of risk adjustment formulas requires a choice of enrollee data, adjustments to deal with time lags, accommodation of other risk sharing and premium features, and specifying how funds are to be reallocated from the sources of funds selected.

[1] Acknowledgement: We thank Arlene Ash and Wenjia Zhu useful input to the paper on an early draft, and above all Tom McGuire and Richard van Kleef for their extraordinarily detailed and useful comments on the chapter.

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 health care 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 one-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, called here the regulator[2], which could be a government, independent agency or employer, willing to reallocate payments to plans based on the expected costs of 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

[2] 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. A profit-maximizing agent will typically not be willing to do this.

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 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. More than ten-fold differences in premiums can emerge based on age and gender alone, with much larger differences possible if health status or other information is allowed for premium setting. Elsewhere in this volume, we explore risk sharing, in which payments reflect combinations of actual costs, lump sum payments, and premiums. The key insight motivating risk adjustment is that risk-adjusted lump sum payments will usually improve selection incentives better than a fixed lump sum payment, and will also reduce the desired premium differentials plans want to charge enrollees. Risk adjustment can never be perfect, but it can still be a useful tool for mitigating selection while maintaining cost containment incentives.

Van Barneveld et al (2001) provide a useful framework for thinking about four key dimensions of risk sharing and health plan payment, which is discussed more fully in the next chapter. Risk adjustment is more complex, since the payment formula need to be estimated in advance, and then typically applied to a different sample at a later date. We amend the four dimensions used for risk sharing into nine dimensions used for risk adjustment shown in Textbox 3.1, and organize our chapter around discussing these dimensions after first discussing the broad goals of risk adjustment in the next section.

Textbox 3.1: Nine dimensions of risk adjustment

Risk adjustment model calibration

1. The sample on whom the risk adjustment model is to be calibrated (e.g. the entire population, or specific subsets of the population)
2. The types of services on 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 (socio-demographic, diagnostic, pharmacy or other information).
4. The timing of the information to be used for predicting annual spending (eg., 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 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 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 newly estimated 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 to 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 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 US 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 health care services.

3.2.1.1 Avoid endogenous signals

A central concern when selecting risk adjusters is that they are not 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 not highly predictive. Variables such as counts of visits or dollars of health care spending for an enrollee are much more predictive, but also more endogenous variables. Diagnoses and pharmaceutical use are also endogenous, although

[3] The DxCG-HCC predictive model, (licensed by Verscend Technologies as Version 4.2), with 484 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.

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 so as to increase payments are less clear. For example, in systems using zero sum risk adjusted payments, upcoding of diagnoses reallocates funds to plans that are more successful at upcoding, but does not increase total spending.

Papers that document or quantify the degree of endogeneity in the US 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 how serious this problem is. Chapters 11 and 14 discuss recent evidence 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, for which Glazer and McGuire (2000) propose a theoretical correction. Variables such as homelessness, income, race/ethnicity, and needing long term care services are examples of risk adjusters that can be predictive, but difficult to verify. Unfortunately, few variables that predict health care 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. Figure 3.1 documents that among the commercially insured in the US; there is a remarkable amount of year-to-year variation in the prevalence of specific chronic conditions. This suggests both noisy coding and the potential for upcoding through greater plan effort.

3.2.1.3 Avoiding incentives to undertreat

A related but slightly different issue for choosing risk adjusters is to avoid incentives for health plans to avoid treating certain conditions because it will only reduce revenues in the future. A common complaint about risk adjustment is that it reduces the incentive for providers to cure expensive conditions since this will result in reduced income the following year (Pope et al, 2004).

3.2.1.4 Maintaining "power"

The efficiency issue that has received the greatest attention by developers of risk adjustors is to maintain incentive 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 that measure the power of risk-adjusted payments. Geruso and McGuire explore how risk adjusted payments are tied costs by simulating the impact of eliminating random spells of treatment and observing how the missing use affects plan revenue. In a non-capitated, cost-based framework, eliminating expenditures will have a power of zero, since both revenue and costs will be reduced equally. 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. They 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 also a prime consideration in discussions of risk sharing in Chapter 4.

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 level of payments

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, Martins and Miller, 2016). This can be the result of either the overall payment rate being too generous or the capitated payment for an individual risk group 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.

3.2.2 Fairness

Although many economists focus solely on efficiency incentives, 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. As illustrated elsewhere in this volume, regulators can desire that – ceteris paribus – these contributions should be equal for people with diabetes and those without, but should be higher in regions with high-supplier induced demand than in regions with low supplier-induced demand. 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 easily be in conflict with fairness of payments across income, geography, or other socioeconomic variables like education and race. In the US, fairness considerations commonly guide the choice of risk adjusters to use in payment formulas (Ash et al., 2000). In Europe, fairness is commonly spoken about in terms of "solidarity" across income or health when discussing risk adjustment (Chinitz et al., 1998; van de Ven and Ellis, 2000; Van Kleef et al., 2009).

The European concepts of solidarity have sometimes been used more formally in *a priori* sense to identify variables that should or should not be used for prediction. For example, Belgium (Chapter 7) classifies risk adjusters into two types: S and N variables. S-type variables are solidarity variables, capturing information for which the regulator wants to insure that plans are neutral in this dimension, and which the regulator is willing to include in the risk adjustment payment formula. Age, gender and health status are S variables. The N variables are non-solidarity variables where the grounds for inclusion in the risk adjustment formula can potentially be challenged. For example, should the risk adjustment formula adjust for obesity, skiing accidents, or heroin addiction, all of which may reflect lifestyle choices? Consumers and policy makers may differ in their judgment on these. Incorporating fairness objectives into risk adjustment models is discussed further below in our discussion of modeling objectives.

3.2.1 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. For example, in the early 2000s, German physicians resisted providing diagnoses on claims until it was clear that their unwillingness would mean that only hospital diagnoses would be used for risk adjustment. In a similar way, the prospect of using office-based diagnoses for risk adjustment in the US for Medicare Advantage risk adjustment led to a remarkable improvement in diagnostic coding practice in the early 2000s when the change was phased in. 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 US to calibrate risk adjustment models in the commercial sector that span different insurers. Switching from commercial to Medicaid or Medicare health insurance creates similar data issues. Feasible risk adjustment in the US 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. Incompatibilities across different electronic medical record systems in the US also make it infeasible to develop universal medical record or medical chart based models, although this could change if suitable regulations were changed.

Policymakers often feel that a simpler system is more feasible to implement. This is revealed in the early efforts in the US Medicare and Germany 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 to this day used in Switzerland (Chapter 16) and elsewhere) in which each person is uniquely assigned to one rate cell, and the mean cost of people in that cell is 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 all other cell predictions. 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 they are constrained by sample sizes to remain simple and have weaker predictive power than additive models.

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 US have become more complex, with separate models for different population groups, and increased numbers and variety of risk adjusters. Should machine learning algorithms be used to develop risk adjustment models in the future, as discussed below, this may potentially greatly increase model complexity, in which the details of the classification system and payment formula is not easily identified by health plans and providers. Rose (2016) has argued that this complexity may be desirable by making health plan selection more challenging.

3.2.2 Ten principles in Pope et al. (2004)

Textbox 3.2 summarizes the ten principles that guided the creation of the diagnostic classification system of the first HCC system for Medicare Advantage, and have remained important 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 they can be used by modelers without having to return to clinicians, statisticians, and policymakers as frequently for guidance.[4]

[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).

Textbox 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: First ten principles are from Pope et al, 2004.

Principle 1 seems obvious but is often violated with 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. 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 list, but reflect recent insights into risk adjustment discussed below: deisgners 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 Textbox 3.1.

3.3 Choice of estimation sample

The first decision in any every 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 the case. The US Medicare Advantage model (Chapter 19) continues to use the traditional Medicare enrollee sample, not Medicare Advantage enrollee data, for calibrating its risk adjustment formula more than 30 years after first adopting risk adjustment. The US Marketplace (Chapter 17) 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 now uses a national sample.

A number of empirical studies have shown that risk adjustment formulas developed on one sample are 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 single US formula over six years and seven plan types. Ellis et al. (2013) found that an HCC formula calibrated using US data had predictive power nearly strong as a simplified HCC formula 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 in which troublesome records have been eliminated. 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, dropping extreme outliers, or estimating on subgroups such as by eliminating 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 is comparable across samples. 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 huge 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.[5] For diseases like chronic heart failure and pancreatic cancer, only including people who survive for an additional twelve calendar months in the estimation sample generates a very biased subset of these populations. Our conclusion is that if the size of the sample allows it, including partial year eligibles in the prediction year should be done when estimating risk adjustment models.

3.3.2 Separate formulas for special populations

It is relatively common to estimate separate regression models for distinct subpopulations as a means of ensuring a level playing field across diverse groups. The 2017 CMS Medicare

[5] In our 2014 MarketScan sample, we discovered 26 people with annualized plan obligations that exceeded 1000 times the sample mean, and hence average monthly 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 modest effect on the mean and CV.

Advantage model uses nine different formulas for different subpopulations (Chapter 19). 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 nine 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 US Marketplace (Chapter 17), Netherlands (Chapter 14), and other countries also use separately calibrated models. The Swiss have separate risk adjustment formulas within each canton (similar to a county in the US), by age, gender and according to whether people are hospitalized or not. Since they use a 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).Risk sharing is used to share costs across cantons.

Estimating separate models for different subgroups is probably a good idea if sample sizes are adequate, and there is strong evidence that cost patterns are distinctly different. Germany, despite having an enormous sample size, uses a single risk equalization formula for the full population, although the formula does includes age-specific HCC terms that allow it to better predict certain age groups (Chapter 11). Estimating a single formula, but including dummy variables - 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) 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 benefit design. 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 not by modeling separate subsamples of the estimation sample population, but rather by simulating different degrees of benefit coverage, since a wide range of actuarial value plans are offered. Note the Marketplace plans use a separate new enrollee model even though its use of a concurrent framework does not necessitate this.

Adjusting payments for differences in benefit design is clearly a good idea for predicting means correctly, although empirically the risk scores from formulas estimated separately on different benefit plans are highly correlated. A significant concern, about which there is relatively little research, is how to incorporate behavioral response to the benefit design differences across plans, which can greatly change spending and hence plan obligations.

3.4 The types of services for which spending is to be predicted

The correct dependent variable is the one capturing the spending that is to be an obligation of each health plan for each of their enrollees, which implies calculating both the services covered and plan obligations after deducting enrollee cost sharing payments. In the US, 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). The Netherlands has separate formulas for subsets of spending rather than subsets of the population. Their main model covers somatic health (medical plus pharmaceutical spending, excluding certain specified categories) that encompasses about 80 percent of total health care spending. 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).

While estimating separate formulas for distinct services may seem attractive to policymakers (called "carve outs" in the US), the danger is that it may encourage inappropriate substitution between different services. Outpatient prescription drugs can be an alternative to medical or surgical therapies (as well as substitute for drugs provided while in the hospital), and short-and long-term mental health services are clearly substitutes. If payments for these services come out of different bundled payments, providers may have an incentive to change care patterns to take advantage of these different payment flows. Carve outs are also frequently funded from a different source than the main formula, adding budgetary complexity. Monitoring and avoiding double counting bills may also be challenging.

3.4.1 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 U.S. is pharmaceutical spending where formularies may include or exclude a wide number of drugs, or include or exclude specific providers. 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 US, while there is meaningful heterogeneity in what services are covered or not covered in Medicare Advantage, prescription drug plans and the Marketplace.

Sometimes, payment formulas are adjusted when there are new technologies or costs that can be 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. Some, but not all, Medicaid and private risk adjustment model users built these additional costs into their risk adjustment formulas, but this was complicated because not all plans chose to cover this expensive drug in full.

Empirical data show that total paid and covered amounts are extremely highly correlated ($\rho = .998$ in our US MarketScan data, whether topcoded at $250k or not) so the differences in risk scores at the aggregate for a given sample is relatively small in these data 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 small and there is little risk sharing by the regulator, the differences in relative risk scores 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 other countries.

3.4.2 Actual versus annualized spending

For research studies, researchers often choose to focus on the cleanest sample, which usually means samples in which everyone is enrolled for all twelve months in a calendar year. For payment purposes, this is not desirable, since partial year enrollments are nonrandom, and have different patterns of disease. This is illustrated in Table 3.1 above. Births, deaths, retirement, and changing jobs and health plans are all correlated with specific disease and levels of health spending, and hence if partial year enrollees are dropped, or this issue is ignored, then serious biases can result.

Since Ellis and Ash (1995), the preferred method for estimating linear risk adjustment models has been to annualize spending and then weight 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 one-to-ten year olds. Suppose that on average in their first year, newborns cost $6000. Unlike most one-to-ten year olds, babies are on average only eligible for coverage for about half of the year. Therefore their average monthly cost should be $1,000 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 US 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 US.

Using unweighted spending can be preferred when health plan eligibility data is 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.[6] With very intermittent use of the benefit, perhaps only every few years, assigning individuals to a geographic region or provider group can be problematic.

Adjustments for partial year enrollment are done differently in various countries. The US and Switzerland use monthly eligibility to annualize spending and perform equalization. Germany and the Netherlands use the fraction 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 in modest sample sizes. The second issue is how premiums and plan revenue payments are paid. In the US, most employers and the government pay health plans a monthly premium for each enrollee, even when an enrollee is only eligible for

[6] 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.

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.

3.4.3 Normalizations to create relative risk scores

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

3.5 Information used for predicting spending (risk adjusters)

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

3.5.1 Age and gender

The classic risk adjusters are age and gender. Figure 3.2 illustrates the one-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 U.S. in 2014 by one-year age intervals for males and females, 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.[7]

Figure 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 five- or ten-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 one-year increments. As long as the overall sample size is large, in the tens of millions as we advocate, then overfitting is minimal and the age gender pattern is accurately without risk of overfitting.

3.5.2 Diagnoses on submitted claims or encounter records

Both in the US and internationally, diagnoses on claims or encounter records[8] submitted by providers is the preferred set of information for risk adjustment, currently in use in the US, Germany, 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

[7] Figure 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, Martins and Zhu, 2017a).

[8] 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 US and abroad some capitated plans do not require claims, and hence only encounter records are available.

much more predictive than simply age and sex. As shown in Table 3.2, the R2 for prospective and concurrent models are 15.3 and 41.5 percent, versus only 1.5 percent for age-sex alone in US commercial data. Improvements in RMSE and 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 is topcoded at $250,000 where we also see that the confidence bands are reduced to close to a zero range. Goodness of fit measures and the value of using out of sample measures are discussed in Chapter 7.

One key issue with diagnoses is that the quality of diagnoses recorded varies across providers and settings, with inpatient diagnoses generally viewed as more accurate than office-based diagnoses. Sometimes non-clinicians (e.g., home health workers or massage therapists) may report diagnoses on claims or encounters, which in the US and in most countries are not officially recognized in plan risk adjustment payments (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 68,000 legal codes in the ICD-10-CM versus 14,000 ICD-9-CM codes) are collapsed into a limited number of categories.

It is worth mentioning that in most years, the World Health Organization (WHO) makes changes in the ICD diagnoses. Most of these ICD changes are just in the 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 (proposed for 2017). Many countries do not adopt the WHO ICD codes immediately or without modification. The US (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 US 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 system consistent with their own coding system.[9]

3.5.3 Pharmacy information

Pharmacy information can also be used for risk adjustment, and was adopted for use in the Netherlands in 2002, even before the introduction of diagnostic information there in 2004. Typically, individual drugs, as identified by their National Drug Code (NDC) in the US, are mapped into categories of drugs, and selections of these broader categories based on chemical entities are then incorporated in risk adjustment models. In 2017, the risk adjustment system in the Netherlands (the Dutch system) is using 33 pharmacy-based cost groups for risk adjustment in addition to diagnostic cost groups and diverse other measures including seven multiple-year high cost spending measures (van Kleef et al., 2017, this volume, Chapter 14). The US is not

[9] Revisions to ICD-9-CM introduced by the ICD-10-CM include:

- Relevant information for ambulatory and managed care encounter, 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.

currently using pharmaceutical information for payment in its risk adjustment system, although there was a proposal to do so for the Marketplace (CMS, 2016).

There are several challenges with using pharmaceutical information for prediction in risk adjustment. One challenge is the large number of different drugs, because the NDC defines a particular drug as combination of distributor, formulation and package type. The US Food and Drug (FDA) administration maintains a directory of allowed drugs with extensive information 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.[10] The Europeans uses a different classification system called the Anatomical Therapeutic Chemical (ATC) to group together drugs, and WHO updates their system only twice per year, however it still has a very large number of categories for mapping.

Even more challenging is that prescription practices and the 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 US 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 in this category until it was recalibrated.

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,

[10] US Food and Drug Administration, National Drug Code Directory, https://www.accessdata.fda.gov/scripts/cder/ndc/

some anti-hypertensive 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. Pharmaceutical information is currently being used to augment diagnostic information in the Germany and the Netherlands (Chapters 11 and 14) and has been proposed for use in Switzerland and in US Medicare Advantage program (Chapters 16 and 19 in this volume).

3.5.4 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^2 of 9.08%. This predictive power can be improved to 14.40% by topcoding spending used on the *right hand side* at $250,000 and further improved to 21.41% by topcoded both the dependent and right hand side variables at this level. The coefficient on the lagged spending variable in this last model is .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 have a power of .50. 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, Jiang and Kuo (2013) 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^2 increased from 10 to 15%). Such models would probably never 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) 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, if allowed.

Although lagged spending is not used directly as a risk adjuster, as described in Chapter 14 of this volume (Van Kleef et al. 2017), Dutch risk adjustment model 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. 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 that already includes diagnoses- and pharmacy-based risk adjusters. Moreover, the Dutch risk adjustment model temporarily includes risk classes based on prior-year spending for two specific services, i.e. home care and geriatric rehabilitation care.

Chapter 4 of this volume discusses the innovation of van Kleef et al. (2015) of using the risk adjustment formula to incorporate supply side cost sharing or reinsurance, via using functions of current year spending as a right hand side variable together with constrained regressions.

3.5.5 Health care 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 (DME), while Switzerland uses a dummy variable for whether or not a person has been hospitalized in the prior year. Moreover, diagnosis based models implicitly are

rewarding for at least one claim with a diagnosis as are pharmacy based models. It is difficult to assess the incentive effects of prior utilization on cost containment incentives, but certainly, they have some effect at reducing 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 shown.

3.5.6 Medical record information

Ever since medical records started becoming computerized there has been a desire to utilize this information for improved risk adjustment (Parkes, 2015). The focus of this chapter is prediction of health care 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[11], and for reconciling and buttressing claims submissions that affect plan payments. There is an active industry in the US advising plans on how to more carefully capture diagnoses, so as to increase plan revenue, but similar efforts to use this information to build even more refined 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 US have not be sufficiently standardized that they can be used across different information systems or merged into a standardized format. At the system level, they are difficult to work with. Second, both

[11] See especially Iezzoni (2013).

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 collected in biased settings. One rarely gets a measure at random times, but instead it tends 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 US or government health systems in Europe and Australia have used medical record information.[12]

3.5.7 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, bias, and predictive power. Feasibility relates to the high cost of surveys relative to using diagnoses from submitted claims, bias relates to the challenges of getting adequate and representative response rates, and predictive power has been the primary subject of various research studies. 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

[12] The first author of this article participated in unpublished exploratory work that attempted to use simple lab test results on a moderately large sample and did not find meaningfully large predictive measures from doing so once diagnoses are used.

activities of daily living (ADL) and instrumental activities of daily living (IADL). 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, 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 to make useful statements for population-based uses. Van de Ven and Ellis (2000) report fit measures (R^2) 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. (2013) report results using a cross section from New South Wales, Australia on 267,188 individuals over a four-year panel data set, yielding a panel size of 787k person-years. The encouraging thing they report is that self-reported measures perform well over even 2 years before or after the survey was taken. The less encouraging news is that adding survey information in the form of 76 responses capturing each of the dimensions discussed above achieved an R^2 of only 10.2%, which was lower than those achieved by coarse diagnostic, pharmacy, or lagged utilization models. Survey results only added .8 percentage 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, although nontrivial variation remained.[13]

3.5.8 Other socio-economic 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 is often not 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 hospital and primary care. Gravelle et al. (2011) demonstrated that diagnosis based risk adjusters largely eliminated the usefulness of most of the demand and supply side variable 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 US to better recognize "social risk factors" is currently mandated by Congress (US Department of HHS, 2016).

A key challenge in using certain socioeconomic variables like race, language, income, or education, is that they may not be attractive to include in a payment model. One possibility is that it may simply be politically infeasible for policymakers to include these variables (e.g., race) directly. More commonly, if discrimination is a problem, a subgroup (say a minority or non-native language group) may currently receive too little health care. A regression-based model without any adjustment will tend to perpetuate this inequity, paying less for this subgroup

[13] 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^2 by 1-3 percentage points depending on model specification.

because it better predicts current spending. The usual solution is to simply omit this information from the predictive model, which makes this underpayment less visible, but does not necessarily change the model tendency to recreate the existing inequity. A better solution proposed by Layton et al. (2016) and further developed in Bergquist et al (2017) is to modify the data so as to compensate for the under provision problem before estimating the risk adjustment model. Ash and Ellis (2012) did something similar in developing a primary care payment model, while Layton et al. (2016) develop a more formal framework for doing so.

A different approach is used by Ash et al. (2017) who explore alternative ways of incorporating socioeconomic information while estimating individual level risk adjustment models for the US Medicaid enrollees in the state of Massachusetts. Using a relatively large sample (N > 800k 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) explore how to collapse a large number of highly correlated socioeconomic measures using principle components. They identified eight variables primarily related to income, which they collapsed into the first principle component, and two variables related to homelessness and frequent changes in mailing address, which they include separately in their regression model. They find that the geography-based variables add information to the model and identify variation that would be meaningful to providers who receive capitated revenue, but the overall impression is that many of the variables identified were of marginal statistical significance.

In Europe, sociodemographic variables are commonly used in risk adjustment models. The Dutch risk adjustment models includes risk adjusters based on household income, household size and employment status (see Chapter 14 of this volume for more details). Similar types of information are used in Belgium (see Chapter 7 of this volume). Though these risk adjusters do generally not lead to substantial increases in R^2, they can redistribute large amounts of money (e.g. from plans with relatively many self-employed to plans with many unemployed), which can be important to achieve a level-playing field for insurers (see Chapter 7 for an extensive discussion of this point). Textbox 3.3 discusses a refinement of how solidarity versus non solidarity risk factors have been considered in the Netherlands.

Textbox 3.3 Adjusting for S versus N (Solidarity versus Non-solidarity) risk factors

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 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 payment. Things are more complicated in case these two types of risk factors are correlated (Schokkaert et al. 2017). An example of such 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 (from a prior year) 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. Schokkaert and Van de Voorde (2004) have proposed to determine the subsidy in two steps. In the first step an expenditure model is estimated with both the S-factors and N-factors included as explanatory variables. In the second step subsidies are calculated using the expenditure model from the first step with the weights for N-factors being neutralized (i.e. put on the population average for all). Van Kleef et al. (2008) and Stam et al. (2010) follow a different approach and modify the spending data on which the S-factor weights are estimated. Whereas Schokkaert and Van de Voorde (2004) regress observed spending on the relevant factors (in their first step), Van Kleef et al. (2008) and Stam et al. (2010) first correct observed spending for N-type variation. After this correction, a simple regression of "corrected" spending on S-factors provides unbiased weights.

3.6 Choice of time frame for data used for prediction

In specifying the different types of information to be used for risk adjustment we have not yet specified the time period over which that information is observed and used for prediction. The time interval over which risk adjustors 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; Kauter, 2014). Several alternatives for choosing the base and prediction periods are possible.

3.6.1 Prospective versus concurrent use of risk adjusters

Two broad frameworks are commonly used for risk adjustment. Prospective models use a base period that precedes and does not overlap with the prediction period. The second common approach is a concurrent model, using 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, sometimes called retrospective risk adjustment, require that one must wait until the end of the year to observe all of the information used for prediction.[14]

The majority of risk adjustment formulas used around the world today are prospective rather than concurrent, although a growing number are using both types of information. Prospective models have more power (Geruso and McGuire, 2016) than concurrent models, and less prone to endogenous signals, since diagnoses for acute conditions that are treated and resolved within one-year matter little for prospective models. On the other hand, prospective models complicate implementation because they create a need to have a separate formula to use

[14] Ash and Ellis reserve the concept of retrospective for models that use a base period that follows the prediction period. For example, one could want to study the costs of an event, such as 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. This would be a retrospective analysis.

with newly arriving enrollees, for whom prior year information is not 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 arrives for payment one 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 to their revenue as new payments come in, even though the same plans readily accept cost uncertainty. Concurrent risk adjustment is particularly attractive when turnover is high, as it is with Medicaid and Marketplace enrollees, so prior year information is commonly missing.

3.6.2 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. Dudley et al. (2003) were perhaps the first to examine such a framework, and introduce the terminology of "hybrid risk adjustment".[15] 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 percent of the population, and used a concurrent framework to pay for this subsample of the population while paying for the remaining 90.7 percent of the population using a prospective HCC framework. Their pioneering early work achieved an R^2 of 26% versus a prospective R^2 of only 8%. Further research in this direction was

[15] Hybrid risk adjustment is also used sometimes to refer to including diverse risk adjusters, that may differ in source and not just timing.

conducted by García-Goñi et al. (2009) to predict drug expenditures using Spanish data with similar gains in predictive power.

The Netherlands is currently using a rich set of risk adjusters that differ in multiple dimensions in the source of information used. Prospective diagnoses and pharmacy information are used, dummies for multiple prior years of spending help identify enrollees with longer-term chronic conditions, while concurrent information about DME and long-term care are used. See Chapter 14 for more details.

Before ending this discussion of hybrid prospective/concurrent risk adjustment, it is worth mentioning that any payment system that is using *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 adjustor and fixing the coefficient at the desired share (making it equivalent to reinsurance) is inherently a hybrid framework. Chapter 4 contains a more detailed discussion on this point.

3.7 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. This section reviews the key concepts relevant to objective functions, and how they are incorporated in risk adjustment model design and selection.

3.7.1 Conventional risk adjustment

The classic approach to risk adjustment, commonly called "conventional risk adjustment," has emphasized accuracy in matching plan revenue to predictable spending for individual persons and policy-relevant subgroups. A commonly stated objective is to "level the playing field" so that health plans (the agents who contract pay providers for health care services) do not gain from attracting profitable enrollees, nor lose from attracting unprofitable

ones (Ash et al., 1989; Pope et al., 2000). Risk adjustment changes health plan profit incentives by paying more for enrollees predicted to cost more and less for enrollees predicted to cost less. Conventional risk adjustment calculates the predicted cost of each individual, and pays the health plans the expected cost of each enrollee, while minimizing the unexplained variation in spending or equivalently, maximizing the model fit. Traditionally, consumer premium contributions, risk sharing, and regulatory constraints are not incorporated directly in the conventional risk adjustment objective function, although issues such as gaming, incentives, and fairness are invariably considered in the process of risk adjustor selection and model design. Although diverse objective functions are often considered, the overwhelming favorite objective function of conventional 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.

3.7.2 Optimal risk adjustment

Recent economic models of risk selection (Glazer and McGuire, 2000; Layton et al., 2017) have argued that conventional risk adjustment – focusing on explaining as much of the variance as possible - will in general not solve efficiency problems related to selection except under strong and implausible assumptions.[16] Glazer and McGuire (2000) show that simply maximizing the fit of a model can still lead to serious 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

[16] Sufficient assumptions so that maximizing the R^2 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).

"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 under-supply (over-supply) certain services. Building on this insight Ellis and McGuire (2007), and more recently McGuire et al. (2014) and Ellis, Martins and Zhu (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 5.

3.7.3 Statistical objective functions.

By far the most common metric of model performance is the squared prediction errors at the individual level. Almost all papers report the conventional R^2, which calculates this squared error and then normalizes it by the total variance and hence captures the proportion of the overall variance in the dependent variable explained by the model (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. We follow Ash et al., (1989, 2000) and report the R^2 as a percentage rather than a ratio. The R^2 can be calculated as

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

where f_i is the prediction for observation i and y_i is the actual value, and $\bar{y}$ is the sample mean ofy_i. Note that the R^2 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^2 measures using our test sample for three alternative dependent variables and four alternative sets of right hand side variables, which we discuss further below.

3.7.4 Categorical versus additive models

Since the origins of risk adjustment in the 1980s, two different frameworks have been advocated: Categorical models place each individual uniquely in a single cell, and reduce the prediction problem to simply calculating the mean for that cell. Categorical models are used in Switzerland, and Colombia, as well as in 3M's Clinically Related Groups (CRG) system in the US. Cell-based approaches are equivalent to a regression with many indicator variables. A regression approach is more flexible in terms of the form in which regressors can be entered, including allowing for interactions. The essential difference in modeling approach is whether predictions are additive in the explanatory factors or fundamentally mutually exclusive, as with a branching structure. Advocates still exist for both approaches: Fuller et al. (2016) advocates for the mutually exclusive categories. Many machine learning algorithms are based on cells.

So far, in head-to-head comparisons of models by independent researchers on large samples, such as those 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^2, 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, which confirm 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 comparisons below, some of which also use a categorical rather than additive framework.

3.7.5 Functional form

3.7.5.1 Transformations of the dependent variable

All risk adjustment performance measures can be influenced 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 topcode the dependent variable at some level such as $250,000.[17] Hence, if Y is total spending, the transformed dependent variable Y^{TC} is the minimum of actual spending or $250,000. This has the effect of minimizing the effect 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.

[17] Topcoding 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.

Topcoding, which retains the risk adjusters that may have resulted in the very high levels of payments is preferred to dropping high cost observations altogether. Alternative values for topcoding ranging from $50,000 to $1 million are also sometimes used.[18]

A second, more dramatic transformation is to use natural logarithms. Since annual health spending is often zero[19], it is common to add one to spending before taking logs. If negative values of Y occur, these must also be eliminated by resetting to one. Hence the natural log of Y, LnY, would typically be calculated as

$$LnY = Ln(\max(1, Y + 1))$$

Tables 3.2, 3.3, and 3.4 above present R^2, 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: untopcoded spending, $250k topcoded spending, and natural log of spending.[20] The R2 shown here are 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

[18] It might seem that a correction for the bias from topcoding 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 relative risk scores, then this bias is immediately rectified once its predicted value, whether Y or Y^{TC}, is divided by its mean.

[19] This typically happens in the US 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. Often there is no easy way correct these negative payments, at least in US claims systems. The US Medicare Advantage risk adjustment program leaves them in as negative amounts rather than resetting them to zero for fear of introducing bias and an increase of total payments.

[20] We also tested a model that predicts untopcoded spending, but uses the results from estimating the topcoded model. However, this model did not improve our results in any statistical measure.

predict concurrent spending, while the third row adds 394 DxCG-HCCs variables to the concurrent model. The final row in each table shows a prospective model results, using the same specification as for the concurrent model. Among the two age sex models, the first one uses 28 Age-Sex groups, as used by the Affordable Care Act Marketplace risk adjustment model[21] while the second row uses 130 age-sex dummies, with sex interacted with one-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 Figure 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 untopcoded spending. 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, topcoding 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 we will discuss in Chapter 4, outlier policies such as reinsurance deal with the same concerns about outliers as topcodings. 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^2 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

[21] 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].

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., 2013).

Transforming the dependent variable also has implications for the precision of the goodness of fit measures. In fact, the confidence intervals around the R^2 for the untopcoded model 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 even removing them completely from the risk adjustment model – not only increases the R^2 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 untopcoded spending might lead to misleading results, as the estimated R^2 is sensitive to the particular draw of observations. In this sense, topcoding the dependent variable before the analysis is a more robust approach to compare different models.

3.7.5.2 Constraints and hierarchies

Even after grouping diagnoses into a manageable number of discrete categories, there are 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 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, however even they

are in danger of sacrificing 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 are 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 hierarchies, which are 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 Hierarchical Condition Category (HCC) that underlies the risk adjustment models used in the US 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 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 the only 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 thirty body systems for prediction.

Consider the following extended example. Assume there are two costly 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 properties 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 are additive: the incremental cost of a broken arm or an allergy diagnosis is about the same regardless of a large number of coexisting conditions.

Another second possibility is that conditions A and B complicate each other. Diabetes, cancer, immune disorders, heart conditions, pregnancy, and liver disorders, for instance, tend to complicate the treatment and hence the cost of other diseases. 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 terms is desirable. Indeed, the risk adjustment models used in the US 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 broad sets of many such conditions.

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, Conditions 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.

The frequency of this third possibility can be greatly inflated in a sample because of imprecise diagnostic coding. Physicians will often have a choice of 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 reimbursing 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 sloppy coding like this, then a regression model will find 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 sloppy or only the more serious manifestation matters, imposing hierarchies can use this knowledge to 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 $D_{B\sim A}$. which is what imposing a hierarchy does: only recognizes B when not accompanied by A. In effect, the hierarchies provide a clinically motivated 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. Without the hierarchies, 57 additional regressors would need to be added to the model to achieve roughly the same statistical results.

Imposing constraints by excluding some variables from the payment model is another way to allow clinical or policy-motivated criteria to help select the preferred risk adjustment model. The 2017 CMS-HCC model started with 189 HCCs, of which 110 were excluded from the payment model, before estimating using the included set of HCCs. Hierarchies were then imposed on these 110 HCCs. Although Kautter et al. (2012) provides a valuable overview of the final CMS-HCC model chosen, details of the process used for its selection are not provided. The ten principles shown in Textbox 3.2 above 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 having omitted HCCs. Principle 6 - not to

include coding proliferation, is another important reason why some HCCs are omitted. In a very small number of cases, constraints are also imposed to insure monotonicity, so that payments do not go down when a new diagnosis is recorded.

The ability to use *a prio*ri 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 US and Germany. A second and equally important rational 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 condition (lung cancer). Purely additive models, without hierarchies, will tend to keep increasing predictions as more (less serious) conditions are reported.[22]

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 to only recognize the insulin prescription when the diagnosis has not been recorded, which is a form of

[22] 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 the 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, adds clinical coherence, reduces overfitting, and reduces the incentive for upcoding.

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.

3.7.6 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 (2017). They demonstrate the value of constrained regression models to simultaneously balance model estimation, with achievement of various social 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 corrected by using constrained least squares regression techniques. If the conventional risk adjustment formula allocates too little money for people receiving home care services, then this can be corrected by imposing constraints while calculating risk adjustment parameters in order to eliminate this inefficiency. (Note that the problem here is reallocating money to an underfunded subset of the sample, not adding additional funds for some subset, discussed below.) They use a large sample of Dutch enrollees to show proof of concept in which underpayment for both physiotherapy and home health care 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 people using these two types of services exactly match total spending on those people. Constraints will change the payments for other groups as well. As Van Kleef et al show, a number of other previously underpaid groups have payments increased with the introduction of the constraint on home care underspending. Funding for other groups is also affected, on average being reduced by enough to fund the increased spending desired.

Constrained regressions can be used for other purposes 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 ten health care services.

Constrained regression risk adjustment is attractive conceptually, and deserving of further research. For practical implementation, it remains to be seen whether objectives embodied in the constraint are acceptable, whether the models are sufficiently understandable and whether the effects on other groups in aggregate is acceptable. Almost all of the other issues discussed in the chapter remain important even with a constrained regression framework.

3.7.7 Machine learning methods

Machine learning algorithms provide automated tools to data-adaptively learn about the relationships between variables. This is 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. This is where the 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 hurdles 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, while making the assumption that each observation is independent and drawn from the same underlying multivariate distribution.

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 conventional risk adjustment, called the general L_2 loss function:

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

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 small to modest.

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. It would also be a formal way to enforce topcoding in the loss function directly. 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.[23] For linear regressions, the function to minimize can be characterized in its simplest form by:

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

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

[23] For a brief economist-accessible description of penalized regressions for prediction, see Kleinberger et al. (2015).

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 health care 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 conventional non-tree-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^2 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^2 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 number of terminal nodes, observations per terminal node, trees, or covariates allowed for each tree.

Random forests is 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 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 performance of 14 algorithm implementations to the super learner based on a validation R^2, 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. 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^2 of 25% for the full set vs 23% for the reduced set). 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^2. 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, the time has not yet come where machine learning will replace more conventional risk adjustment models for payment purposes. Notably, machine learning methods may result in prediction functions that clinicians and policymakers find unintuitive or hard to explain, although this lack of transparency could be advantageous to prevent strategic responses to the risk adjustment formula, such as by "upcoding" diagnoses or under-supplying 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 issues related to issues related to the implementation of risk adjustment formulas, which is commonly called risk equalization. Risk equalization involves choosing the set of 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 will want to read the individual country/sector chapters in Part II of this volume. Here we try to touch on a broad overview of challenges and selected solutions.

3.8.1 The population groups for which risk is to be equalized

In Textbox 3.1 we note that in addition to choosing the sample on which to estimate the formulas, one must also define the population set to whom the formula is to be applied, which need not be the same. In the US, 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. Similarly, risk adjustment in Switzerland is done at a Canton level, although risk sharing is done at the national level.

The choice of whether to do regional areas, demographic subsets, or a single region for risk equalization is often driven by political considerations. From a risk perspective, and for fairness objectives, using a national population 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 persons average risk score. Adjustments are also made for 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. Textbox 3.5 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.

Textbox 3.5:Hypothetical risk equalization with guaranteed (average) payment

Health Plan	Number of eligible months	Average relative risk score (RRS)	Re-normalized 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 Netherlands, Germany and the US Marketplace, is called "zero sum" risk adjustment in that risk equalization payments sum up to zero across plans. 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 zero sum payment 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 Textbox 3.6 illustrating how premium revenue to four health plans from the previous example (Textbox 3.5) 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 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, as is done in the Netherlands.

Textbox 3.6:Hypothetical risk equalization with "zero sum" payment

Health Plan	Number of eligible months	Average relative risk score (RRS)	Re-normalized 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	$400	$400	

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.

3.8.3 Accommodating lags between model estimation and implementation

In every risk adjustment payment system used so far, the payment formula has been estimated using historic data and the implemented on more recent data.[24] 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 US, 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. The CMS-HCC model adds in new diagnoses to existing payment categories approximately every two years to three when the payment formulas are updated to help keep the HCCs better up to date. The HHS-HCC risk adjustment model has been updated for 2017 using 2014 data, which enabled changes in coding and cost patterns to be incorporated.

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 health care cost inflation needs to be estimated and used to

[24] 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, Ash, and Ellis (2013) for one health plan in Albany New York. High quality data and speedy action would be needed, along with tolerance for delayed payments.

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 is always made or at least adjusted after the end of the year. This is a particular challenge with concurrent risk adjustment formulas, since it can take months for claims to arrive and to be validated. 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 percent of payments for risk adjustment and 6.9 percent of payments for the reinsurance program (US Department of Health and Human Services, 2016). Together this means that 14 percent of plan revenue was withheld pending final reconciliation of risk adjusted payments and reinsurance.

Newhouse (2017) raises an important issue often overlooked which partially distinguishes when 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 US this includes traditional Medicare (with a 70% market share – see Chapter 19), and the private insurance outside of the Marketplace (Chapter 17), or in Germany (Chapter 11), the private, nonstatutory insurance plans retains 10 percent of the market and does not participate in the insurance risk equalization. Newhouse highlights that if the payment system

includes corrections for adverse selection, then a zero sum payment program will not suffice, and the plan payments will need additional revenue to allow plans to break even.

3.8.4 The sources of funds used for equalization

In many countries a diverse set of sources of funds are used to generate revenue for 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 US), 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).

3.8.5 Integrating risk adjustment with risk sharing

A key theme of this volume is that risk sharing is an important alternative to risk adjustment for reducing risk selection incentives, and reducing plan level risk. Using constrained regressions to achieve fairness or efficiency goals is another opportunity for modifying conventional risk adjustment approaches. Several of the strategies for risk sharing discussed in

the next chapter can also be achieved with payment modifications to the risk adjustment formula. The observation to make here is that the separation between risk adjustment and risk sharing is becoming blurry, and any implementation of a risk adjustment formula should automatically take into account the risk sharing program that is also planned. Moreover, when multiple payment strategies are being taken, it is valuable to consider the full set of all systems when calibrating and implementing risk adjustment. It can often be appropriate to adjust payments, the risk adjustment formula, or consumer premiums to take into account risk sharing or reinsurance so as to avoid double counting certain types of expenditures. Too often this has not been done in the past.

As a simple example, if 7% of total payments are designated to be used for risk sharing, then this amount should be deducted from total payments that are equalized through risk adjustment, whether using a zero sum or guaranteed payment structure. This should be done both for risk adjustment estimation and for implementation.

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 the interested future risk adjusters. We wanted to 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 prospectively (year t-1) information is used to predict and determine year t payments. Constrained regression techniques are another exciting new direction for hybrid models. The statistical and incentive properties of these new approaches are just beginning to be understood. 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, there is enormous diversity across countries in the risk adjusters and methods used. Opportunities exist for cross-fertilization and a convergence in their approaches. Third, new machine learning algorithms show great progress 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. Fourth, to our knowledge, none of the existing risk adjustment models have fully taken advantage of the rich new diagnostic detail included in the new ICD-10 diagnosis system, (only implemented in the US in 2014) or of the rich new information contained in electronic medical records or consumer self-reported information. Fifth and finally there is a new emphasis on incorporating 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.

References

Agency for HealthCare Policy and Research (AHRQ), 2017. Clinical Classifications Software (CCS) for ICD-9-CM Fact Sheet. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp

Ash, A.S., Ellis, R.P., 2012. Risk-adjusted Payment and Performance Assessment for Primary Care. *Medical Care* 50 (8): 643-653.

Ash, A.S., Ellis, R.P., Pope, G., Ayanian, J., Bates, D., Burstin, H., Iezzoni, L., Mckay, E., Qu, W., 2000. Using Diagnoses to Describe Populations and Predicts Costs. *Health Care Financial Review* 21 (3), 7–28.

Ash, A.S., Mick, E., Ellis, R.P., Kiefe, C., Clark, M., 2017. Adding Social Determinants of Health Factors to Medically-Based Risk Adjustment Improves Risk Equalization Payment

in a US Low-Income Population. Working paper, University of Massachusetts Medical School, Worcester, MA

Ash, A.S., Porell, F., Gruenberg, L., Sawitz, E., Beiser, A., 1989. Adjusting Medicare Capitation Payments Using Prior Hospitalization Data. *Health Care Financing Review* 10 (4), 17-29.

Bergquist, S.L., Layton, T.J., McGuire T.G., Rose S. 2017. Intervening on the Data to Improve the Performance of Health Plan Payment Methods. Working paper, Department of Health Care Policy, Harvard Medical School, September 12, 2017.

Breiman, L., 1996. Stacked Regressions. Machine Learning 24 (1), 49–64.

Buchner, F., Wasem, J., Schillo, S., 2017. Regression Trees Identify Relevant Interactions: Can This Improve the Predictive Performance of Risk Adjustment? *Health Economics* 26 (1), 74-85.

Centers for Medicare &Medicaid Services, 2016. Patient Protection and Affordable Care Act: Benefit and Payment Parameters for 2018. 45 CFR Parts 144, 146, 147, 148, et al. September 6, 2016. https://www.gpo.gov/fdsys/pkg/FR-2016-09-06/pdf/2016-20896.pdf

Chinitz, D., Preker, A., Wasem, J., 1998. Balancing competition and solidarity in health care financing. In: Saltman, R.B., Figueras, J., Sakellarides, C. (Eds.), *Critical Challenges for Health Care Reform in Europe.* Open University Press, Buckingham, U.K., 55-77.

Cid, C., Ellis, R.P., Vargas, V., Wasem, J., Prieto, L., 2016. Global Risk-Adjusted Payment Models. In: Scheffler, R. (Ed.), *Handbook of Global Health Economics and Public Policy* 11 (1), 311-362.

Dixon, J., Smith, P., Gravelle, H.S.E., Martin, S., Bardsley, M., Rice, N., Georghiou, T., Dusheiko, M.A., Billings, J., Lorenzo, M., Sanderson, C., 2011. A Person Based Formula for Allocating Commissioning funds to General Practices in England: Development of a Statistical Model. *BMJ* 343 (nov22 1), pp.d6608-d6608.

Dorn,S., Garrett, B., Marks, J., Holtz-Eakin, D., Holt, C., Book, R., and O'Neill Hayes, T. 2017. Stabilizing the Individual Market: Risk Adjustment and Risk Mitigation. June 28, Downloaded 9/25/2017 from https://www.americanactionforum.org/research/stabilizing-individual-market-risk-adjustment-risk-mitigation/.

Drozd, E., Cromwell, J., Gage, B., Maier, J., Greenwald, L., Goldman, H., 2006. Patient Casemix Classification for Medicare Psychiatric Prospective Payment. *American Journal of Psychiatry* 163 (4), 724–32.

Dudley, R.A., Medlin, C.A., Hammann, L.B., Cisternas, M.G., Brand, R., Rennie, D.J., Luft, H.S., 2003. The Best of Both Worlds? The Potential of Hybrid Prospective/Concurrent Risk Adjustment. *Medical Care* 41 (1), 56-69.

Dunn, D.L., Rosenblatt, A., Tiaira, D.A., Latimer, E., Bertko, J., Stoiber, T., Braun, P., Busch, S., 1996. *A Comparative Analysis of Methods of Health Risk Assessment. Final Report to the Society of Actuaries*. SOA Monograph M-HB96-1.

Einav, L., Finkelstein, A., 2011. Selection in Insurance Markets: Theory and Empirics in Pictures. *Journal of Economic Perspectives* 25 (1), 115-138.

Ellis, R.P., Ash, A.S., 1995. Refinements to the Diagnostic Cost Group (DCG) Model. *Inquiry* 32 (4), 418-429.

Ellis, R.P., Fiebig, D.G., Johar, M., Jones, G., Savage, E., 2013. Explaining Health Care Expenditure Variation: Large Sample Evidence Using Linked Survey and Health Administrative Data. *Health Economics* 22 (9), 1093–1110.

Ellis, R.P., Jiang, S., Kuo, T., 2013. Does Service-Level Spending Show Evidence of Selection Across Health Plan Types? *Applied Economics* 45 (13), 1701-1712.

Ellis, R.P., Martins, B., Miller, M.M., 2016. Provider Payment Methods and Incentives. In: Heggenhougen, H.K., Quah, S. (Eds), *International Encyclopedia of Public Health*, Second Edition.

Ellis, R.P., Martins, B., Zhu, W., 2017a. Health care demand elasticities by type of service. *Journal of Health Economics*

Ellis, R.P., Martins, B., Zhu, W., 2017b. Demand elasticities and service selection incentives among competing private health plans. *Journal of Health Economics*.

Ellis, R.P., McGuire, T.G., 2007. Predictability and Predictiveness in Health Care Spending. Journal of Health Economics 26 (1), 25–48.

Friedman, J., Hastie, T., Tibshirani, R., 2001. *The Elements of Statistical Learning*. New York: Springer.

Fuller, R.L., Averill, R.F., Muldoon, J.H., Hughes, J.S., 2016. Comparison of the Properties of Regression and Categorical Risk-Adjustment Models. *Journal of Ambulatory Care Management* 39 (2), 157-65.

García-Goñi, M., Ibern, P., Inoriza, J.M., 2009. Hybrid Risk Adjustment for Pharmaceutical Benefits. *European Journal of Health Economics* 10 (3), 299–308.

Geruso, M., Layton, T., 2015. Upcoding: Evidence from Medicare On Squishy Risk Adjustment. NBER Working Paper 21222.

Geruso, M., McGuire, T.G., 2016. Tradeoffs in The Design of Health Plan Payment Systems: Fit, Power and Balance. *Journal of Health Economics* 47, 1-19.

Glazer, J., McGuire, T.G., 2000. Optimal Risk Adjustment in Markets with Adverse Selection: An Application to Managed Care. *American Economic Review* 90 (4), 1055-1071.

Glazer, J., McGuire, T.G., 2002. Setting Health Plan Premiums to Ensure Efficient Quality in Health Care: Minimum Variance Optimal Risk Adjustment. *Journal of Public Economics* 84 (2), 153-173.

Gravelle, H., Dusheiko, M., Martin, S., Smith, P., Rice, N., Dixon, J., 2011. Modeling Individual Patient Hospital Expenditures for General Practice Budgets. CHE Research Paper 73.

Gruber, S., van der Laan, M., 2010. A Targeted Maximum Likelihood Estimator of a Causal Effect on a Bounded Continuous Outcome. *The International Journal of Biostatistics* 6 (1), A26.

Hileman, G., Steele, S., 2016. *Accuracy of Claims-Based Risk Scoring Models*. Society of Actuaries.

Iezzoni, L.I. (Ed), 2013. *Risk Adjustment for Measuring Healthcare Outcomes,* Fourth edition. Ann Arbor, Michigan: Health Administration Press.

Jones, A.M., 2011. Models for Health Care. In: Clements, M.P., Hendry, D.F. (Eds.), Oxford *Handbook of Economic Forecasting*. Oxford University Press, New York.

Kautter, J., Ingber, M., Pope, G.C., Freeman, S., 2012. Improvements in Medicare Part D Risk Adjustment: Beneficiary Access and Payment Accuracy. *Medical Care* 50 (12), 1102-1108.

Kautter, J., Pope, G.C., Ingber, M., Freeman, S., Patterson, L., Cohen, M., Keenan, P., 2014. The HHS-HCC Risk Adjustment Model for Individual and Small Group Markets Under the Affordable Care Act. *Medicare & Medicaid Research Review* 4 (3), E1-E11.

Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction Policy Problems. *American Economic Review: Papers & Proceedings* 105 (5), 491-495.

Laffont, J-J., Tirole, J., 1993. *A Theory of Incentives in Procurement and Regulation*, vol. 1, 1 ed., The MIT Press.

Layton, T.J., Ellis, R.P., McGuire, T.G., and Van Kleef, R.C. 2017 (forthcoming). Measuring Efficiency of Health Plan Payment Systems in Managed Competition Health Insurance Markets. *Journal of Health Economics*.

Layton, T.J., McGuire, T.G., 2017. Marketplace Plan Payment Options for Dealing with High-Cost Enrollees. American Journal of Health Economics. 3(2): 165-191.

Layton, T.J., McGuire, T.G., Van Kleef, R.C., 2016. Deriving Risk Adjustment Payment Weights to Maximize Efficiency of Health Insurance Markets. NBER Working Paper 22642.

McGuire, T.G., Newhouse. J.P., Normand S.L., Shi, J., Zuvekas, S., 2014. Assessing Incentives for Service-Level Selection in Private Health Insurance Exchanges. *Journal of Health Economics* 35, 47-63.

McCullagh, P., 1983. Quasi-likelihood Functions. *The Annals of Statistics* 11, 59–67.

Medicare Payment Advisory Commission, 1998. *Report to the Congress: Medicare Payment Policy*. Volume 2.

Newhouse, J.P., 1996. Reimbursing Health Plans and Health Providers: Efficiency in Production Versus Selection. *Journal of Economic Literature* 34 (3), 1236–1263.

Newhouse, J.P., 2017. Risk Adjustment with an Outside Option, *Journal of Health Economics*.

Newhouse, J.P., Buntin, M.B., Chapman, J.D., 1999. *Risk Adjustment and Medicare*. The Commonwealth Fund.

Parkes, S., 2015. Producing Actionable Insights from Predictive Models Built Upon Condensed Electronic Medical Records. *Health Watch*.

Pope, G.C., Ellis, R.P., Ash, A.S., Ayanian, J.Z., Bates, D.W., Burstin, H., Iezzoni, L.I., Marcantonio, E., Wu, B., 2000. Diagnostic Cost Group Hierarchical Condition Category Models for Medicare Risk Adjustment, Final Report. Health Care Financing Administration, Contract No. 500-95-048.

Pope, G.C., Kautter, J., Ellis, R.P., Ash, A.S., Ayanian, J.Z., Ingber, M.J., Levy, J.M., Robst, J., 2004. Risk Adjustment Of Medicare Capitation Payments Using The CMS-HCC Model *Health Care Financing Review* 25 (4), 119-141.

Relles, D., Ridgeway, G., Carter, G., 2002. Data Mining and the Implementation of a Prospective Payment System for Inpatient Rehabilitation. *Health Services and Outcomes Research Methodology* 3 (3), 247–266.

Rose, S., 2016. A Machine Learning Framework for Plan Payment Risk Adjustment. *Health Services Research* 51 (6), 2358–2374.

Rose, S., Shi, J., McGuire, T., Normand, S.L., 2015. Matching and imputation methods for risk adjustment in the health insurance Marketplaces. *Statistics in Biosciences*, Advance online publication. doi:10.1007/s12561-015-9135-7.

Rose, S., Zaslavsky, A.M., McWilliams, J.M., 2016. Variation in Accountable Care Organization Spending and Sensitivity to Risk Adjustment: Implications for Benchmarking. *Health Affairs* 35 (3), 440-448.

Schokkaert, E., Steel, J., Van de Voorde, C.2017. Out-of-Pocket Payments and Subjective Unmet Need of Healthcare. *Applied Health Economics and Health Policy* 15 (5): 545-555.

Schokkaert, E., Van de Voorde, C. 2004. Risk Selection and the Specification of the Conventional Risk Adjustment Formula. *Journal of Health Economics* 23, 1237–1259.

Shrestha, A., Bergquist, S., Montz, E., Rose, S., 2017. Mental Health Spending Risk Adjustment Using Clinical Categories and Machine Learning. Working Paper.

Song Z. S., Safran, D.G., Landon, B.E., He,Y., Ellis, R.P., Mechanic, R.E., Day, M.P., and Chernew, M.E. (2011) "Health Care Spending and Quality in Year 1 of the Alternative Quality Contract" *New England Journal of Medicine*. July 13, 2011 (10.1056/NEJMsa1101416).

Stam, P. J., van Vliet, R. C., & van de Ven, W. P. (2010). Diagnostic, pharmacy-based, and self-reported health measures in risk equalization models. *Medical care*, 448-457

U.S. Department of Health and Human Services, 2016. Patient Protection and Affordable Care Act; HHS Notice of Benefit and Payment Parameters for 2018. *Federal Register* 81 (17), 61456- 61535.

U.S. Department of Health & Human Services, Office of the Assistant Secretary for Planning and Evaluation. *Report to Congress: Social Risk Factors and Performance Under*

Medicare's Value-Based Purchasing Programs. https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs. Washington, DC: 2016. Accessed May 3, 2017.

Van Barneveld, E.M., Lamers, L.M., R.C.J.A. van Vliet and W.P.M.M van de Ven, (2001) Risk Sharing as a Supplement to Imperfect Capitation: A Tradeoff Between Selection and Efficiency, *Journal of Health Economics*, 20(2): 147-168.

Van der Laan, M.J., Polley, E., Hubbard, A., 2007. Super Learner. *Statistical Applications in Genetics and Molecular Biology* 6 (1), A25.

Van Kleef, R.C., Beck, K., Van de Ven, W.P.M.M., Van Vliet, R.C.J.A., 2008. Risk Equalization and Voluntary Deductibles: A Complex Interaction. *Journal of Health Economics* 27 (2): 427-443.

Van Kleef, R.C., Eijkenaar, F., Van Vliet, R.C.J.A., Van de Ven, W.P.M.M., 2017. Health Plan Payment in the Netherlands. In: McGuire, T.G., Van Kleef, R.C. (Eds.), *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets*: Theory and Practice.

Van Kleef, R.C., Van Vliet, R.C.J.A., 2012. Improving Risk Equalization Using Multiple-Year High Cost as a Health Indicator. *Medical Care* 50 (2), 140-144.

Van Kleef, R.C., McGuire, T.G., Van Vliet, R.C.J.A., Van De Ven, W.P.M.M., 2016. Improving Risk Equalization with Constrained Regression. *The European Journal of Health Economics* 1-20.

Van Kleef, R.C., Van de Ven, W.P.M.M., Van Vliet, R.C.J.A., 2009. Shifted deductibles for high risks: more effective in reducing moral hazard than traditional deductibles. *Journal of Health Economics*, 28, 198-209.

Van de Ven, W.P.M.M., Ellis, R.P., 2000. Risk Adjustment in Competitive Health Plan Markets. In: Culyer, A., Newhouse, J. (Eds.), *Handbook of Health Economics*. Amsterdam: North Holland: 755–845.

Van Veen, S.H.C.M., Van Kleef, R.C., Van De Ven, W.P.M.M., Van Vliet, R.C.J.A., 2017. Exploring the Predictive Power of Interaction Terms in a Sophisticated Risk Equalization Model using Regression Trees. *Health Economics*, forthcoming

Van Veen, S.H.C.M., Van Kleef, R.C., Van De Ven, W.P.M.M., Van Vliet, R.C.J.A., 2015a. Improving The Prediction Model Used in Risk Equalization: Cost and Diagnostic

Information from Multiple Prior Years. *European Journal of Health Economics* 16 (2), 201-218.

Van Veen, S.H.C.M., Van Kleef, R.C., Van De Ven, W.P.M.M., Van Vliet, R.C.J.A., 2015b. Is There One Measure of Fit That Fits All? A Taxonomy and Review of Measures-Of-Fit for Risk-Equalization Models. *Medical Care Research and Review* 72 (2), 220-243.

Vats, Sonal, Ash, A.S. Ellis, R.P. (2013) "Bending the Cost Curve? Results from a Comprehensive Primary Care Payment Pilot." *Medical Care.* 51(11):964-969, November 2013.

Ware, J.E., Sherbourne, C.D., 1992. The MOS 36 item short form health survey (SF-36). *Medical Care* 30 (6), 473–483A.

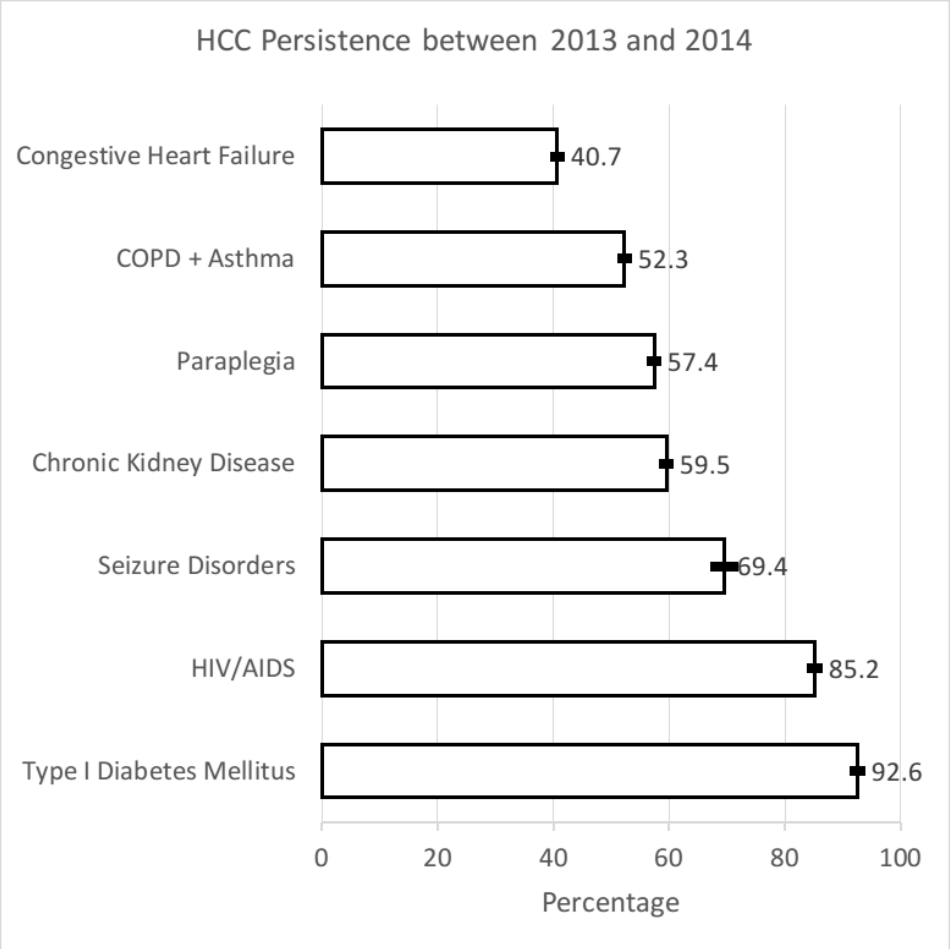
Wedderburn, R., 1974. Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika* 61, 439–447.

Wennberg, J.E., Staiger, D.O., Sharp, S.M., Gottlieb, D.J., Bevan, G., Mcpherson, K., Welch, H.G., 2013. Observational Intensity Bias Associated With Illness Adjustment: Cross Sectional Analysis Of Insurance Claims. *BMJ* 346, F549.

Winkelman, R., Mehmud, S., 2007. *A Comparative Analysis Of Claims-Based Tools For Health Risk Assessment.* Society Of Actuaries.

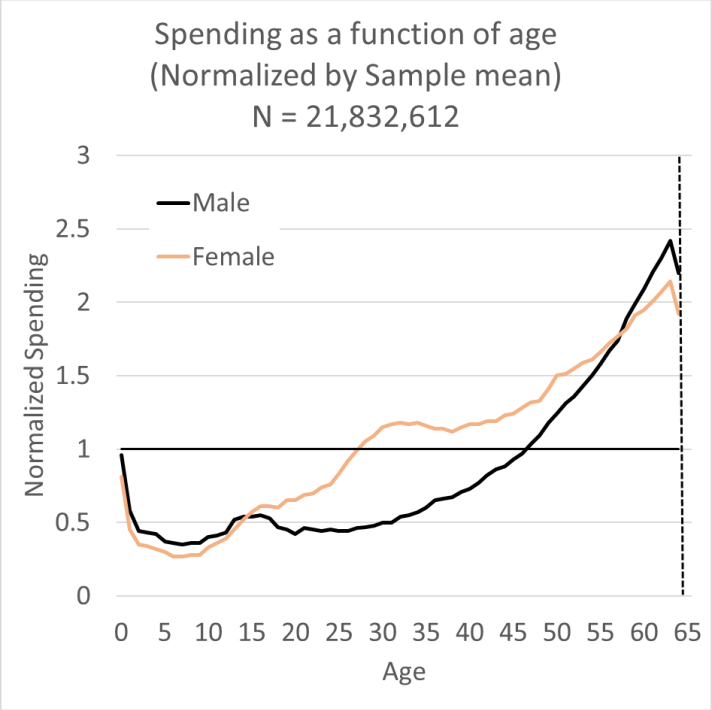
Wolpert, D.H., 1992. Stacked Generalization. *Neural Networks* 5 (2), 241–259.

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.

Figure 3.2. – US normalized spending by age and sex



Note: Figure shows normalized spending by one-year age increments, for males and females, aged 0 to 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.

Table 3.1 Alternative estimation sample summary statistics on 2014 plan payments per enrollee

	Number of Obs.	Mean Spending	C.V.	Skewness	Kurtosis
Full Sample	21,832,612	4,429	1,660	184.85	219,008.77
Removed if less than 12 months eligible in 2014	18,041,199	4,322	1,521	36.36	5,061.04
Above, plus removed if less than 12 months eligible in 2013	15,710,699	4,416	1,507	35.80	5,135.17
Above, plus removed if aged 0 to 21	10,894,520	5,473	1,322	29.11	4,070.94
Above, plus removed if spending more than 1000 times mean	10,894,517	5,471	1,305	21.34	1,177.09

Notes: 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.

Table 3.2. – Risk-adjustment model results - R^2 (in percentages) with 95% confidence intervals

Model	Untopcoded Spending	Spending Topcoded at $250.000	Log(1 + Spending)
Within Sample Results			
Concurrent			
Age-Sex (Marketplace age groups)	1.4	2.9	10.68
	(1.2, 1.6)	(2.9, 3.0)	(10.6, 10.7)
Age-Sex (1 year increment)	1.5	3.0	11.0
	(1.2, 1.7)	(3.0, 3.0)	(11.0, 11.1)
DxCG-HCC with Age-Sex (Marketplace age groups)	41.5	57.9	59.1
	(35.2, 46.5)	(57.8, 58.0)	(59.1, 59.2)
Prospective			
DxCG-HCC with Age-Sex (Marketplace age groups)	15.3	23.2	29.7
	(12.9, 17.3)	(23.1, 23.3)	(29.7, 29.7)

Note: Each cell provides the 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 Markeplace age groups or one year age increments. DxCG-HCC refers to DxCG 394 Hierarchical Conditional Categories. 95% confidence intervals (in parentheses) are based on 500 bootstraps.

Table 3.3. – Two alternative risk-adjustment model measures of fit

	Root Mean Squared Error		Mean Absolute Errors	
Model	Untopcoded Spending	Spending Topcoded at $250.000	Untopcoded Spending	Spending Topcoded at $250.000
Concurrent				
Age-Sex (Marketplace age groups)	14.7	9.8	1.20	1.13
	(13.8, 16.1)	(9.8, 9.8)	(1.19, 1.20)	(1.13, 1.13)
Age-Sex (1 year increment)	14.7	9.8	1.20	1.13
	(13.8, 16.1)	(9.8, 9.8)	(1.19, 1.20)	(1.13, 1.13)
DxCG-HCC with Age-Sex (Marketplace age groups)	11.3	6.5	0.71	0.65
	(10.2, 13.0)	(6.4, 6.5)	(0.71, 0.72)	(0.65, 0.65)
Prospective				
DxCG-HCC with Age-Sex (Marketplace age groups)	13.6	8.7	1.00	0.93
	(12.7, 15.1)	(8.7, 8.7)	(0.99, 1.00)	(0.93, 0.94)

Note: Each cells 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 Markeplace age groups or one year age increments. DxCG-HCC refers to DxCG 394 Hierarchical Conditional Categories. 95% confidence intervals (reported in parentheses) are based on 500 bootstraps.