February 2, 2018

Chapter 5: Evaluating the Performance of Health Plan Payment Systems

Timothy J. Layton, Randall P. Ellis, Thomas G. McGuire and Richard C. van Kleef

Abstract

Health plan payment systems serve important objectives of social policy, including encouraging the efficiency of the health insurance and health care markets, containing health care costs and promoting individual affordability of health insurance coverage. This chapter elaborates on the meaning of efficiency in the context of regulated competition and reviews the methods and measures by which a plan payment system can be evaluated. We discuss the strengths and weaknesses of measures commonly used for evaluation, such as measures of statistical fit of risk adjustment and over- and undercompensation for selected particular population groups. We propose improvements in some methods and propose specific alternative practical measures. We explain the circumstances in which newer measures are preferred to the traditional metrics.

5.1 Introduction

Social objectives for health plan payment systems include efficiency and fairness, each with multiple dimensions. Efficiency is concerned with matching the form of insurance to consumer preferences, and encouraging provision of efficient health care. Fairness has to do with individual affordability of health insurance and health care, access to high quality providers, and with the distribution of the burden of financing health insurance.[1] This chapter deals primarily with efficiency, although fairness implications are noted as they arise. We explain the nature of efficiency goals and then review methods for evaluating a plan payment system against these goals. We look for evaluation metrics that satisfy two criteria. First, measures should be *valid*, that is, be linked to an objective of the health plan payment system. Second, measures should be *practical*, that is, feasible to construct with the data typically available to researchers charged with design of payment system methods.[2]

Other objectives for health plan payment systems are covered elsewhere in this volume. Fairness and access concerns are discussed extensively in Chapter 2. Risk adjustment is often a crucial part of health plan payment. Some criteria for evaluation specific to the risk adjustment component of plan payment, such as that the risk adjustment scheme should not be "gameable," are covered in Chapter 3.

As in the rest of this Volume, the institutional setting for our discussion of methods for evaluation is regulated competition. Individuals choose their health insurance plan from among

[1] Some aspects of "fairness" are related to efficiency. Specifically, redistribution from healthy to sick consumers provides implicit insurance against the financial consequences of shifting from a healthy to a sick state, which can be welfare enhancing (Handel, Hendel, and Whinston 2015).
[2] For a technical presentation of some of the ideas in this chapter, see Layton, Ellis, McGuire and Van Kleef (2017).

a set of competing insurers offering products subject to premium and benefit regulation. Regulation notwithstanding, plans may have the ability to discourage/encourage membership by, among other things, distorting some elements of their coverage and services. We assume throughout that regulation includes open enrollment provisions, which, though perhaps working imperfectly, require that health plans accept all applicants.

Our primary perspective is at the market design phase: with data on patterns of utilization representative of the population to be covered, researchers and regulators need to assess how well a payment system – meaning the set of policies regulating both the premium structure and the plan payment scheme – will achieve social objectives of efficiency and fairness. The market-design phase is when most evaluations of plan payment methods take place. Statistical analysis and simulations prior to putting a payment system in place are the primary way regulators evaluate and decide on payment systems for the U.S. state-based Marketplaces, Medicare's payment system for private health plans, and plan payment systems in the Netherlands, Switzerland, Israel Germany and elsewhere.[3]

The (*ex ante*) market-design phase contrasts with the post-market-performance phase commonly studied in the empirical literature in economics, where econometric methods are used *ex post* to study the impact of payment system changes. While this form of research obviously feeds into choice of plan payment system, evaluations of changes in complex health care systems (such as, for example, U.S. Medicaid or Marketplace expansions) usually cannot identify the causal effects of distinct design components of a health plan payment system, necessarily relying on *ad hoc* model calibrations, and therefore fall short of answering questions critical to

[3] See as examples, Kautter et al., (2014) on U.S. Marketplaces; Pope, et al., (2011) on U.S. Medicare; Shmueli, et al. (2010) on Israel; Beck, Trottman and Zweifel (2010) on Switzerland; Breyer, Heineck and Lorenz (2003) on Germany; Van Kleef, Van Vliet and Van de Ven (2013) on the Netherlands.

regulators.[4] We include some discussion of these *ex post* evaluation studies as we go, using them to substantiate that the selection-related distortions payment systems are designed to combat actually play out in insurance markets.

5.1.1 Efficiency Problems in Individual Health Insurance Markets

Individual health insurance markets are vulnerable to economic inefficiencies caused by adverse selection, the tendency of sicker, higher-cost consumers to choose more generous coverage. This natural pattern of demand causes two central problems: 1) equilibrium premiums reflect selection as well as coverage differences, leading to pricing distortions that cause consumers to choose the "wrong" plans (Einav and Finkelstein, 2011), and 2) insurers distort the coverage of their health plans, or take other discriminatory actions, to make them less attractive to unprofitable (typically sicker) enrollees (Glazer and McGuire, 2000). The relative importance of these two forms of inefficiency varies across regulated competition markets. In the U.S. Medicare program, sorting of beneficiaries between the private managed care plans (Medicare Advantage plans) and traditional Medicare has received the most attention (see Chapter 19), whereas in the national health insurance system in the Netherlands with common regulation and coverage for the entire population, underprovision of some services (e.g. exclusion of high-quality doctors or health care facilities from provider networks) is the larger concern (see Chapter 14). Other markets, such as the Marketplaces established in the U.S. as part of the Affordable Care Act (ACA) (Chapter 17) feature both concerns: inducing participation among those eligible to purchase coverage on the Marketplace (Newhouse, 2017) and ensuring that plans provide adequate coverage for all conditions (Shepard, 2016; Geruso, Layton, and Prinz, 2017).

[4] Exceptions occur when there is variation in program implementation geographically, over time, or across eligible populations that may enable the impact of specific reform features to be identified.

For many years, motivated by concerns with adverse selection, studies of and reports on health plan payment methods have focused on the R-squared from a risk adjustment regression as the main metric of health plan payment system performance. Some papers and official reports also include ratio or difference measures of over/under compensation for specific groups. In the U.S., researchers tend to use the ratio of predicted costs to actual costs for selected groups in the population ("predictive ratios"), such as those with a chronic illness, whereas in Europe researchers tend to use the difference between projected revenues and costs ("over- and undercompensation"). Typically, in calibration of risk adjusted payments, R-squared is given primacy. The statistical regression procedure maximizes R-squared, and then under/overcompensation for various groups is checked to see if it is satisfactory. One goal of this chapter is to explain when and what modifications of these measures are called for to assess the efficiency consequences of health plan payment systems.

The health plan payment system in all countries and sectors is also expected to help with the moral hazard – or cost control – problem in health care: the tendency of providers and patients to decide on "too much" health care when the patient is close to fully insured and does not bear the full cost of the care she receives. The health plan payment system should pay health plans so as to give them incentives to discourage overutilization of health care, where overutilization is defined as care for which the cost exceeds the value consumers place on it. In discussions of regulated competition, beginning with Enthoven, this objective motivated the idea of paying plans "prospectively," that is, independent of the quantity of health care an individual uses during the current year. If, at the plan level, revenues are set in advance, any costs incurred by an individual reduce net revenue of the plan. In this way, while the consumer may not care about cost control, the plan will, and the plan will take actions to restrain spending, such as

setting copays and deductibles, managing care, negotiating efficient prices from providers, or creating networks of selected providers.

It turns out, however, that in the complex payment systems in use in many countries, "prospectiveness" of a payment system is not a yes/no characteristic, but a matter of the degree to which revenues depend on costs incurred. It has been infrequently recognized that most health plan payment systems are not fully prospective. Less common still is the application of measures of prospectiveness to health plan payment systems. While some payment system features such as reinsurance obviously incorporate some amount of cost-reimbursement, other features like risk adjustment also incentivize use though in less transparent ways, making accurate measurement of this aspect of payment critical. An objective of this chapter is to explain how researchers and policymakers can measure the degree of prospectiveness of a health plan payment system, and bring forward for discussion this policy-relevant aspect of health plan payment.

5.1.2 Plan of the Chapter

Table 5.1 previews treatment of four efficiency issues associated with plan payment methods covered in the next four sections. The purpose of each section is to propose valid, practical metrics for each dimension of efficiency.

[Insert Table 5.1 here]

5.2 Measures of Fit and Incentives at the Individual Level

This section explains the rationale for a measure of fit at the individual level as a metric of the efficiency properties of a health plan payment system. After review of the rationale for the

R-squared from a risk-adjustment regression, we present a generalization of the R-squared measure that is easy to compute and takes into account other aspects of the health plan payment system, not just the predicted values from a risk-adjustment model. This generalization is desireable when health plan payment systems contain other features in addition to a capitation rate based on a risk adjustment regression, such as premium categories and risk sharing. Assuming that fit at the individual level is a valid and relevant metric for the efficiency of health plan payment, the generalized fit measure we propose integrates other health plan payment features.

5.2.1 Rationale for R-squared from a Risk Adjustment Regression

By far the most commonly reported measure of the performance of a health plan payment system is the R-squared from a regression of spending at the individual level on the variables used as risk adjustors. Letting Y_i be the actual spending of individual i in the data used for calibrating the risk adjustment model, $\overline{Y}$ be the average spending in the population, and $\widehat{Y}_i$ be the predicted spending from the regression of Y_i on the risk adjustors, the R-squared of the risk adjustment model is:

$$R^2_{reg} = 1 - \frac{\sum_i (Y_i - \widehat{Y}_i)^2}{\sum_i (Y_i - \overline{Y})^2} \qquad (5.1)$$

We label this R-squared with a subscript "reg" to indicate that it comes from the risk adjustment regression (most commonly in practice a variant of ordinary least squares (OLS)).[5] The denominator in (5.1), $\sum_i (Y_i - \overline{Y})^2$, is the total sum of squares of individual spending, with higher values indicating that spending is more dispersed around the mean $\overline{Y}$. The numerator, $\sum_i (Y_i -$

[5] Most real-world risk adjustment models use weighted least squares (WLS) to accommodate partial year enrollees or population sampling weights. WLS weights then enter into the R-squared formula in the usual way. For simplicity, we ignore those issues here and refer simply to "ordinary" least squares.

$\widehat{Y}_i)^2$, is the "residual" sum of squares measuring the dispersion of spending in relation to the value predicted in the risk-adjustment regression. The better the risk adjustment model does fitting predicted to actual spending, the smaller is the residual sum of squares. If predicted values exactly fit actual values, the numerator is zero and $R^2_{reg} = 1$. If the regression equation explains none of the variation in individual costs, the predicted value is just the mean, implying the numerator equals the denominator and $R^2_{reg} = 0$. Real-world risk adjustment models typically fall somewhere in between with $0 \leq R^2_{reg} \leq 1$. It is common to report the R-squared as a percentage of the total variance explained by the model, so that they range between 0% and 100%.[6]

The R-squared from risk-adjustment regressions can strike those new to the field as being surprisingly low. Although age and gender are important predictors of health care costs, a regression with age and gender cells often explains only 1-4% of the variation in health care costs. This is due to the enormous variation in spending even within an age-gender cell. The risk adjustment formula used in the U.S. Medicare program based on a previous year's diagnoses raises the R-squared from a least squares regression to around 12%. The Dutch risk adjustment model for somatic care, which is the most sophisticated prospective risk adjustment system in the world, using 186 variables from a number of domains, including prospectively defined clinical variables, has an R-squared of about 31% (see Chapter 14). Chapter 3 contains an extensive discussion of the methods behind and the results of various risk adjustment models.

A higher R-squared is generally regarded as an improvement in the performance of a risk adjustment model. Buchner et al. (2013) for Germany, Beck et al. (2010) for Switzerland and

[6] Although the within-sample R^2_{reg} is always guaranteed to be nonnegative, outside of sample (validation) measures, or measures generated using simulation models that change the model specification can have negative values.

Van Veen et al. (2014) for the Netherlands use R-squared to compare individual-level fit across different risk adjustment models. Examples from the U.S. include the risk adjustment work by Kautter et al (2012, 2014) and the assessment of alternative risk adjustment specifications in the Society of Actuaries evaluation by Hileman and Steele (2016). While R-squared is by far the most common, it is not the only statistic used to evaluate risk adjustment models: each of these studies also includes others. The mean absolute prediction error (MAPE) uses the absolute value of the difference between actual and predicted spending, rather than squaring that difference as is done with an R-squared measure.[7] Arguments for the less-common alternatives to R-squared are generally made on statistical rather than economic grounds.[8]

The R^2_{reg} is the right metric to use to evaluate efficiency of a payment system under four assumptions: 1) health plans can take actions to encourage or discourage enrollees at the individual level (this is why fit would be figured person-by-person), 2) any inefficiency associated with those actions is proportional to the square of the gains and losses associated with the revenue and cost for the person (this is why we square the prediction-cost deviation),[9] 3) observed spending levels (Y_i) are the socially efficient spending levels, and 4) the predicted values from the regression are the exclusive basis for health plan payment (otherwise the

[7] See, for example, Van Barneveld et al. (2001) and Ettner et al., (2001). Van Veen et al. (2015) summarize fit measures used in this literature.

[8] Some papers propose an empirical measure of "how much of health care costs are predictable" by using extensive sets of information that consumers might have available for prediction, such as five years of past health care spending in Van Barneveld et al (2001) or something similar in Newhouse et al (1989) who estimate individual fixed effects based on several years of data. These predictions may of course under- or overstate how much consumers can actually predict. Researchers then compare the R-squared from a particular risk adjustment formula to this "maximum explainable R-squared."

[9] For example, suppose a health plan can direct treatment resources at the individual level and responds to the incentive to spend too much or spend too little based on whether the individual is a winner or a loser. In that case, a consumer's declining marginal benefit curve implies squaring the measure of incentive at the individual level is correct. See Layton, Ellis, McGuire and Van Kleef (2017) for a formal development including other assumptions necessary for the R-squared to be the exact metric to compare payment models in terms of incentives for economic efficiency.

predictions don't fully represent the actual payment model).[10] We discuss each assumption in turn before presenting our generalized measure.

Assumption 1), health plans can discriminate at the individual level for all potential enrollees, is unlikely to hold. Economic analysis of the dangers associated with adverse selection regard the health plan as discriminating in favor or against groups of enrollees, not individual enrollees; for example, persons with a certain diagnosis who are underpaid in the risk adjustment formula, or persons using a certain service (such as home care). If the plan acts at a group rather than an individual level, a group level measure of fit is the appropriate one. For example, a plan might be underpaid by 20% for users of home care, and have incentives to underprovide this care. It is not important to plan incentives that some people within the group of home-care users are underpaid more or less leading to the 20% underpayment.[11] Risk adjustment researchers are aware of this issue and often present group-level measures of fit (such as group over/undercompensation) and group R-squared to supplement reports of model fit at the person level.

Assumption 2) has a sound basis in welfare economics, where it is normally assumed that the efficiency cost of a distortionary incentive is proportional to the square of the distortionary incentive. A distortion may move a decisionmaker (consumer, producer, plan) away from the optimal decision in some linear fashion, but a small movement near the optimum may have little efficiency effect whereas the same size movement far away from the optimum will have a large efficiency effect. Figure 5.1 illustrates the rationale for squaring the price distortion as a

[10] And the assumption that plan actions to discriminate in favor/against some enrollees is the main efficiency issue. The R-squared measure is not well-suited to measuring efficiency incentives with respect to enrollee choice of plan or incentives for cost containment.

[11] An assumption here is that within-group variation of profits and losses is not correlated with differences in consumer response to selection actions. We come back on this assumption in Section 5.3.5.

measure of inefficiency in the familiar context of a tax. If a tax t is imposed, price rises t above marginal cost (MC). The welfare loss associated with price MC + t is shown in the Figure. Also shown is how the welfare loss quadruples (squaring) with a doubling of the tax to 2t.

Figure 5.1: Efficiency Effects of a Price Distorton go up with the Square of the Distortion

[Insert Figure 5.1 here]

Assumption 3) is unlikely to be true. Observed spending levels are likely to be different from efficient levels unless the optimal payment system is in place and competition is perfect, among other things. However, because efficient spending levels are typically unknown, and the efficient levels are the correct benchmark for welfare analysis (see Figure 5.1) the researcher must specify some spending to be efficient. Observed spending, especially observed spending from a well-functioning setting (such as employer provided insurance in the U.S.), has sometimes been assumed to be efficient by researchers.[12]

Assumption 4) will be true in some institutional circumstances and not in others. It is reasonable to assume that the degree of fit of revenues to costs is captured by the R-squared from a regression in Germany, Israel and the U.S. Medicare Advantage program where a plan's revenue is tied closely to the empirical risk adjustment model. In other contexts where premium categories influence payment (U.S. Marketplaces, Ireland, Switzerland) or where there is risk sharing (U.S. Marketplaces, Ireland, Switzerland, Australia), or where the risk equalization payment is made up of more than one predictive model (the Netherlands) the payments a plan receives for a person depend on more than the statistical fit of the risk adjustment formula. In

[12] This is a rationale for why data from traditional Medicare are used to calibrate payment models for Medicare Advantage plans. See Bergquist et. al. (2018) for discussion of this issue.

Switzerland, a plan receives a risk equalization payment and a payment for each day an enrollee is hospitalized. Any incentives for or against individuals or groups are generated by the full set of payments a plan gets, not just from one feature of the payment system.

Judging how well the full payment system fits costs ideally includes taking all features into account. Even if the purpose of an analysis is to assess only the risk adjustment methodology, taking account of the other features of payment is necessary to more accurately gauge the incremental contribution of risk adjustment.

5.2.2 Generalizing the R-squared: Payment System Fit

Providing a rationale for a fit measure at the person level requires acceptance of the first three assumptions: 1) plans can discriminate at the person level, 2) efficiency loss goes up with the square of the distortionary incentive, and 3) observed spending levels are equal to optimal levels. Our generalization has to do with assumption 4); more specifically, our modified metric generalizes the R-squared to account for other payment system features. In health care systems where predicted values from the risk adjustment model nearly fully capture payments, our generalized metric reduces to the R-squared from the risk adjustment model. When other payment features (e.g., risk sharing) are present, the metric takes them into account in a way consistent with assumptions 1) - 3).

The generalization is based on the simple idea that incentives are created by the relationship of plan revenues to costs. Revenues to a plan for a person are what matters for how much the plan chooses to allocate to that person, and the revenue function can have more components than the predicted values from the risk adjustment regression model. Payment System Fit (PSF) is constructed by substituting the revenue a plan would receive for a person for the predicted value from the regression. Then, an R-squared-type measure describes the

population-level individual fit of payments to costs. PSF measures the "explained variance" in costs accounted for by the full set of payment system features, not just the variance explained by the risk adjustment model.

Textbox 5.1: Payment System Fit

Payment system fit substitutes the simulated payment that a plan would receive for enrolling an individual for the predicted value from the risk adjustment regression model. In relation to the formula for the regression R-squared presented above (R^2_{reg}), payment system fit replaces the predicted value $\widehat{Y}_i$ with the revenue R_i a plan receives for each person. Thus,

$$PSF = 1 - \frac{\sum_i (Y_i - R_i)^2}{\sum_i (Y_i - \overline{Y})^2} \qquad (5.2)$$

This is analogous to an R-squared and is in fact equal to the R-squared if risk adjustment is the only factor determining plan payment. It differs from R-squared from a regression if plan revenues depend on other payment system features, such as through premium categories or risk sharing.

Geruso and McGuire (2016, page 9) compare conventional R-squared fit with Payment System Fit in the U.S. Marketplaces (using data from Marketscan, which are those used to calibrate Marketplace risk adjustment). Concurrent risk adjustment alone has an R-squared of .37. During 2014-2016, plan payments in Marketplaces also included reinsurance. Adding the 2014 version of reinsurance (100% coverage after $45k in annual expenses), increases the Payment System Fit to .61. A conventional R-squared measure has no way to consider the fit of both elements when used in tandem.

5.2.3 Comments on Individual-Level Fit Measures

In spite of its tenuous basis as an economic efficiency metric, the R-squared from a risk adjustment regression remains a natural and easy-to-compute metric for the performance of a risk adjustment model. It is intuitive that better fit at the person level should improve the performance of a payment system with respect to selection problems. In settings in which only relative risk scores from a regression model determine payments, and discrimination at the individual level is an issue, the R-squared has a sound basis in economics. It is a short hop from there to account for other payment system features, should they exist, within a concern for individual-level discrimination. Our proposed Payment System Fit makes that hop.

Replacing predicted values from the risk adjustment regression model by revenues a plan receives for an individual is called for even if the deviations are not squared and summed as in the R-squared. Other metrics of individual fit, such as the mean absolute prediction error (MAPE), also benefit by the generalization to payment system fit. Replacing simple predicted values with revenues that reflect predictions minus imputed premiums and risk sharing at the person level is also part of what we recommend for measures of fit at the group level, a topic we turn to next.[13]

5.3 Measures of Fit and Incentives at the Group (or Action) Level

Restrictions on risk rating of premiums and open enrollment provisions in individual health insurance markets are intended to prevent health plans from discriminating on the basis of price or access to health insurance at the individual level. Health plans can, however, still take actions to discourage or encourage enrollment by targeted groups of consumers, referred to in the

[13] Chapter 4 discusses empirical methods for incorporating the presence of premium categories and risk sharing into the estimation of the risk adjustment model. The payment system fit measure remains the relevant one because it incorporates the "explanatory power" of all payment system features.

research literature as "indirect selection," "service-level selection," "supply-side selection," or "cream skimming."[14] The potential for this type of insurer behavior raises two key questions: "What groups?" and "What actions?" The answers to these questions will depend on the market being studied. For example, in the Netherlands, individuals reporting low health status or multiple chronic illnesses have been identified as potential targets for plan underservice (Van Kleef et al. 2013; Eijkenaar et al. 2017). In the U.S., researchers have studied users of particular classes of drugs (Carey 2017a; Carey 2017b; Han and Lavetti 2017; Geruso, Layton, and Prinz 2017), users of certain hospitals (Shepard 2016), users of certain types of services (Ellis and McGuire 2007; McGuire et al. 2014), and population subgroups such as nursing home residents and amputees (Pope et al, 2011).

At the close of this section we will recommend that for this purpose groups be defined on the basis of discriminatory actions available to plans in the market under study. We will also explain that for a tactic to be effective as a selection device, it must be recognized by consumers (otherwise they do not respond). We begin with a discussion of some of the general issues regarding measurement of incentives to discriminate against (or in favor of) a group of potential enrollees.

The risk of under and overservice for certain groups of enrollees is well-recognized by architects of health plan payment systems. In Europe, incentives to serve certain groups (e.g., those with multiple chronic illnesses) is typically assessed by measuring over and undercompensation for a group. Researchers in the U.S. concerned with the same issue form a ratio rather than a difference between predicted values and costs.

[14] The literature on service-level or "supply-side" selection began with studies of the incentives of insurers to distort service-level offerings to attract good risks based on models of health plan profit maximization. Geruso and Layton (2017) provide a recent review of this literature.

This section first presents the rationale for group-fit measures such as over/undercompensation and predictive ratios. We note some shortcomings of these measures and suggest three lines of improvement: 1) recognizing other elements of the payment system (as in Section 5.2 and fit at the person level), 2) developing a comprehensive plan-wide measure of group fit covering the service of interest as well as all others, and 3) improving the measure of plan incentives by recognizing that incentives created by a given amount of over and undercompensation will differ for different people.

5.3.1 Rational for Over/Undercompensation and Predictive Ratio Measures

Presently used metrics to assess incentives at the group level compare predicted values from a risk-adjustment regression to actual costs for a defined group of consumers. We will thus refer to these as measures of group (as opposed to individual) level fit. An example would be a group of consumers who used home care in a previous period. The question these measures address is, "Does the payment system adequately pay plans for enrollees who used home care in a previous period?" The concern is that if the system does not pay adequately, a plan might take actions to discourage membership from among this group, by, for example, unduly restricting access to home-care services.

As in the previous section, let $\hat{Y}_i$ be the predicted value from the risk adjustment regression for individual i, and Y_i be i's actual cost. Let $i \in g$ indicate the individuals in the group, g, of concern, and n_g be the number of consumers in group g. A commonly used measure of possible over- or undercompensation for group g is:[15]

[15] This over/undercompensation measure is the negative of the more familiar mean prediction error (MPE) which is widely used in statistics. Using the negative makes positive values correspond to positive profits when the predictions are thought of as a measure of revenue.

$$\text{over/undercompensation} = \frac{\sum_{i \in g}(\hat{Y}_i - Y_i)}{n_g} \qquad (5.3)$$

The over/undercompensation measure is the average for group g and is measured in monetary terms (e.g. Euros or dollars). When (5.3) is positive it indicates overcompensation and when it is negative, undercompensation.

A predictive ratio uses the same elements:

$$\text{predictive ratio} = \frac{\sum_{i \in g} \hat{Y}_i}{\sum_{i \in g} Y_i} \qquad (5.4)$$

The predictive ratio is a unit-free number. When (5.4) is greater than 1.0 it indicates overcompensation, and when less than 1.0, undercompensation.

Over/undercompensation and predictive ratios are both useful measures of group-level incentives. We will, however, argue in favor of modifying them to better reflect the full set of payment system features. The usual interpretation of these metrics is that if over/undercompensation is near zero, or the predictive ratio is near one, a plan has little incentive to discriminate in favor or against members of group g. As overcompensation grows more positive (negative) or the predictive ratio goes above (below) one, a plan has an incentive to attract (deter) members of the group. Expression (5.3) makes clear that over/undercompensation is a group-level measure, which is appropriate if insurer actions operate at the group level.

Over/undercompensation, either in the form of a difference or a ratio, is routinely assessed for selected groups in many risk adjustment contexts. For example, Van Kleef et al. (2013) merged survey information with health claims for a subset of people in the Netherlands to calculate "undercompensation" (defined as the difference in costs and predicted revenue rather than their ratio) for various groups of people, including those with low physical and mental health scores and those with chronic conditions. They compare seven different risk adjustment

models with different sets of explanatory variables. For the risk-adjustment model used in U.S. Marketplaces, Kautter et al. (2014, E22) computed predictive ratios for various subgroups defined by predicted costs. In their evaluation of the CMS-HCC model, Pope et al. (2011) report predictive ratios for a large number of subgroups, including groups defined by disease, numbers of prior hospitalizations, demographic characteristics, and others.

Other papers assess the evidence for service-level distortions without measuring the incentives to engage in service-level selection. Cao and McGuire (2003) in Medicare and Eggleston and Bir (2009) in employer-based insurance find patterns of spending on various services consistent with service-level selection among competing at-risk plans.

Some papers do both, assessing incentives and checking for evidence of under/oversupply. Ellis, Jiang and Kuo (2013) rank services according to incentives to undersupply them. Consistent with service-level selection, they show that HMO-type plans tend to underspend on predictable and predictive services (in relation to the average) just as the selection index predicts. This pattern of spending is not observed among enrollees in non-HMOs.

A number of recent papers focus on groups defined by use of a certain class of drugs. The "action" here is a plan's decision to cover a group of drugs generously or not by tier placement on the drug formulary. This active area of recent research confirms that with respect to this readily measured "action," payment models create incentives and plans respond. In particular, plans distort coverage to attract the healthy and avoid the sick. Carey (2017a, 2017b), and Han and Lavetti (2017) study incentives for selection in Medicare Part D and document evidence that Part D insurers respond to those incentives when designing their drug formularies. Other recent work has focused on identifying evidence of service-level selection among

Marketplace plan formulary contracts. Geruso, Layton, and Prinz (2016) use data on Marketplace plan and self-insured employer plan formularies to determine whether differences between Marketplace formularies (where selection incentives are strong) and employer formularies (where there are no selection incentives) correspond to the strength and the direction of the selection incentive associated with a particular drug class. They find that selection incentives are minimal in this setting due to a well-functioning payment system, but for the drugs where payment "errors" exist, they find robust evidence that Marketplace plans severely limit coverage and access for drug classes that are used by the most unprofitable enrollees.

Finally, another recent paper analyzes groups defined by their use of a particular "star" hospital system in Boston. Shepard (2016) shows that people who switch plans in response to one plan's decision to drop the hospital system from its network have costs that greatly exceed the revenue they bring to the plan. Using counterfactual simulations, he finds that in equilibrium, this underpayment would lead to this star hospital system being dropped from all health plan provider networks, a finding that has effectively played out in this market in recent years.[16]

5.3.2 Identifying Potential Actions and Groups of Interest

In thinking about group-level measures, what groups are relevant? How should a population be grouped with respect to incentives for plans to act at the group level? Some years ago, Newhouse (1993) defined risk selection as "actions by consumers and health plans to exploit unpriced risk heterogeneity…" A key word in this definition is "actions." Plan actions to exploit unpriced risk heterogeneity consist of tactics to discourage enrollment of the unprofitable and encourage enrollment of the profitable. Groups should therefore be defined as those that may be affected by a plan action. For example, if plans can *only* take actions that

[16] See also Kuziemko, Meckel, and Rossin-Slater (2014) for a study of Medicaid managed care plans attempting to attract lower cost births based on the race-ethnicity of the mother.

discriminate between people under the age of 65 and those above the age of 65, these become the groups of concern when it comes to (measuring) risk selection (incentives). If plans can only discriminate on the basis of "yes/no chronic condition" then these are the two relevant groups. If health plans can discriminate on combinations of "yes/no >65" and "yes/no chronic condition", there will be four groups of concern, and so on.

Some research defines groups according to geography under the thinking that a health plan might favor or disfavor certain regions because of systematic regional differences in medical spending, as was done in a study of risk selection in Germany by Bauhoff (2012). Other research defines groups according to the services used, the idea being that a health plan could favor or disfavor primary versus some kinds of specialty care, for example, to encourage/discourage potential enrollees anticipating making use of those services.[17] Studies of selection and drug formulary design discussed in the previous section typically assume that insurer actions take place at the level of the drug class (Carey 2017a; Carey 2017b; Geruso, Layton, and Prinz 2017; Han and Lavetti 2017). Studies of selection and network design assume insurer actions take place at the level of the hospital or physician group (Shepard 2016).

Since the instruments for health plans to engage in risk selection differ across health care schemes, there is no universal set of relevant groups. Thus, an important step for evaluating incentives for risk selection in a particular setting is to identify the possible selection actions in that setting and to derive the relevant groups. For example, in the Netherlands health plans are unable to discriminate at the individual level due to open enrollment requirements. On the other hand, plans can discriminate across groups on the basis of network design. For example, contracting with first-best physicians for treatment of disease X will attract patients with disease

[17] See Ellis and McGuire (2007) for implementation of this approach in Medicare and McGuire et al. (2014) for its application in Marketplaces.

X; conversely, a poor network in terms of quality or convenience will deter patients in that disease group. When a plan can make a network decision hospital-by-hospital, study of groups defined by those using individual hospitals may be called for.

Van de Ven et al. (2015) identify a number of specific selection actions in the Netherlands that can occur as a consequence of over/undercompensation, including selective advertising, offering choice of deductible, making supplementary insurance (un)attractive for certain groups, offering group contracts and quality skimping on certain services. To measure the incentives involved requires a designation of the group affected. Advertising may be targeted to certain populations, e.g., young families, or group contracts may be offered to only selected groups among the population.

An important corollary of this discussion is that if there is no action a plan can take with respect to a group, there is no point, and indeed, it may be misleading, to construct incentive measures for that group.

5.3.3 Generalizing Over/Undercompensation and Predictive Ratios to Include Other Elements of Plan Payment

Once the simulated payment amount for each person is available, ratio and difference measures of over and undercompensation can be easily modified to incorporate other plan payment features, such as risk sharing, a modification that improves the validity of the measures of incentives at the group level. Incentives to a plan to attract/deter members of a group are governed by net revenues. Inclusion of all elements of net revenues yields the valid measure of these incentives.

McGuire et al. (2014) modify predictive ratios incorporating premium differences and risk sharing in the U.S. Marketplaces. The numerator of the "payment system predictive ratio"

for a subgroup is the sum of the payments for the group (which can depend on all payment system features) rather than the regression predicted values. The denominator in these predictive ratio measures remains the actual costs for the groups. Geruso, Layton, and Prinz (2016) modify predictive ratios and under/overcompensation measures in the same way.

5.3.4 Generalizing Group Fit to the Entire Population

Studies of fit at the group level typically report under/overcompensation or predictive ratios for a subset of the population (e.g., those with a chronic illnesse). When predictive ratios are computed for the entire population (e.g., those with a chronic illness and those without a chronic illness), the statistics are not summed or aggregated in any way to provide an *overall measure* of fit at the group level. By contrast, the payment-system fit measure noted above for assessing fit at the person level summarizes fit for the entire population (in the form of the reduction in sum of squares of the payment-cost residuals).

A summary measure may be useful for group fit as well. While we can agree that reducing undercompensation for a group of interest is an improvement for that particular group, what if a payment system alternative decreases undercompensation for one group but increases it for another? Which alternative is preferred? If payment alternatives are all subject to the same overall budget constraint, moving payments more towards one group inevitably lowers payments for another group. This could be a good thing if the group experiencing lower payments was initially overpaid; it would be a bad thing if the group were initially underpaid and the policy change exacerbated an underpayment problem.

A group-level measure analogous to the individual level measure discussed above is a natural way to summarize group fit at the population level (Van Kleef et al., 2017). Suppose

potential actions by a plan allows health plans to discriminate among G mutually exclusive groups indexed by g with g = 1, …, G. We can then use data to determine:

s_g	the share of the population in group g, with $\sum_g s_g = 1$,
$\overline{R}_g$	the average plan revenue for a person in group g,
$\overline{Y}_g$	the average plan cost for a person in group g,
$\overline{R}_g - \overline{Y}_g$	the average under/overcompensation for a person in group g.

Given these parameters, under and overcompensations can be summarized in several different ways. One possibility is $\sum_g s_g|\overline{R}_g - \overline{Y}_g|$, i.e. the sum of absolute under and overcompensations weighted by the share of the affected population. As this metric falls, fit improves.

Most closely analogous to the payment system fit measure above, however, is a group fit measure that weights the squared group-level payment-cost residuals and is scaled to fall between zero and one (like an R-squared or a payment system fit). Our measure is analogous to the one presented by Ash et al (1989) who measured regression fit at the group level by a Grouped R-squared. We generalize this measure and call it the Group Payment System Fit (GPSF) because it incorporates other payment features.

$$\text{GPSF} = 1 - \frac{\sum_g s_g\left(\overline{Y}_g - \overline{R}_g\right)^2}{\sum_g s_g\left(\overline{Y}_g - \overline{Y}\right)^2} \qquad (5.5)$$

The denominator of (5.5) is the total sum of squared residuals at the group level. The numerator is the sum of squared group-level residuals after the payment system is in place. Analogous to

an R-squared or payment system fit measure at the individual level, $0 \leq GPSF \leq 1$, with higher values indicating the payment system is doing a better job at matching revenues to costs at the group level.

Squaring the group level payment-cost residuals has grounding in welfare economics, where the efficiency loss associated with a price distortion (such as a tax) is proportional to the square of the distortion at the group-level. A related argument supporting raising the group-level residual to a power greater than 1.0 comes from Van Barneveld et al. (2000) who contend that small predictable profits and losses are likely to be irrelevant for a health plan. Selection can be costly and the net benefits are uncertain, and small incentives may simply not induce a health plan to act.

Depending on the institutional circumstancs, other functions of the group-level payment-cost residuals may be justified. Van de Ven et al. (2015) point out that overcompensation may lead to an improvement in quality whereas undercompensation leads to a deterioration of quality. It may be that undercompensation is worse than overcompensation. A metric to represent this would be the group population weighted sum of only the negative deviations (squared or not), similar to that used in Shen and Ellis (2002).[18]

In the end, while we believe squaring and summing group-level errors with population weights is a natural way to measure incentives around group fit, depending on the circumstances, researchers may justify and choose other functions of the weighted residuals.

5.3.5 Taking Account of Consumer Response

Measures based on predictive ratios or under/overcompensation are missing a key element of selection incentives: how consumers (in a group) will respond to the action in

[18] A related argument is made by Lorenz (2014) who also identifies empirical methods that weight over and undercompensation asymmetrically.

question. If consumers cannot or simply do not respond to the action in question, the plan has no incentive to take it, even if the group in question is under or overcompensated. Here is a simple example. Suppose the targeted group is young families for whom a plan is overcompensated. The action is advertising in newspapers and television. If young people do not respond (perhaps because they get their news elsewhere) to newspaper advertising, in spite of the overpayment, plans have no incentive to take the action of newspaper advertising.

The same point applies to health care services. Unless consumers respond to skimping or overprovision of services, the plan has no incentive to take the action. Another example is the following: suppose plans are undercompensated for members who use ambulance services during a year. But suppose also that use of an ambulance cannot be anticipated by consumers. Specifically, consumers do not know whether they are at high or low risk for using an ambulance. In that case, skimping on ambulance services will not disproportionately discourage enrollment by the group for which the plan was undercompensated. Indeed, the more consumers can correctly anticipate that they will or will not be users of a certain service, the more effective an action on that service will be with respect to separating risks. Well-baby care will be very appealing to young families anticipating have a child, but irrelevant to young couples who have decided not to have children. Young families might well know into what group they fall. More generally, it is the profitability of the consumers whose choice to enroll in a plan is marginal to the plan's decision of how much of a particular service to provide who matter for plan incentives (Veiga and Weyl 2016). Consumers whose plan choice does not depend on the plan's actions with respect to the service do not matter, even if they are heavy utilizers of the service in question.

With that qualification in mind, it remains true that services affecting those with a chronic illness are likely to be effective selection tools. The idea of a chronic illness is that it is persistent, and therefore likely to be anticipated. Those with diabetes this year are very likely to have diabetes next year, and these people are likely to be well-aware of their health situation. In this case, plan choice of those with diabetes are very likely affected by the level of diabetes-related services offered by the plan. Restricting access to care important to consumers with diabetes is thus likely to be an effective strategy, should plans be undercompensated for this group.

A key factor in determining whether a consumer is likely to respond to changes in the level of a service offered by a plan is likely to be the "predictability" of the service. Research papers in health economics have studied the role of predictability of health care use by consumers and its role in incentives to plans to under/over provide services (Ellis and McGuire, 2007). Because predictability is measureable, at least in part, the concept has played a prominent role in measuring which services are more likely to affect consumer plan choices. In the research literature, the total selection incentive is measured by combining a measure of over/undercompensation with a measure of how well consumers can anticipate their use of a service. Research papers show that the incentive to select against a service is a function of its predictability, how well it predicts profitability (termed its "predictiveness"), the variation of profits, and the demand elasticity.[19]

5.3.6 Summary Comments about Action/Group-Level Measures of Incentives

[19] The theory of plan incentives to use services to affect selection is presented in Frank, Glazer and McGuire (2000) and Ellis and McGuire (2007). The ideas are developed and applied empirically in McGuire et al. (2014) and Ellis et al. (2017).

In summary, one of the key distortions that plan payment systems are designed to combat is distortions to health insurance contracts to attract profitable enrollees and deter enrollment by unprofitable ones. An insurer's incentive to distort its plans in such a way is related to the extent to which a group is over/undercompensated. The extent to which over/undercompensation matters for a given group depends on (1) whether an insurer can target the group in some way, (2) whether members of the group would respond to distortions targeted at them, and (3) the size of the over/undercompensation. We developed measures of incentives for insurers to engage in these distortions where the researcher determines (1) and (2) while (3) is estimated from data.

5.4 Measures of Incentives for Consumers to Choose the Right Plan

Many individual health insurance markets allow for a variety of plan types (e.g. more/less coverage, variation in network design, or other differences) for the purpose of serving consumer preferences and rewarding plans for successful innovation. In the Netherlands, consumers can choose among plans with a range of deductibles, with lower premiums associated with higher-deductible plans. Dutch plans also contract with different networks of providers. In the U.S. Marketplaces, consumers choose among metal levels with a gold plan covering a larger share of costs than a silver plan, which in turn covers more than a bronze. Marketplace plans also construct different provider networks. An element of the efficiency of the market for health insurance is to encourage consumers to enroll in the "right" plan for them, defined as the plan that offers them the most net benefits over cost. This is not only a static issue; it is also important in a dynamic framework with innovation. As consumers move to plans with better value, incentives are conveyed to the plans to innovate in ways to improve value to consumers.

Using consumer choice to reward high-value plans and punish low-value plans is the essence of the "competition" element of Enthoven's vision of "managed competition." If plans innovate by improving value, it might be more expensive and they would have to charge more, but if the innovation is worth more than the cost consumers will reward the plan with enrollment. If plans innovate by reducing costs, competition will press them to pass on the savings to consumers in the form of lower premiums. In this setting, with consumers facing different prices approximating the costs of the alternative insurance products, Enthoven argued that the market would lead to an efficient allocation of consumers across plans. There are two factors that interfere with a market producing the premiums that lead to efficient choice of plans. We explain these, and how to measure the problems they cause, after first describing what efficient pricing looks like.

5.4.1 Efficient Premium Pricing

We now turn to the role that premium pricing plays in influencing efficient enrollee plan choices. It is worth recognizing up front that no health insurance market could realistically achieve the set of premium prices necessary to fully meet the ideal of efficient sorting of consumers across plans. To see this, consider a simple setting with just two plans, Plan A and Plan B, with somewhat different characteristics. For concreteness, suppose Plan A offers full coverage and Plan B has a large deductible. Consumers are heterogeneous in their costs (at each plan) and in their tastes with respect to the presence of a deductible (because of risk aversion or other reasons).

Consumer 1 "should be" in Plan A if that consumer's extra or incremental valuation of Plan A with the full coverage is greater than the incremental plan cost Consumer 1 would incur in a plan without a deductible compared to one with a deductible. This rule for efficiency leads

consumers to the plan in which their benefits most exceeds the costs.[20] The same statement of efficient sorting could be made for consumers, 2,3,…..N. It is immediately clear that in the case where each consumer faced a price difference (referred to in the literature as an incremental price) between the two plans equal to *that particular consumer's cost difference in plan cost* between the plans, consumers will sort efficiently.[21] If consumer 7 certainly will go over the deductible, the incremental price to 7 is approximately the deductible,[22] and, at that price, 7 would choose whether Plan A is preferred at that price, and this choice would be efficient. If consumer 11 faces a low probability of having any medical spending, the price to 11 should be much lower. 11 might prefer choice and be willing to choose A at the premium right for her. If consumers 7 and 11 face the same incremental price for Plan A, the resulting choices may not be efficient. This is the "single-premium problem" in the research literature (Bundorf, Levin, and Mahoney 2012; Geruso 2017). Generally, no single premium can sort consumer efficiently between two plans.[23]

5.4.2 First Source of Deviation from Efficient Pricing: Limited Premium Categories

[20] Note that the efficiency rule has to do with plan costs, not total costs. If a deductible only shifts costs from the plan to the consumer, only the portion of spending covered by the plan should be part of the efficient incremental premium. All plan costs, including administrative costs, should be considered when evaluating efficiency. We proceed by effectively assuming that administrative costs for a given consumer are constant across plans.

[21] This argument is made in Keeler, Carter and Newhouse, (1998), among other places. The argument is the same as that for prices generally: when consumers face prices equal to costs, utility-maximizing consumers make socially efficient choices. For now, we ignore the distinction between expected and realized cost. We will recognize the importance of expected costs in developing our proposed measure below. We also ignore deviations between willingness-to-pay (demand) and underlying valuation that may be caused by behavioral frictions (Spinnewijn 2017).

[22] This is the incremental plan cost for the plan without a deductible. All other costs are covered similarly in the two plans.

[23] A single premium can sort efficiently in some special cases. For example, if all heterogeneity in preferences is perfectly collinear with expected costs, a single premium can achieve efficient sorting. This is the special case in Cutler and Reber (1998).

It is obviously not realistic to expect a market, regulated or not, to generate incremental premiums to be person-specific, even in this simple setting of just two plans. Asymmetric information can be one barrier. Consumers may know if they are likely to be high-cost but some of this information may be unavailable to plans. More important, however, is that regulation constrains plan risk rating. As shown in part 2 of this volume, in most markets with regulated competition, risk-rating of premiums is proscribed for purposes of fairness and access. It is understood that such regulation comes at some cost in terms of efficiency,[24] but this cost is generally regarded as tolerable in exchange for the gain in fairness achieved by having the healthy subsidize the sick in health insurance purchase.[25] Even so, there are alternatives for regulating premiums in pursuit of fairness, e.g., subsidizing the sick by a special risk pool, allowing age bands or not, restricting differences between the old and the young, etc.; but there are always unavoidable tradeoffs between fairness and efficiency. It is therefore worthwhile to be able to measure the comparative efficiency in terms of sorting of the various approaches to fairness.

The discussion here is in terms of *incremental* plan costs, *incremental* benefits, and *incremental* premiums. Incremental plan costs of a person are the *difference* in plan costs (not the total or out-of-pocket cost) for that person between plans; incremental benefits refer to the person's individual subjective *difference* in the valuation of alternative plans. Incremental premiums are differences in premiums in a market and will be group- rather than person-specific.

[24] For an early treatment, see Pauly (2008).
[25] Gains in fairness may also represent efficiency gains. The transfers from healthy to sick consumers induced by limited premium categories also effectively provide insurance against the financial consequences of transitioning from a healthy to a sick state. Indeed, Handel, Hendel, and Whinston (2015) show that in a setting similar to the ACA Marketplaces, efficiency gains from limiting "reclassification risk" exceed efficiency losses due to adverse selection when comparing risk-rated premiums to a single premium policy.

For example, if the community-rated premium for Plan A is 800 Euros per month and the community-rated premium for Plan B is 600 Euros, the incremental premium – the amount a consumer can save by choosing the lower-priced plan -- is 200 Euros.[26]

The intuition behind the measure we propose for the inefficiency associated with limited premium categories can be illustrated in Figure 5.2, which shows, in the case of two plans (again, Plan A and Plan B), the distribution of the incremental cost of consumers in Plan A compared to Plan B along the horizontal axis. We assume these differences are all positive (Plan A costs more for everyone), and that the incremental costs are distributed uniformly on the line. (Neither of these simplifications are important for the argument.) As was said in the previous section, if each consumer faced an incremental premium between the plans equal to her specific difference in cost, all consumers would sort themselves efficiently between Plans A and B.

Figure 5.2: A Community Rated Premium Leads to a Large Average Gap Between the Incremental Premium and Incremental Costs

[Insert Figure 5.2 here]

Suppose premiums are community-rated so that premiums are different for Plan A and Plan B, but everyone pays the same premium for each plan. If Plan A drew a representative set of consumers from the population, the diffence in average cost between Plan A and Plan B would be the average of the incremental costs. With competition, the premium difference for

[26] The analysis here assumes that consumers choose on the basis of premium differences across plans, i.e., would make the same choices if plan premiums were 100 Euros and 300 Euros as when the premiums were 600 Euros and 800 Euros. If consumers react to relative prices rather than differences in absolute prices, this assumption is questionable. Douven et al (2018) question whether consumers decide on the basis of price differences independent of the level of prices.

Plan A compared to Plan B would then also be the average of the incremental cost differences, shown in Figure 5.2 as a vertical line at the midpoint (the average) of the incremental costs. Except in special cases, this incremental premium will not lead to efficient sorting: the premium difference is too high for the consumers to the left of the vertical line and too low for the consumers to the right. Too few of the low-cost consumers are likely to choose Plan A and too many of the high-cost consumers will choose Plan A (we come back to the implications of this for premiums shortly).

Measures of the misaligned incentive from community rating are based on the gap between the efficient incremental premium (the person's incremental plan cost) and the incremental premium they face because of community rating, a measure analogous to the "price distortion" measure common in welfare economics. This gap can be summed (or averaged) in a linear accounting of the distortion, or greater deviations can be given more weight by squaring the gaps before summing. The mean absolute deviation of the distribution of incremental cost, the expected value of this divergence, is the linear measure of the gap. The mean absolute deviation is shown in Figure 5.2. The quadratic measure would square the differences before summing to yield the variance of the incremental cost distribution.

We now illustrate how this premium efficiency measure can be used to compare policy alternatives. Suppose instead of community rating, regulators set two premium categories, one for the young and one for the old. Assume there are equal numbers of young and old in the population. Since the old tend to be more costly, raising the incremental price for the old and lowering it for the young will tend to improve the match between the incremental premium consumers face and their incremental cost. More young, facing a lower incremental premium,

will be induced to correctly choose Plan A and fewer old, facing a higher incremental premium, will be correctly discouraged from choosing the more generous plan.

The change in the average gap between prices and incremental costs measures the improvement in sorting incentives achieved by risk rating by old and young. Figure 5.3 shows consumers divided into the young and old, and to keep the illustration simple, the young are assumed to be to the left with lower incremental costs and the old are to the right with higher incremental costs. If we allow risk rating by these two age categories, the incremental premium for the young will fall at the midpoint of the young distribution of incremental costs and the incremental premium for the old will fall in the middle of the old distribution, as shown in the figure. The figure also shows that the new mean absolute deviation (which is the same for the young and the old), is smaller than the standard deviation with community rating. In this example, the mean absolute deviation falls to exactly half of the previous value. We could also use a squared measure of the deviations in terms of the variance of the gap between incremental costs and premiums, and that would also fall. In this example, the variance would fall to one-quarter of the previous level.

Figure 5.3: Risk Rating by Age Reduces the Average Gap Between Incremental Premiums and Incremental Costs

[Insert Figure 5.3 here]

This example makes clear that a measure of the fit of incremental premiums to incremental costs is the natural way to measure how well a set of premium categories conveys

efficient incentives to consumers regarding choice of plan.[27] Following the approaches proposed earlier in this chapter, we measure fit of (incremental) premiums to (incremental) costs with a linear and a quadratic metric.

The linear measure is the mean absolute prediction error (MAPE) associated with a set of premium categories, normalized by the mean absolute prediction error with a single community-rated premium (equal to the difference in average cost). Taking one minus this measure transforms it into a measure with range 0 to 1 with higher values better (like other measures of fit):

$$\text{Premium MAPE} = 1 - \frac{\sum_i |\Delta Y_i - \Delta P_i|}{\sum_i |\Delta Y_i - \overline{\Delta Y_i}|} \qquad (5.6)$$

The Premium MAPE in (5.6) linearly accumulates the individual-level price distortions associated with a given payment system. In making such a summation, it may be easier to think of i as representing types of individuals, who have an average or expected incremental cost, than i representing each person.

A case can be made that the metric should be quadratic rather than linear. In that case the measure is simply the R-squared from a regression of incremental costs on incremental premiums. Obviously, one incremental premium (the average) explains none of the variance. As the premium categories bring incremental premiums closer to incremental costs, the fit improves. We call the quadratic measure Premium R-squared, also normalized and subtracted from one (like a conventional R-squared):

[27] Incremental costs can be thought of in terms of expectation if the metric developed here is applied ex ante. The expectation is an objective expectation (not necessarily what the consumer might be able to forecast). In implementing this idea we use data on actual costs averaged over the types of interest to estimate expected costs.

$$\text{Premium } R^2 = 1 - \frac{\sum_i (\Delta Y_i - \Delta P_i)^2}{\sum_i (\Delta Y_i - \overline{\Delta Y_i})^2} \qquad (5.7)$$

Although simple and intuitive, there are three notable limitations to both the linear and quadratic measures presented in (5.6) and (5.7), respectively, which limit the applicability of our metric in many settings. First, as is the case with all the *ex ante* measures presented in this chapter, (5.6) and (5.7) stop short of measuring welfare loss due to adverse selection. Instead, they measure price distortions. A measure of welfare requires information (or assumptions) about demand/willingness to pay, and demand response. In particular, these price distortions do not *necessarily* lead to inefficient sorting: While a price equal to the individual's incremental cost ensures efficient sorting, there is an infinite number of other prices that will result in the individual making the same choice. However, while other prices induce efficient sorting only under a specific level of demand/willingness-to-pay, the incremental cost induces efficient sorting under *any* level of demand, making it the natural benchmark with which to evaluate price distortions.

Second, measures (5.6) and (5.7) require information not readily available in claims data, including information about incremental costs and equilibrium premiums under different regulatory regimes. Estimating incremental costs between plans would likely need to be done by making assumptions about the demand response of different individuals to alternative plan designs.[28] Evaluation of payment systems at the market design stage, prior to observing market equilibrium, or evaluating alternative payment systems different from the system currently in use also requires an estimate of the equilibrium incremental premiums, as these premiums are not

[28] Assumptions about demand response to metal tier plans in the U.S. Marketplaces are made prior to estimating the separate models for each metal tier.

observed. The analyst needs to make an assumption about what they would look like under the modified payment system rules.

Finally, individual health insurance markets are likely to have a number of types of plans. The expressions (5.6) and (5.7) can be generalized to multiple plan types, by, for example, measuring incremental costs for all plans relative to a base plan, but this might get complicated fast as the number of plan types proliferate. We recommend the analyst choose two plans, those representing the most important choice facing the largest body of consumers. In some markets, this will be natural – Medicare Advantage versus traditional Medicare, a Silver plan versus a Gold plan in the Marketplaces, a high versus a low deductible plan in the Netherlands – but in other settings, some consideration will be necessary to make the choice. Once the two plans are chosen, *incremental premiums* can be observed for each person. For two plans in the community-rated Netherlands, for example, this is simply the difference in the plans' premiums.

5.4.3 Second Source of Deviation from Efficient Pricing: Adverse Selection

The second source of deviation from efficient pricing is due to adverse selection. As we noted above in our discussion of Figure 5.2, with community rating (or, even with other limited premium categories), too few of the low-cost and too many of the high-cost members of an insurance pool will tend to choose Plan A, the more generous plan in our example. With the more generous plan drawing an adverse selection of the risks, it must price higher not just for its more generous coverage, but for the higher costs of the risks the plan attracts. Efficient sorting is promoted when Plan A prices higher for its more generous benefits -- this is the incremental plan cost standard. Efficient sorting is undermined when Plan A prices higher because it draws more expensive risks (Einav, Finkelstein and Cullen, 2010). Some of the highly developed theoretical and empirical literature on this subject is summarized in Textbox 5.2.

Textbox 5.2: Theoretical and Empirical Literature on Adverse Selection and the Inefficiency of Plan Pricing

Building on work by Cutler and Reber (1998), Einav and Finkelstein (EF) and their colleagues have proposed an elegant and influential model of sorting between two plan types (Einav and Finkelstein, 2011; Einav Finkelstein and Cullen, 2010). The population is ordered by their willingness to pay for the more generous plan. Adverse (or favorable) selection is related to the slope of the average cost curve (the average of the incremental costs as a function of incremental price of the more generous plan). With competition and average cost pricing, the more generous plan sets too high a price and too few consumers join the plan. This form of pricing with a feedback loop between selection and pricing can lead to the dreaded "death spiral" for the generous option in health insurance markets (Cutler and Reber, 1998).

With empirical estimates of the shapes of the demand and cost curves, the EF model can be used to estimate a welfare triangle related to the inefficiency of pricing due to adverse selection. The EF framework has been frequently applied to study premiums and efficient sorting of consumers among plans. For example, Hackmann, Kolstad and Kowalski (HKK, 2015) use the EF model to evaluate the welfare consequences of the Massachusetts health care reform of 2006, the precursor to the national reform. Kowalski (2014) applies the HKK version of the EFC model to estimate the welfare consequences of the implementation of the ACA.

EF-type models have also recently been applied to Medicare Advantage (MA). Cabral, Geruso, and Mahoney (2014) use a modified version of the EF framework to estimate the extent of selection into MA, finding little evidence of selection on the margin. Curto et al. (2014) also estimate important structural elements of demand and cost using changes in MA premiums over time, again finding little evidence of selection into MA at the margin (as premiums move up and down) but on average, costs were lower in MA, even after risk adjustment, by 2-3%. More recently, Glazer and McGuire (2017) use the EF conceptual framework to derive the implications for setting the level of subsidy to MA plans.

Note that the EF model is concerned with the second form of pricing problem, that due to selection, and not the first, due to limited premium categories. Welfare losses estimated with the EF framework thus are only a partial measure of the welfare loss from inefficient premium pricing. Bundorf, Levin, and Mahoney (2012) and Geruso (2017) use a more general framework to study the interaction between premium regulation and selection, highlighting the issues we discuss here.

We do not anticipate that policy researchers will have, *ex ante*, a measure of the degree of adverse selection in the relevant individual health insurance market. Rather than basing a

measure on the extent of the problem (which will typically not be known), we base our measure on the degree to which the payment system, including risk adjustment and any risk sharing, addresses the problem. Risk adjustment, and other plan payment features such as reinsurance, transfer funds from plans attracting healthier enrollees to plans attracting sicker, more expensive enrollees. This transfer requires the plans with the healthier pool (likely the plan with lower premiums/less generous coverage) to raise its premium and enables the plans with the sicker pool (those with more generous coverage) to lower their premiums. Risk adjustment transfers thus counteract the adverse selection effect, raising the incremental premium in the less generous plan and reducing the degree to which the incremental premium difference is affected by selection.

The previously discussed Payment System Fit (PSF) from (5.2) is a suitable measure of how much the payment system contributes to blunting the problems adverse selection causes for premium setting. Our approach here is to note that there is a second reason to be interested in PSF: it is also a good measure of how the payment system contends with adverse selection and premium pricing.

5.4.4 Summary Comments on Measures of Incentives for Efficient Sorting

The issue of the efficiency of plan sorting has different importance in different institutional settings. The underlying problem of efficient sorting is less salient in health insurance markets where the differences among the plans are small, as for example in Germany where plans have the same regulated benefit package and do little in terms of selective contracting or managed care. In such a setting, the fairness associated with community rating comes at a small cost in inefficient sorting. In other setings, however, available plans differ considerably, and there the efficiency cost of the fairness of community rating is higher. In some cases some fairness can be maintained even with efficient incremental pricing. Textbox 5.3

explains this with an example showing the fairness improvement with some risk adjustment and risk rating only the incremental premium.

Textbox 5.3: Risk Rating of Incremental Premiums Can Combine with Fairness-Related Subsidy of Higher-Cost Groups

Suppose the population is half young and half old, with cost structure in Plans A and B as follows:

	Costs	
	Young	Old
Plan A	100	200
Plan B	150	300

Risk rating of Plan A and B premiums would lead the old to have to pay more than the young for both Plan A (200 versus 100) and for Plan B (300 versus 150). This would lead to efficient sorting of both groups between the plans because the incremental premium for the young to join Plan B (50 =150-100) and for the old (100 = 300-200) is equal to the incremental cost for each group. It may be regarded to be unfair that the old pay more than the young.

An alternative is to make Plan A free to both groups (preserving fairness). The regulator then pays Plan B 100 for every young that joins and 200 for every old (these are the group costs in Plan A). If we then allow Plan B to risk rate, the young will be charged 50 for Plan B and the old will be charged 100, ensuring efficient sorting.

5.5 Measures of Incentives for Cost Control

An important objective of managed competition is conveying incentives to health plans to control costs. Making a fixed payment to a plan per person per month is intended to do just that. It is clear, as we discussed in Chapter 4, that risk sharing in plan payment with, for example, reinsurance, reduces a plan's incentives to control costs. Traditionally, diagnostic-based risk adjustment has not been seen as also sacrificing incentives for cost-control incentives, under the

premise that risk adjustment compensates for patient characteristics rather than services provided (Pope et al., 2011). This position, however, is not correct. In risk adjustment formulas with diagnosis-based risk adjustment, plan revenues are *not* independent of cost. After a very brief reminder about risk sharing and incentives for cost control, this section explains why diagnosis-based risk adjustment also dilutes incentives, and then proposes a metric, incorporating the effects of both risk sharing and risk adjustment, for measuring the degree to which a payment system deviates from full incentives to control costs.

5.5.1 Risk Sharing Affects Incentives for Cost Control

Section 4.4.1 was titled, "The Share of Dollars/Euros Touched by Risk Sharing Measures Incentives Affected," telling the story of how the effect of risk sharing on incentives can be measured. For example, a common form of risk sharing is reinsurance. If the reinsurance attachment point implies 5% of the costs are above the attachment point, and if the reinsurance share is 80%, the share of dollars/euros touched by risk sharing is 80% of 5%, or 4%. As we say in Section 4.4.1, the idea of this simple measure is that a plan would have incentives to reduce costs to the degree that it was responsible for those costs, and it is responsible in this example for 96% of the costs, so would retain, with this reinsurance example, very strong incentives to contain costs.

5.5.2 Diagnosis-Based Risk Adjustment Affects Incentives for Cost Control

It is important to dispel the belief that paying plans by health status risk-adjusted capitation leaves cost-control incentives unaffected. The belief would hold true if the capitation payment were based on age and gender (or other characteristics independent of utilization), but predominantly, risk classification systems are based on diagnoses that emerge with health care treatments. In practice, the conditions used to determine risk adjustment are established during

provider-patient interactions in which a claim is generated. For example, in the CMS-HCC system paying Medicare Advantage plans in the U.S. Medicare system, a single physician office visit at which a patient receives a new diagnosis of "diabetes without complications" changes a patient's risk score and results in an additional payment of approximately $1,500 annually. The visit generating the diagnosis, and the follow-up events the visit triggers such as further diagnostic testing, are components of cost to the plan, creating a link between payments a plan receives from risk adjustment and the plan's realized costs. Thus, utilization affects both costs and risk-adjusted payments, implying that insurers are compensated at least in part for their patients' utilization. The diabetes case is of course not an isolated example; diagnoses emerge only in the course of diagnostic visits or treatment so that plans are paid more when new diagnosis-generating health care takes place. This is true in concurrent risk adjustment systems, such as in the U.S. Marketplaces, and in prospective systems used in most other settings in which diagnoses are from a previous period, so long as the enrollee has some likelihood of remaining enrolled in the plan.

The connection between a prospective capitation payment and plan spending is obvious in systems that use some explicit indicator of costs as a risk adjustor. As noted in Chapter 4, Van Kleef and Van Vliet (2012) compared a form of high-risk pooling with the option of including "membership in the high-risk group" as a risk adjustor in the Dutch equalization model. Presently, the Dutch system includes risk adjuster variables that provide insurers with additional compensation for consumers who were in the top-15% of the spending distribution in each of the three preceding years. As long as there is some continuity in plan enrollment, it may be financially attractive for plans to provide extra care to individuals in order to induce them to exceed the cost threshold and assure assignment to the high-risk group.

The degree to which cost-containment incentives are affected by health-status based risk adjustment can be measured (and ultimately compared to the incentive of other plan payment options such as risk sharing). Building on analysis of the incentive effects of hospital "prospective payment" by McClellan (1997) who measured the *de facto* risk sharing in the DRG payment system to hospitals, Geruso and McGuire (2016) in the U.S. Marketplaces and Schmid and Beck (2016) in Switzerland measure the *de facto* cost sharing in risk adjustment payment systems by simulation methods. These papers ask: suppose some component of health care were not provided to a patient during a year. How much would this affect the payment the plan receives for the person? Averaged over the experience of a group of enrollees, the payment reduction associated with the cost reduction describes the portion of the costs shared by the regulator (or by the market depending on the risk sharing modality). In a fully cost-based system, payments to the plan would go down one-for-one with any reduction in services. In a pure prospective system (where, for instance, capitation was based only on age and gender as in Israel), payments would not fall at all and there would be no risk sharing. But in capitation payment systems using health status indicators, for some patients randomly removing certain office visits (for example) along with their costs and any diagnoses generated in that visit does lead to a payment reduction, indicating and measuring the degree of *de facto* risk sharing built into the capitation payment system.

Schmid and Beck (2016), for Switzerland, find that the de facto risk sharing in the Swiss risk adjustment model is .09, meaning that on average 9% of plan spending is returned in the form of higher payments given the risk adjustment model.[29] Reinsurance with a cut off of 60,000 CHF and a reinsurance rate of 80% increases the return to about 17% on average. This

[29] Schmid and Beck (2016) report the "power" of the payment models which is 1 minus the de facto cost sharing.

8% additional reduction in cost control incentives, however, boosts the fit of the payment model at the individual level from .15 to .41 (Schmid and Beck (2016, Table 1)), serving as a reminder of the essential tradeoff between reducing incentives for risk selection against the loss of incentives for cost control in having costs drive revenues.

"Upcoding" of diagnoses, another persistent problem with risk adjusted payment systems, is another indicator of the incentive effects of diagnosis-based risk adjustment. Upcoding takes place not only when plans change codes in isolation, but decide to "do stuff" in order to generate codes, such as engaging in home-based diagnostic visits. Under the prospective health status-based risk adjustment model in Medicare Advantage, an individual generates a risk score in a Medicare Advantage plan that is about 6-7% larger than the risk score the same person would generate in Traditional Medicare by some combination of more services and simple upcoding. (Geruso and Layton, 2017)

5.5.3 Defining Incentives for Cost Control: The Power of a Plan Payment System

We base our measure on the concept of *power* as the term is used in contract theory, the share of costs at the margin born by the health plan.[30] Power in health insurance contracts is tightly linked to the goal of cost control, as it describes the payment system's impact on the insurer's marginal incentive to limit healthcare spending. Power characterizes how a plan's expenditures impact a plan's net payment from the regulator. This connection is obvious with risk sharing features of plan payment and present but not-so-obvious when it comes to risk adjustment. Our definition and method of operationalizing power is intended to expose the full incentives in a payment scheme.

[30] Power is maximized with a fixed price contract and decreases as the price is tied to realized costs. See Laffont and Tirole (1993, p. 11).

If an insurer's payment R_i is invariant to changes in realized costs Y_i, as it would be in a plan paid by an age-gender only risk adjustment system, the power of the payment system would be at the maximum of 1.0. That is, the share of costs born by the plan at the margin is 1.0. Conversely, in a cost-based system where payment tracked costs exactly, the power would be 0. Away from these polar cases of payment systems, the change in payment for a person with respect to a change in cost for a person could vary over people, vary over ranges of cost, and vary over types of services. For example, the first health care event in a diagnostic area will trigger higher payment, but subsequent ones may not.

Imagine a thought experiment in which 10% of each person's health care costs were reduced by randomly eliminating part of the use of that person during a year. Individuals' risk scores would fall and the revenue of the plan would go down by X%. The share of the cost reduction kept by the plan, the power of the payment system, is $1 - \frac{X}{10}$. More formally, we define power as:

$$\text{Power} \equiv 1 - \frac{1}{N}\sum_i \frac{dR_i}{dY_i} \qquad (5.8)$$

where N is the number of enrollees in a plan, and $\frac{dR_i}{dY_i}$ is the derivative of payment for person i with respect to a marginal change in their utilization (Geruso and McGuire 2016).

5.5.4 Measuring Power in Plan Payment Systems

Power, as defined in (5.8), has been measured in research studies in Switzerland and in the U.S. Marketplaces, but at present, the technology of power measurement is not easily applied by policy researchers. It is likely that the simulation methods proposed by Geruso and McGuire (2016) will be refined (or replaced) as research continues. In the meantime, policy researchers can put to use the basic findings from the research literature.

The power of a prospective system is greater than a concurrent system because diagnoses given last year are less predictive of costs this year than diagnoses received this year. Roughly speaking, Geruso and McGuire (2016) find that the power of the HCC-based concurrent risk adjustment system is about 70%, and the power of the same diagnostic system when applied prospectively is about 80%.[31] High year-to-year turnover does not affect power in a concurrent system (because the past doesn't matter), yet strengthens the power of a prospective system (because a diagnosis made for a person last year returns nothing to a plan if the person leaves the plan). The 70% power for a concurrent system and 80% power for a prospective system (before figuring in turnover) are the best power numbers for the two types of HCC-based systems, in our view, based on current research. We look for both more conceptual and empirical research to refine these estimates and to extend them to other institutional settings and risk adjustment models.

A complete power analysis takes account of any risk-sharing features of the plan payment model. For example, suppose the policymaker was considering a prospective system with a reinsurance component that affected 5% of total costs. We have argued above that the power loss from such a risk-sharing policy is 5%. Prospective risk adjustment plus the reinsurance policy would have a power of 75% (80%-5%). This metric can be useful in the following way: based on the (limited) research literature, the prospective system plus reinsurance has greater power than a concurrent system (with no reinsurance). Policy researchers can readily compare the fit at the individual and group level of the two alternatives. If the fit is at least as good in the

[31] Geruso and McGuire (2016) in Table 2 report power for specific disease areas, and then a range for inpatient and outpatient overall. Our 80% and 70% are summary numbers not found in the Geruso and McGuire table.

prospective plus reinsurance system, the policymaker can obtain better fit with more power in the prospective plus reinsurance system than in a concurrent risk adjustment system.[32]

5.5.5 Final Comments on Measuring Incentives for Cost Control

The main message of this section is that diagnosis-based risk adjustment systems, prospective as well as concurrent, weaken incentives for cost control. A patient must have a medical encounter in which a diagnosis is made in order to turn on the diagnosis flag in a risk adjustment model. The higher revenue associated with the appearance of the diagnosis rewards the encounter where the diagnosis takes place, weakening incentives to control costs. This is true both for a concurrent risk adjustment system where the plan is paid more this year if the flag goes on, and for a prospective risk adjustment system, where the plan is paid more next year if the patient stays in the plan.

Once this point is accepted, it becomes a matter of degree to which a particular plan payment system maintains the power of cost containment incentives. We have proposed a way for analysts to conduct such an assessment without need to undertake extensive new research. We recognize the "evidence base" for the power of alternative systems is thin. The simulation methods used to assess power in the few papers doing so also need reconsideration and refinement. Much more work is needed on alternative systems in different settings in order to gain an appreciation of how payment models affect incentives, and ultimately to quantify the fundamental tradeoff between incentives for cost control and incentives related to selection that have been recognized to be the fundamental issues in plan payment design (Newhouse 1996).

[32] This is the comparison made by Geruso and McGuire (2016) in the context of U.S. Marketplaces. There a prospective HCC-based system with Marketplace reinsurance policy fits better than concurrent risk adjustment (as memasured by Payment System fit) and preserves higher power.

5.6 Summary and Discussion

In a nutshell, this chapter intends to equip researchers and regulators with a toolkit for practical *ex ante* evaluation of health plan payment systems. We hope to advance the field by proposing simple modifications of currently used approaches that, in many circumstances, do a better job than conventional metrics of measuring incentives for efficiency.

5.6.1 When Does It Matter to Make Use of the Proposed Metrics?

Table 5.2 briefly summarizes our views about when our proposed metrics will be more informative than existing metrics. With respect to both individual and group-fit measures, using predictions from a risk adjustment regression to compute R-squared and conventional measures of over/undercompensation is fine if the payment system is fully described by those predictions. We regard this to be essentially true in Germany and Medicare Advantage in the U.S., but not elsewhere. In Switzerland, risk sharing figures into plan payment; in the Netherlands there are four predicted values that need to be aggregated to describe plan payment; in Marketplaces, premium categories and reinsurance features play a role. Payment System Fit is what is called for generally, and in special circumstances this will be approximated by regression predicted values.

[Insert Table 5.2 here]

We proposed two metrics related to consumer sorting. Our first, based on the gap between incremental premiums and incremental costs, is problematic from a practical standpoint in many settings, and may not be of interest in a context where policymakers are committed to community rating. Our second measure, capturing the inefficiency in pricing and sorting caused

by adverse selection, is simply the individual-level fit measure, Payment System Fit. A necessary condition for either of our sorting metrics to be worthwhile is that there are some meaningful differences in plans that affect their cost. This latter condition is true in most settings but not all (again, Germany is in the minority).

Measuring power is important when alternatives being considered might differ in the degree to which they affect cost-control incentives. For example, adding diagnoses from outpatient claims in a prospective risk adjustment model will affect power. Adding a risk sharing feature will affect power.

5.6.2 Mathematical Properties of the Measures

We close this chapter with a couple of brief reminders about the mathematical properties of the measures proposed.

First, we present linear and quadratic forms of measures of fit of plan payments to plan costs at the individual and group level, and linear and quadratic forms of the fit of incremental premiums to incremental costs. Our power measure is simply linear. Though we lean towards the quadratic forms on the basis of the general property of the increasing economic harm from price distortions, the analyst needs to make a choice about whether they would regard the linear or the quadratic (or some other power) most helpful.

Second, in the quadratic form, all of our measures are between 0 and 1. They are unit-free numbers that cannot be added. The measures are useful for identifying potentially dominant policies. In the example discussed above, if prospective risk adjustment plus a small amount of reinsurance yields better Payment System Fit and higher power, it can be regarded as superior to the alternative of concurrent risk adjustment. The measures could be useful for designing a policy that is equivalent to another policy in one dimension so as to focus policy choice on the

other. Sticking with the same example, our method for assessing power would allow the analyst to identify the degree of reinsurance that, when paired with prospective risk adjustment, yields the same power as concurrent risk adjustment. With this in place, the analyst can compare various fit properties of the models.

The metrics cannot be added (this would be a meaningless number) nor, in the presence of tradeoffs – ie one payment system is better on one metric but worse on another – can the metrics value the tradeoff involved. A .01 change in power cannot be compared with a .01 change in Payment System Fit. This comparison must be based on the values of the decisionmaker.

References

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.

Bauhoff S., Fischer, L., Göpffarth, D., Wuppermann, A.C., 2017. Plan responses to diagnosis-based payment: Evidence from Germany's morbidity-based risk adjustment. Journal of Health Economics.

Beck, K, Trottmann, M & Zweifel, P 2010 'Risk Adjustment in Health Insurance and its Long-term Effectiveness', Journal of Health Economics, vol. 29(4): 489-98.

Breyer, F., Heineck M. & Lorenz, N., 2003, "Determinants of Health Care Utilization by German Sickness Fund Members – with Application to Risk Adjustment," Health Economics 12(5): 367-76.

Buchner, F, Göpffarth, D & Wasem, J 2013, 'The new risk adjustment formula in Germany: Implementation and first experiences', Health Policy, vol. 109, pp. 253-62.

Bundorf, M. K, J. D. Levin and N. Mahoney. (2012) "Pricing and Welfare in Health Plan Choice," American Economic Review 102(7): 3214-3248.

Cabral, M., Geruso, M., and Mahoney, N. (2014) "Does privatized health insurance benefit patients or producers? Evidence from Medicare Advantage." NBER Working Paper 20470.

Carey, C. (2017a). "Technological Change and Risk Adjustment: Benefit Design Incentives in Medicare Part D," American Economic Journal: Economic Policy, 9(1): 38-73.

Carey, C. (2017b) "Time to Harvest: Evidence on Consumer Choice Frictions from a Payment Revision in Medicare Part D," Working Paper. Retrieved from https://drive.google.com/file/d/0B2TeS7lispKBQ21RLV9PckE2eWs/view.

Curto, V., Einav, L., Levin, J. and J. Bhattacharya, (2014) "Can Health Insurance Competition Work? Evidence from Medicare Advantage," National Bureau of Economic Research Working Paper 20818, December.

Cutler D.M. and Reber S.J. (1998) "Paying for Health Insurance: The Tradeoff between Competition and Adverse Selection," The Quarterly Journal of Economics 113(2): 433-466.

Douven, Rudy, R. Ron van der Heijden, R., McGuire, T., and Schut, E. (2017) "Premium Levels and Demand Response in Health Insurance," National Bureau of Economics Research Working Paper 23846.

Eggleston K. and Bir A. (2009), "Measuring Selection Incentives in Managed Care: Evidence from the Massachusetts State Employees Insurance Program," Journal of Risk and Insurance 76: 159-175.

Eijkenaar F, R.C.J.A. van Vliet, and R.C. van Kleef. (2017). "Diagnosis-based Cost Groups in the Dutch Risk-equalization Model - Effects of Clustering Diagnoses and of Allowing Patients to be Classified into Multiple Risk-classes", Medical Care, forthcoming.

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

Einav, L., Finkelstein A. and Levin, J. (2010) "Beyond Testing: Empirical Models of Insurance Markets," Annual Review of Economics 2:311-36.

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., Zhu, W., 2017b. Demand elasticities and service selection incentives among competing private health plans. Journal of Health Economics. December
Ellis, R.P. & McGuire, T.G., 2007. Predictability and Predictiveness in Health Care Spending. Journal of Health Economics 26 (1), 25–48.
Enthoven A.C. (1993) The History and Principles of Managed Competition, Health Affairs 12: 24-48.
Ettner, S., Frank, R., McGuire, T. and Hermann, R. (2001), "Risk Adjustment Alternatives in Paying for Behavioral Health Care Under Medicaid," Health Services Research 36(4): 793-811.
Frank R.G., Glazer J. and McGuire T.G. (2000) "Measuring Adverse Selection in Managed Health Care," Journal of Health Economics 19(6): 829-854.
Geruso, Michael. (2017) "Demand Heterogeneity in Insurance Markets: Implications for Equity and Efficiency." Quantitative Economics, forthcoming.
Geruso M. and Layton, T. (2017) "Selection in Insurance Markets and Its Policy Remedies" Journal of Economic Perspectives. 31(4): 23-50.
Geruso, M., Layton, T., 2015. Upcoding: Evidence from Medicare On Squishy Risk Adjustment. NBER Working Paper 21222.
Geruso M., Layton, T., and Prinz D. (2016) "Screening in Contract Design: Evidence from the ACA Health Insurance Exchanges," NBER Working Paper 22832.
Geruso, M. and 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. and McGuire, T.G. (2017), "Paying Medicare Advantage Plans: To Level or Tilt the Playing Field," Journal of Health Economics December.
Han, T. & Lavetti, K. 2017. "Does Part D Abet Advantageous Selection in Medicare Advantage?" Journal of Health Economics, December.
Handel B., Hendel I. and Whinston M.D. (2015), "Equilibria in Health Exchanges: Adverse Selection vs. Reclassification Risk," Econometrica 83(4): 1261-1313.
Hileman, G., Steele, S., 2016. Accuracy of Claims-Based Risk Scoring Models. Society of Actuaries.
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.
Keeler, E., Carter, G. and Newhouse, J. (1998), "A Model of the Impact of Reimbursement Schemes on Health Plan Choice," Journal of Health Economics 17(3): 297-320.
Kuziemko I, Meckel K, Rossin-Slater M. (2014) "Do Insurers Risk-Select Against Each Other? Evidence from Medicaid and Implications for Health Reform." NBER Working Paper No. 19198.
Laffont, J-J., Tirole, J., 1993. A Theory of Incentives in Procurement and Regulation, The MIT Press.

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

Lorenz, N., 2015: The interaction of direct and indirect risk selection, Journal of Health Economics, 42, 81-89.

McClellan, M. (1997). Hospital reimbursement incentives: An empirical analysis. Journal of Economics and Management Strategy, 6(1), 91-128.

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.

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. December.

Pauly M. (2008) "Adverse Selection and Moral Hazard: Implications for Health Insurance Markets," Incentives and Choice in Health Care, Sloan F. and Kasper, H. (eds) Cambridge, MA: MIT Press, 2008

Pope, G. C., Kautter, J., Ingber, J.J., Freeman, S., Sekar, R., Newhart, C. (2011), Evaluation of the CMS-HCC Risk Adjustment Model, In Final Report, RTI Project Number 0209853.006, RTI International, March.

Schmid, C & Beck, K. 2016. "Reinsurance in the Swiss Health Insurance Market: Fit, Power and Balance," Health Policy. 120(7): 848-55.

Shen, Yujing, and Ellis, Randall P. (2002) "Cost minimizing risk adjustment" Journal of Health Economics. 21(3): 515-530.

Shepard, M. (2016) "Hospital Network Competition and Adverse Selection: Evidence from the Massachusetts Health Insurance Exchange," Harvard University. unpublished.

Shmueli, A., Messika, D., Zmora, I. and Oberman, B. (2010) "Health Care Costs During the Last 12 Months of Life in Israel: Estimation and Implications for Risk Adjustment," International Journal of Health Care Finance and Economics 10(3): 257-73.

Spinnewijn, J. (2017). "Heterogeneity, Demand for Insurance, and Adverse Selection." American Economic Journal: Economic Policy, 9(1): 308-43.

Van Barneveld, E.M., Lamers, L.M., R.C.J.A. van Vliet and W.P.M.M van de Ven, (2000) Ignoring Small Predictable Profits and Losses: A New Approach for Measuring the Incentives for Cream Skimming, Health Care Management Science. 3 131-140.

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 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., McGuire, T.G., Van Vliet, R.C.J.A., Van De Ven, W.P.M.M., 2017. Improving Risk Equalization with Constrained Regression. The European Journal of Health Economics 18(9): 1137-1156.

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., Van Vliet, R.C.J.A. and Van de Ven, W.P.M.M., 2013. "Risk equalization in the Netherlands: An empirical evaluation", Expert Review of Pharmacoeconomics & Outcomes Research , 13(6), 829-839

Van Kleef, R.C., van Vliet, R.C.J.A. and van de Ven, W.P.M.M. (2013) "Risk Equalization in The Netherlands: an Empirical Evaluation," Expert Review of Pharmacoeconomic Outcomes Research 13(6): 829-39.

Van Veen, S.C.H.M. , Van Kleef, R.C., Van de Ven, W.P.P.M. &. van Vliet, R.C.J.A. 2014. "Improving the prediction model used in risk equalization: cost and diagnostic information from multiple prior years", The European Journal of Health Economics, 16: 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., 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.

Veiga, André, and E. Glen Weyl. 2016. "Product Design in Selection Markets." Quarterly Journal of Economics, 131(2): 1007–1056.

Einav, L, Finkelstein A, and M.R. Cullen (2010), "Estimating Welfare in Insurance Markets Using Variation in Prices," Quarterly Journal of Economics, 125(3): 877-921.

Layton, T., R. Ellis and T. McGuire (2015), "Assessing Incentives for Adverse Selection in Health Plan Payment Systems," NBER Working Paper 21531.

Newhouse, J.P., W.G. Manning, E.B. Keeler and E.M. Sloss (1989), "Adjusting Capitation Rates using Objective Health Measures and Prior Utilization," Health Care Financing Review 15(1): 39-54.

Van de Ven, W.P.M.M., and R. P. Ellis, (2000) "Risk Adjustment in Competitive Health Plan Markets," in A. Culyer and J. Newhouse (eds.), Handbook of Health Economics, Volume 1, Elsevier, pp. 755-846.

Zhu, J., T. Layton, A. Sinaiko and T. McGuire (2013) "The Power of Reinsurance in Health Insurance Exchanges to Improve the Fit of the Payment System and Reduce Incentives for Adverse Selection," Inquiry 50(4): 255-74.

Table 5.1: Metrics for Evaluating Efficient Performance of Health Plan Payment Systems

Section of Chapter: Dimension of efficiency	Traditional Metric	Modified/New Metric
5.2: Selection: Fit at the Individual Level	R-squared or other fit statistics from a risk-adjustment regression	Payment System Fit using predicted payments (not regression predictions)
5.3: Selection: Fit at the Group or Action Level	Predictive ratios (U.S.); under and overcompensation (Europe)	Group Payment System Fit; measure for potential plan actions
5.4: Selection: Individuals Choose the Right Plan	No incentive metric in common use	Summary measure of the gap between efficient incremental premium and the actual premium; Payment System Fit
5.5: Incentives for Cost Control	No incentive metric in common use	Power of the payment system

Table 5.2: Circumstances in Which Proposed Metrics Do Better at Measuring Incentives for Efficiency

Section of Chapter: Issue	Proposed/Traditional Metrics	Circumstances Where New Metrics Called For
5.2: Selection: Fit at the Individual Level	Payment System Fit preferred to R-squared from a risk-adjustment regression	Payment system includes other elements than predictions from a single risk adjustment model
5.3: Selection: Fit at the Group or Action Level	Group Payment System Fit preferred to predictive ratios or under and overcompensation	Payment system includes other elements than predictions from a single risk adjustment model
5.4: Selection: Individuals Choose the Right Plan	New metric: gap between incremental premiums and incremental costs; Payment System Fit	Major differences in characteristics and premiums of plans available to consumers, and, changes in premium regulation are under consideration; a second reason to recognize Payment System Fit
5.5: Incentives for Cost Control	New metric: power of a payment system (in retaining incentives for cost control)	When diagnosis-based risk adjustment, risk sharing, or possible major change in encounter-based variables in model are under consideration

Figure 5.1

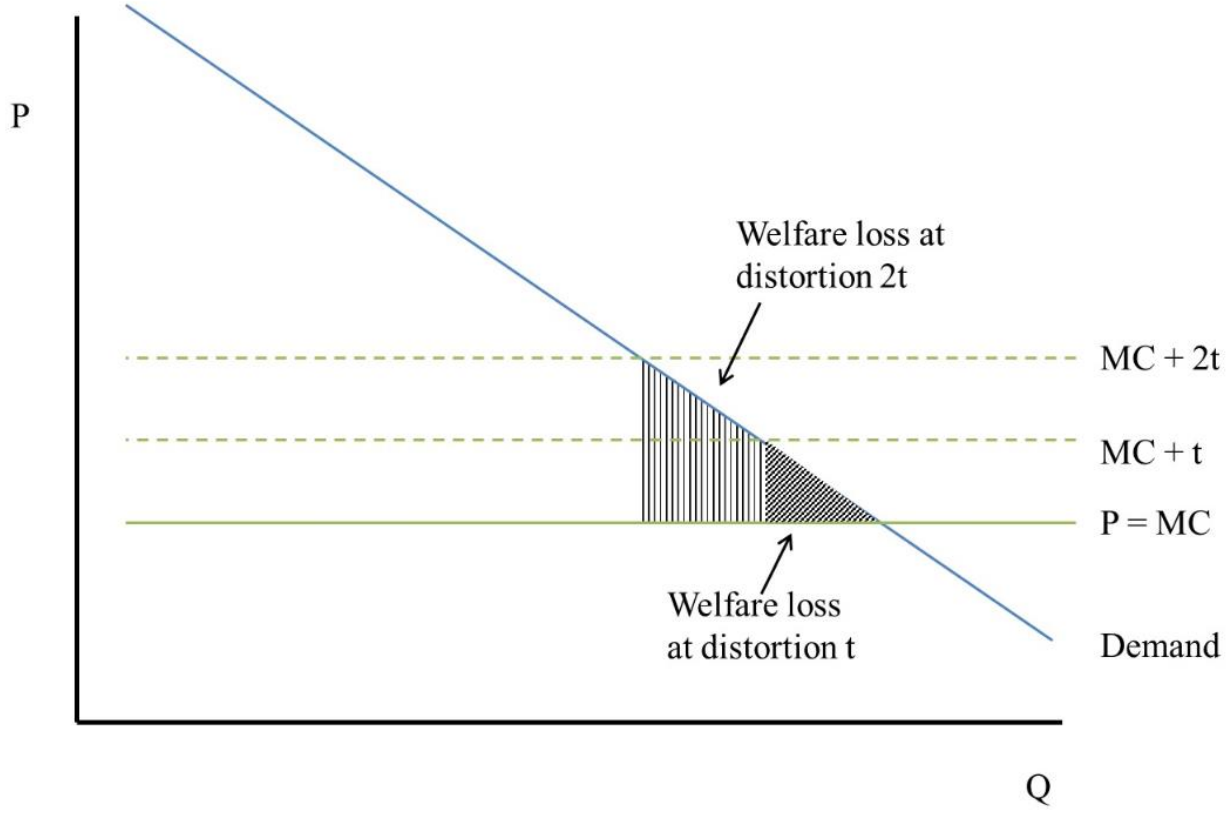


Figure 5.2

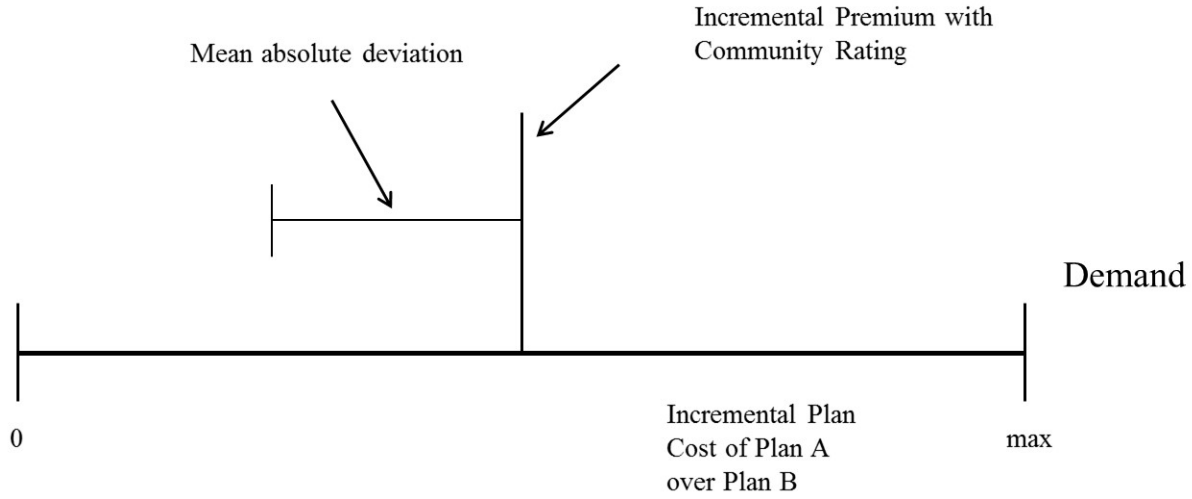
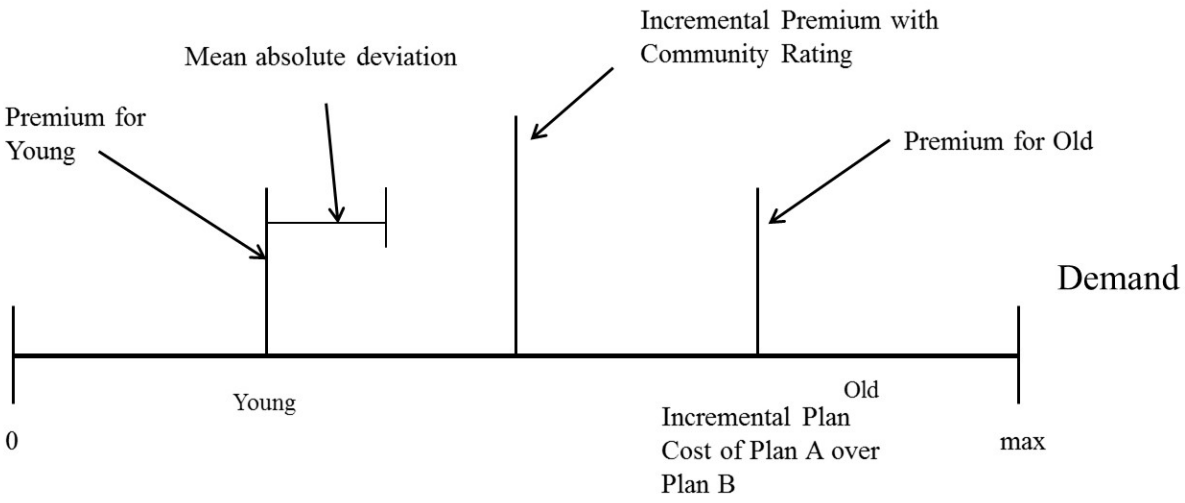


Figure 5.3


Incremental Premium with Community Rating
Mean absolute deviation
Premium for Young
Premium for Old
Demand
Young
Old
0
Incremental Plan Cost of Plan A over Plan B
max