2018

Nonlinear Physician Performance Contracts

Amelia Mcfadden Bond
University of Pennsylvania, ambond@wharton.upenn.edu

Nonlinear Physician Performance Contracts

Abstract

Policymakers have increasingly focused on the design of provider contracts to reduce health care costs and increase care quality. Many of these contracts provide bonus payments to providers contingent on meeting externally set performance threshold levels. Using data from a large insurer in Hawaii, this paper estimates physician responsiveness to two features of these contracts - 1) threshold level and 2) bonus amount - for ten preventative process measures. I estimate provider performance response using a large discrete change in a single measure's threshold level and bonus amount during the sample period. I also estimate a pooled provider performance response across all measures using two instrumental variables. I find that a one percentage point increase in threshold location leads to a 0.3 to 0.5 percentage point increase in performance the subsequent quarter. I do not detect an average response to bonus size. Next I explore mechanisms for these responses. I find heterogeneous responses based on prior performance: low performing physicians are more responsive to threshold level, and high performing physicians are responsive to bonus amount. I do not find evidence for physicians increasing effort near the end of a time period alleviating concerns over decreased access for other types of patients. Finally, I find some evidence that the additional patients screened are higher risk and higher cost suggesting the marginal benefit of screening increases. My results demonstrate that the marginal bonus amount has little effect on provider effort and incentivizes already high-performing physicians. Small increases in threshold levels improves performance without increasing cost. These results have implications for innovations in physician payment models and contract designs.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Health Care Management & Economics

First Advisor

Ashley Swanson

Keywords

Financial incentives, Health care, Nonlinear contract

Subject Categories

Economics | Health and Medical Administration

NONLINEAR PHYSICIAN PERFORMANCE CONTRACTS

Amelia M. Bond

A DISSERTATION

in

Health Care Management and Economics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2018

Supervisor of Dissertation

Ashley Swanson, Assistant Professor of Health Care Management

Graduate Group Chairperson

Catherine M. Schrand, Celia Z. Moh Professor, Professor of Accounting

Dissertation Committee

Guy David, Gilbert and Shelley Harrison Associate Professor of Health Care Management

Hanming Fang, Class of 1965 Term Professor of Economics

Amol Navathe, Assistant Professor of Medical Ethics and Health Policy

NONLINEAR PHYSICIAN PERFORMANCE CONTRACTS

For my parents

ACKNOWLEDGEMENT

This dissertation would not have been possible without the support from many mentors, peers, friends, and family. First, I want to thank my committee. I thank Ashley for all of the early morning meetings, for pushing my ideas further, and for her constant support. I thank Guy for his guidance throughout my entire time at Wharton. Guy was always willing to give hours to talk through any topic - course work, paper ideas, or job market preparation. I thank Hanming for his ability to isolate parts of a model or empirical section that could readily be expanded. I thank Amol for including me so fully in the many projects and for his equal enthusiasm for physician payment topics.

I also thank Hawaii Medical Service Association (HMSA) for generously allowing me to use their data for this project. I also want to specifically thank Kim and Emily with HMSA as well as Amanda, Lin, and especially Kristen at Penn for their detailed answers to my many questions.

Heartfelt thanks to my fellow graduate students for helping me keep a healthy work-life balance, especially Susan, Emma, Ambar, Vicki, Bill, Kyle, Matt, Madison, Amanda, and Ben. A special thank you to Joanne for her help navigating through this program.

Finally, I want to thank my family for their encouragement and acceptance of my insistence to stay in school for so long and Rachel for her presence, wisdom, and love.

ABSTRACT

NONLINEAR PHYSICIAN PERFORMANCE CONTRACTS

Amelia M. Bond

Ashley Swanson

Policymakers have increasingly focused on the design of provider contracts to reduce health care costs and increase care quality. Many of these contracts provide bonus payments to providers contingent on meeting externally set performance threshold levels. Using data from a large insurer in Hawaii, this paper estimates physician responsiveness to two features of these contracts - 1) threshold level and 2) bonus amount - for ten preventative process measures. I estimate provider performance response using a large discrete change in a single measure's threshold level and bonus amount during the sample period. I also estimate a pooled provider performance response across all measures using two instrumental variables. I find that a one percentage point increase in threshold location leads to a 0.3 to 0.5 percentage point increase in performance the subsequent quarter. I do not detect an average response to bonus size. Next I explore mechanisms for these responses. I find heterogeneous responses based on prior performance: low performing physicians are more responsive to threshold level, and high performing physicians are responsive to bonus amount. I do not find evidence for physicians increasing effort near the end of a time period alleviating concerns over decreased access for other types of patients. Finally, I find some evidence that the additional patients screened are higher risk and higher cost suggesting the marginal benefit of screening increases. My results demonstrate that the marginal bonus amount has little effect on provider effort and incentivizes already high-performing physicians. Small increases in threshold levels improves performance without increasing cost. These results have implications for innovations in physician payment models and contract designs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction, background, and overview

1.1. Introduction

Rising U.S. health care costs represent an ever-pressing policy issue – the U.S. spent $3.2 trillion or almost 18 percent of its economy on health care in 2015 (Centers for Medicare and Medicaid Services (CMS), 2015). Furthermore, significant amounts of variation exist in spending and health outcomes across and within geographic markets. Health care systems, insurers and policy makers have long sought to decrease variation in total spending and have more recently also sought to decrease variation in health outcomes in part due to the increasing ability to collect outcome data. There are two general approaches to accomplish these goals: one approach is to focus on demand-side levers; the other is to focus on supply-side levers.

On the demand side, research has shown that brute-force cost containment methods, such as high deductible health plans, lead to a decrease in health care services, but also to a decrease in the use of services that are both superfluous and necessary (Lohr et al., 1986; Brot-Goldberg et al., 2017). Such measures ultimately miss the mark by decreasing spending at the expense of quality. One alternative to brute-force cost containment methods that is slowly gaining momentum is “value-based” insurance contracts, where consumers face a cost-sharing schedule based on the “value” of various services (Baicker et al., 2015). For example, consumers could face $0 or even a negative cost (they receive money) for preventative services whereas they would pay more for low or uncertain quality care. As these schemes are quite new, research has little to say, however one could imagine this scheme as at least partially filling the gaps of the brute-force methods.

On the supply side, which is where this paper focuses, brute-force cost containment methods have been used for years. Similar to demand-side methods, these approaches typically do not preserve quality levels and at times miss reductions in spending. At the federal level, Medicare has changed payment levels as well as altered the amount of risk placed on providers over time (alternatively known as the level of payment prospectivity). However, providers meet each change with behavior that is frequently counter to its purpose. For example, bundling all inpatient charges for one visit into a single payment known as Diagnosis Related Groups (DRGs) in the late 1980's resulted in shorter lengths of stays in hospitals and a shift from inpatient to outpatient care potentially decreasing quality in the short term (Coulam and Gaumer, 1992; Ellis and McGuire, 1993). When Medicare cuts payment for all services, the total number of services provided increases (Clemens and Gottlieb, 2014). When Medicare cuts payment for specific potentially over-provided services, physicians increase the provision of those same services if the income effect is large enough or substitute to different services when it is not (e.g., Yip, 1998; Gruber et al., 1999). The former effect both increases spending and decreases quality as the marginal patient is less likely to benefit from the service. The latter effect will typically have little impact on spending and uncertain quality effects depending on substitution patterns.

"Value-based" contracts also exist on the supply side and attempt to incentivize increased quality and decreased spending by offering a bonus payment conditional on supplying a number of certain services. For example, a physician receives a bonus for supplying a large enough number of high-value services such a preventative screenings or have a bonus withheld if she supplies too many low-value services such as MRIs for lower back pain.[1] Some contracts also include bonus dollars for meeting spending targets. Ideally "value-based" contracts increase the provision of high-value services, decrease the provision of low-value care, and reduce total spending. This dissertation focuses on changes in the quality of service provision and briefly comments on changes in total spending as the contract studied does not directly incentivize spending reductions.[2]

[1] Contracts generally do not directly change the prices paid for individual services. This is likely due to the fact that the underlying fee-for-service schedule does not align with current quality measures. First, low-value or high-value services are frequently defined for a specific type of patient rather than the entire population. For example, MRI payment should not change for all patients; payment should only change when a physician prescribes the intervention for a person with lower-back pain. Similarly, annual eye exams are only recommended for diabetic patients. Second, a fee-for-service payment is paid to the billing physician. In both cases noted above, the referring physician is responsible for generating the two services and should be directly incentivized, but the ophthalmologist or radiologist would be paid.

[2] Recent work has pointed to likely difficulty in increasing the quality of care along with decreasing

Iterations of payment schemes that change the price of services based on quantity have existed for decades under the name Pay-for-Value (P4V). These performance pay contracts are traditionally nonlinear, featuring thresholds where an agent receives a bonus payment only when their performance exceeds a specified level. Nonlinear pay contracts are theoretically suboptimal, yet are still frequently utilized for contacts in a variety of settings (Mirrlees, 1971). Studies of the introduction of P4V have consistently found mixed results and often found small or not statistically significant effects (Rosenthal et al., 2005; Eijkenaar et al., 2013; Greene et al., 2015). These studies identify average physician response to performance pay programs, but do not consider how specific contract features affect performance, potentially due to lack of detailed contract information (Rosenthal and Frank, 2006; Young et al., 2007a; Ryan and Damberg, 2013). Unlike previous work, this paper directly considers physician responses to these contract features.

This dissertation demonstrates the limited agent responsiveness to two key features in nonlinear contracts - a horizontal component (threshold level) and a vertical component (marginal bonus amount) - in a large performance pay program. Changes in the marginal bonus payment have little effect on provider effort. Small increases in thresholds improve performance without increasing cost. It is important to note that this work does not estimate performance without a contract, only performance with changes to a contract once it is in place. Additionally, all results are relevant to local changes in the contract features. Nevertheless, the setting is a highly powered performance pay scheme relative to others in the literature and should be viewed as an upper bound in physician responsiveness.

The rest of Chapter 1 provides an overview of performance pay contracts in health settings and discusses relevant literature. The chapter concludes by detailing the specific performance pay setting and defining the data. Chapter 2 is the main theoretical and empirical analysis of the dissertation. It begins with a theoretical model that motivates the dissertation's focus on the two contract features - a horizontal component (threshold level) and a

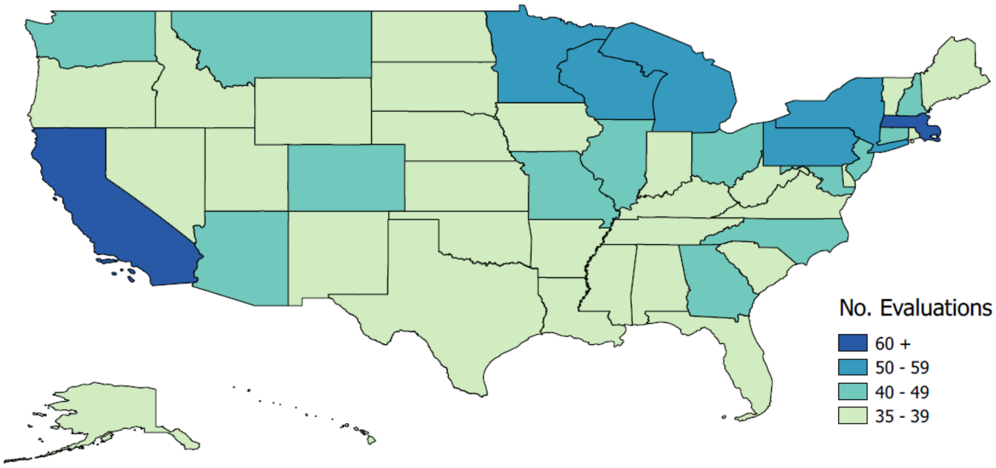
spending (e.g., Burns and Pauly, 2018)

vertical component (marginal bonus amount) - and provides rationale for the second empirical specification. Two sets of empirical analyses follow. The first set of analyses applies a difference-in-difference framework to two natural experiments that represent plausibly exogenous changes in, respectively, the threshold level and marginal bonus amount. The second set of analyses directly estimates physician response to changes in threshold levels and marginal bonus amount. To account for various sources of biases in direct estimation, I construct two instruments that leverage plausibly exogenous changes in a patient's performance measurement status. Physician level fixed-effect specifications are also considered.

Results from both sets of analyses align well. Physicians are responsive to changes in distance to a threshold. From the preferred direct estimation specification, an increase in thresholds by one percentage point results in quarterly performance increases of 0.5 percentage points. This estimate has a similar magnitude to results from the difference-in-difference analysis. Also, I find that on average physicians are not responsive to marginal bonus payment in the quarter after a payment change, again similar to the difference-in-difference results. In the difference-in-difference setting, the decrease in bonus pay represents a transfer of over one million dollars to bonus pay for other measures with no average impact on performance. Using the standard errors in the instrumented specification to bound the bonus amount finding, I cannot detect an increase in quarterly performance of 0.1 percentage points following a one percentage point increase in pay.

Chapter 3 examines various mechanisms for the results found in Chapter 2. I find that low performing physicians are more responsive to the threshold level, and high performing physicians are responsive to the marginal bonus amount. I explore whether physician effort changes over time as work in other nonlinear contract settings has found. I do not find evidence for physicians increasing effort prior to the end of a bonus time period, which alleviates concerns over negative access spillovers. Finally, I examine the risk and total spending of the additional patients who are screened as a proxy for the marginal cost and benefit of screening. I find suggestive evidence that the marginal patient is a higher risk and

Figure 1: Number of Performance Pay Evaluations by State



Notes: Adapted from McClellan et al. (2017).

cost patient, which positively suggests the marginal patient is both more costly to screen and receives a higher benefit from the screening. Finally, Chapter 4 concludes.

1.2. Performance pay contracts

Performance pay contracts between physicians and insurers are extremely common. While specific figures on the number of contracts or number of participating physicians are difficult to find, evaluations of individual contracts can be used to approximate their prevalence. Figure 1 presents the number P4V evaluations by state demonstrating that contracts are pervasive across the US particularly in California and Massachusetts. One of the largest performance pay programs, Medicare's Merit-based Incentive Payment (MIPS), recently began in 2017. This performance pay program is national in scope and the Center for Medicare and Medicaid Services (CMS) estimates 37% of all clinicians will participate in 2018 (Wilensky, 2018).

Though performance pay contracts commonly incentivize a large set of under-provided, high-value preventative services,[3] the average commercial patient was only receiving be-

[3] As recommended by the US Preventative Task Force.

tween 40 to 85% of those services in 2011, the beginning of the study period.[4] A large body of literature evaluates the introduction of performance pay programs intended to ameliorate this discrepancy (see Section 1.3.3 for a more detailed review). This literature produces mixed results, indicating small or not statistically significant physician response to performance programs, perhaps due to the small size of bonus payments. These studies identify average physician response to performance pay programs, but not in response to specific contract features (Young et al., 2007a; Rosenthal and Frank, 2006). Unlike previous work, this paper estimates physician responses to specific contract features, such that results can be applied to other settings including MIPS, and has the added advantage of a setting which contains meaningful bonus sizes.

Traditionally, studying performance pay in health care has been hindered by the complex nature of contracting - physicians are typically not directly employed and have contracts with many insurers. Tying bonuses to physicians' overall performance would require insurers' cooperation. Additionally, contract incentives are weakened by indirect bonus pay since insurer payments are frequently distributed to the physician group rather than an individual physician. Physician groups may not distribute bonus pay directly to physicians, weakening individual incentives and raising concerns of moral hazard (Gaynor and Gertler, 1995).[5] This dissertation circumvents these existing challenges by analyzing a performance pay system within an extremely concentrated insurance market and insurers pay bonuses directly to individual physicians.

Widespread adoption of performance pay has occurred in a range of industries outside of health care, including banking and manufacturing. Many performance pay contracts in these industries as well as in health care are nonlinear, featuring thresholds where an agent receives a bonus payment only when their performance exceeds a specified level. Paradoxically, nonlinear performance pay contracts are both theoretically and empirically

[4] Values are based on commercial Healthcare Effectiveness Data and Information Set (HEDIS) national percentiles in 2011.

[5] Few studies of performance pay schemes exist that directly pay physicians (e.g., Coleman et al., 2007; Rosenthal et al., 2008).

suboptimal, yet are still frequently utilized. Traditional models find nonlinear contracts suboptimal under the assumptions that effort is continuous, agents are sophisticated and more complex contract structures are costly to construct (Mirrlees, 1971).

In health care, nonlinear contracts are often rationalized by the desire of stakeholders including participating physicians to not reward low quality physicians.[6] Some nonlinear contracts also have numerous nonlinearities where the marginal bonus increases as the quantity of services increases, such as in Massachusetts' Alternative Quality Payment Contracts (Song et al., 2012). This increasing marginal bonus could be rationalized as accounting for the increasing marginal cost of seeing patients, however this feature is rarely incorporated as most programs have single thresholds. Previous research on salespeople has examined how nonlinearities in contracts affect the timing of effort (Oyer, 1998; Larkin, 2014), which this dissertation explores in Section 3.3. There is currently no empirical research of how nonlinear features affect responses to performance pay contracts in health care.

1.3. Previous literature

This dissertation is relevant to four strands of literature. The first is the theoretical contract literature space where I will focus on threats to optimal contracting that arise from moral hazard. The second is provider responses to financial incentives, which is specific to health economics and encompasses a mix of theoretical and empirical work. The third is the study of pay-for-value, which is purely empirical and primarily exists in the health services research domain. The final space is largely empirical and focuses on estimating parameters of interest in the presence of nonlinear budgets.

1.3.1. Contract Theory

Contract Theory literature typically focuses on principal-agent models with one of two threats to optimal contracting — moral hazard or adverse selection. The contract structure in this paper is required for all providers (i.e., no adverse selection) and compares individual

[6] During an interview with a member of the insurer's Payment Transformation team, the concern of providing bonuses to low quality physicians was voiced repeatedly.

providers to national benchmarks. Because of this framework, I will focus on the moral hazard literature with the assumption that asymmetric information arises after rather than before a contract is signed. Additionally, I will focus on instances when there are multiple tasks and no teams or multiple agents. Two models fit the empirical setting somewhat well. While both assume a linear payment scheme, which is not present in this setting, the models provide applicable insight.

Baker (1992) develops a model where a firm wants to optimally pay a worker based on productivity, but the worker's contribution to a firm, x, is imperfectly measured. The firm pays the worker salary and productivity components and must choose the weight between the two. Distortions result when a firm's observable measure y does not perfectly capture x. Baker's model implies that the weight a firm places on y depends on the alignment between the agent's marginal performance (y) with respect to effort and the firm's marginal value (x) with respect to effort. Note that the weight on y is not based on the alignment between x and y, but rather the marginal changes in these with respect to effort.

Holmstrom and Milgrom (1991) think instead about multiple efforts or tasks, x_i. If these tasks are independent, the model simplifies where effort and bonus for each task are decreasing in the marginal cost of effort, risk aversion and riskiness of output, x_i. If the tasks are not independent, then the weight placed on x_i's depend on the substitutability or complementarity between the tasks. If the two tasks are compliments, less weight can be placed on the x_i's. Conversely, if the two tasks are substitutes, more weight should be placed on the x_i's.

Both papers highlight important ongoing issues in the Pay-for-Value debate. Baker (1992) raises two points — the importance of a principal's ability to accurately measure what she values and the importance of placing weight on metrics where an agent is responsive. Parallels to P4V include whether contracts adequately measure value and include measures that are responsive to provider action. Holmstrom and Milgrom (1991) introduces the possibility of measures interacting and how weights should respond to substitution patterns.

This is a somewhat new point in P4V — how contract structures should take relationships between measures into account. This paper will describe the responsiveness of physicians to individual measures. Future work could explore the role of effort complementarity or substitutability by focusing on the effect of adding and dropping measures over time on other measure effort. Finally, determining whether measures are truly measuring value is a more difficult proposition. In order to demonstrate this, I would need to find significant enough responses to measures and determine whether longer term outcome measures or total cost of care responds over time. Overall, a longer panel of data would be needed to answer this question.

1.3.2. Response to financial incentives

Literature around provider responses to financial incentives proliferated after the introduction of Medicare's prospective payment system (PPS) in the hospital setting in 1983. This work generally used a simple pre/post design as the payment change was national in scope. Researchers found that hospitals responded by decreasing the lengths of stay as well as shifting services from the inpatient (where PPS was instituted) to the outpatient setting (where fee-for-service (FFS) was still in effect) (Coulam and Gaumer, 1992). Researchers also found that total Medicare spending declined, but do not agree as to whether this decline is a one-time change or a change in spending growth, which would suggest a change in provider practice patterns (Coulam and Gaumer (1992); Ellis and McGuire (1993)).

Few studies have specifically examined responses to financial incentives by physicians. The most convincing evidence comes from Clemens and Gottlieb (2014) who use geographic area-specific shocks resulting from changes in Medicare's payment formula to estimate physician response to a price change for all services. They find on average a 2 percent increase in price led to a 3 percent increase in health care services. They also find a higher response for more elective procedures.

Slightly more similar to my setting, a separate set of earlier papers looked at payment changes to a single or small set of services. McGuire and Pauly (1991) develop theory on physician response to a fee change for one service incorporating a physician's ability to induce demand as well as substitute across different services. The theory predicts inducement after a fee decline when a physician's income effect is large enough.[7] One would expect in settings where a service has a large impact on physician income, the income effect would overwhelm the substitution effect resulting in a quantity increase when there is a price decrease. Alternatively, in settings where the service has a relatively small impact on income, the substitution effect would dominate resulting in a quantity decrease when there is a price decrease. The setting of this paper includes only price increases. The same logic applies to a price increase - quantity decreases only if the income effect is large enough.

Empirical work to date has typically focused on price decreases. Yip (1998) and Jacobson et al. (2010) find support for this model as physicians respond to a price cut by increasing quantity provided for those same services (here cardiac bypass surgeries and certain cancer drugs respectively). Gruber (1994) use a different income shock - the dramatic decrease in the U.S. fertility rate - to test for physician response to financial incentives and finds a correlation between the negative income shock and substitution to higher margin cesarean deliveries. Some more recent work also finds physicians substituting to alternative and more expensive places of service (outpatient rather than office based setting) after a fee cut or alternative billing codes with the complete elimination of a service billing code (Song et al., 2015, 2013).

From the literature, we have seen physicians increase the provision of a service or substitute to alternative services with price decreases. In the P4V context, prices generally increase and often represent a small portion of the provider's income. While basic economic theory and the McGuire and Pauly (1991) model suggests that quantity should increase, this is surprisingly not well documented.[8] Furthermore, we know physicians respond to price

[7] More precisely the change in the marginal utility of income must be large enough.

[8] For the McGuire and Pauly (1991) model to hold, the income effects need to be small enough. P4V has to date represented a small portion of a provider's income and therefore P4V schemes would likely satisfy this condition. However, this setting examines a P4V scheme that represents a relatively larger portion of

changes when the price is constant. (In all of the above contexts, physicians are paid on a FFS schedule). However, this paper seeks to determine whether physicians are responsive to a fee schedule that shifts based on the quantity of services provided.

1.3.3. Pay-for-Value

One of the first studies of P4V programs compared a California insurer that rewarded participating physician groups for five ambulatory care and five patient reported measures to physician groups in the Pacific Northwest (Rosenthal et al., 2005). The researchers found a relatively small but statistically significant increase in only one of the three measures examined — cervical cancer screening (3.6%, p = 0.02). Payment was at the provider group level and at maximum represented 5% of that insurer's payment or 0.8% of the group's total revenue. Another early P4V study in western New York state also used a difference-in-difference approach and found more positive results with five out of six diabetic process measures improving (Beaulieu and Horrigan, 2005). This study compared 34 primary care physicians who opted into the program to other non-participating diabetic patients in the health plan.

Some of the largest P4V programs to date were a part of the Robert Wood Johnson Foundation and California HealthCare Foundation's Rewarding Quality program. The majority of these interventions did not have adequate control groups or pre-data to perform a difference-in-difference analysis (Young et al., 2007a). One study evaluating the program in Rochester, New York, which included one large insurer and physician group with 334 physicians, used a pre-post design and found an increase in only one out of four diabetic measures (Young et al., 2007b). Overall, these programs found modest increases in quality. As noted earlier, these pilots typically had a minority share of a patient panel participate and incentive dollars that ranged from 1% to 5% of a physician's income (Young et al., 2007b; Felt-Lisk et al., 2007).

a providers income. I will need to assume that income effects are small enough to predict quantity should rise based on the model.

Two more recent studies evaluating programs tying a higher proportion of a provider's pay to quality metrics found mix results. One study evaluated the Fairview Health Services P4P program, which tied 40% of a primary care provider's compensation to five clinical quality metrics, and found no overall increase in these metrics compared to other practices in Minnesota (Greene et al., 2015). The second study focused on Michigan Blue Cross Blue Shield's Physician Incentive Program (PGIP) (Lemak et al., 2015). Possible financial rewards included a bonus of up to 20% of a group's fee-for-service amount for meeting certain cost and quality metric goals. This early evaluation found an increase in seven out of the 14 quality measures relative to non-participating physicians. Both programs incorporate payment for quality metric scores and non-clinical measures such as total cost of care, panel size and patient-experience scores. Neither of these studies directly evaluated whether bonus size or use of thresholds attributed to physician responses to quality metrics making it difficult to determine whether results were driven by individual program attributes or bonus size/thresholds.

1.3.4. Behavioral responses to nonlinear budget sets

The last literature space is empirical work focused on estimating parameters of interest in the presence of nonlinear budget sets. This final section describes one popular empirical strategy developed first in labor and public economics as well as specific identification strategies in health economics.

The standard model of behavioral responses of earnings to taxes suggests that bunching (or excess mass) of individuals should exist at convex kink points in the tax schedule.[9] In a seminal paper, Saez (2010) develops an econometric method that uses the presence of bunching to estimate the compensated elasticity of income with respect to (one minus) the marginal tax rate. Many subsequent papers in the taxation literature have used bunching to estimate this elasticity (e.g., Chetty et al., 2011, 2013). Furthermore, the methodological

[9] More specifically, individuals must have convex preferences that are smoothly distributed among the population (following (Saez, 2010)). Without any kink points or nonlinearities, the distribution of individuals along the tax schedule should be concave.

insight has been applied to various other settings including education (e.g., Diamond and Persson, 2016), social security (e.g., Brown, 2013) and minimum wages (e.g., Harasztosi and Lindner, 2015).

Focusing on the health care setting, economists have built structural models that account for nonlinear budget sets (e.g., Einav et al., 2015, 2017a,b) or non-parametrically estimated behavioral responses across various budget segments (e.g., Abaluck et al., 2015). Einav et al. (2017b) demonstrate the difficulty of directly applying the Saez elasticity parameters to counterfactual estimates as different modeling assumptions can account for bunching. In particular, the authors find very different counterfactual results when incorporating a patient's lumpiness in drug need across a year and uncertainty relative to a model that assumes no frictions. The referenced dynamic structural models generally use the additional mass at kink points as moments in GMM estimation. This paper does not construct a structural model nor have the power to estimate a nonparametric model across budget segments. The paper develops a model that takes the nonlinear schedule into account, however this model is not estimated. Additionally this paper uses the nonlinearities to develop a set of instruments to help causally estimate the parameters of interest.

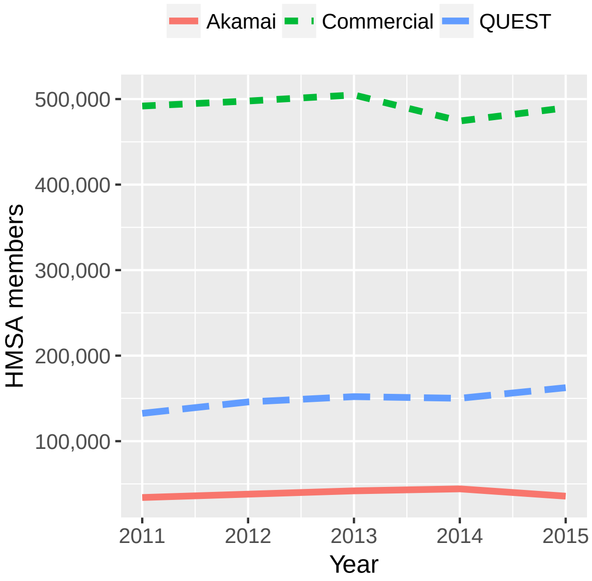
1.4. Context

1.4.1. Hawaii and HMSA

Hawaii's health insurance market has a high degree of managed care penetration - almost 50% of commercial plans and all Medicaid plans are managed care. Additionally, Medicare Advantage plans make up over 50% of the Medicare market. The Hawaiian Medical Services Association (HMSA), the Blue Cross Blue Shield plan of Hawaii, is the predominant private insurer in the state covering about 65% of all commercial patients and about 50% of all Medicaid Managed Care and Medicare Advantage patients. Kaiser Permanente, a closed-panel HMO (Kaiser physicians only see Kaiser patients), has about 25% commercial market share. This extremely consolidated market implies that a non-Kaiser physician's commercial panel is predominantly composed of HMSA patients, which is an important feature of the market.

HMSA plans covers half of Hawaii's total population, approximately 700,000 lives, between their three lines of business: Commercial (PPO and HMO products), Akamai Advantage (Medicare Advantage) and QUEST (Medicaid managed care). Figure 2 describes the number of lives in each line over time. Note the bulk of members are in commercial plans, about 550,000 lives, with approximately 70 to 80% in PPO plans during any given year.[10]

Figure 2: HMSA Membership over time by Line of Business



Notes: Figure includes HMSA members enrolled for at least one month during the year.

1.4.2. Payment scheme

HMSA has run a pay-for-value (P4V) program in some form since 1998. Up until 2011, physicians would have to select into this program and could receive up to 7.5% of their base pay per year for quality metric performance. Participating physicians received between $3,700 and $4,200 per year on average. This program ranked providers using four com-

[10] The commercial market share remains relatively constant across time so variation in members is mainly due to market size fluctuations. The Medicaid market size grows over time as does Quest's market share. Finally, HMSA's Medicare Advantage plan lost significant market share in 2015 to Kaiser. HMSA had expected lower Star ratings the previous year resulting in lower than expected CMS payments and an increase in HMSA premiums.

ponents: clinical performance metrics, patient satisfaction, business operation (electronic health record (EHR) use and participation in multiple HMSA lines of business) and health care utilization. The current P4V program focuses solely on the clinical performance component. The historic program began with 12 clinical performance measures and shrank to seven by 2010 due to changing HEDIS specifications (described below) and annual HMSA P4V working group decisions. The final seven measures included various cancer screenings, vaccinations and one diabetic measure (Hemoglobin A1c testing). A number of studies evaluated this program finding improvement for originally lower performing physicians after three years of the program, but little average effect (Gilmore et al., 2007; Chen et al., 2010, 2011).

The focus of this study is the P4V programs that began in 2011 and 2012. The commercial program rolled out in 2011 and the Medicare and Medicaid programs began one year later. The P4V programs only include primary care providers (PCPs) and required provider participation. This is unlike the previous P4V program where all providers, no matter the specialty, had to opt in. The commercial PCP fee schedule froze in 2011 (no medical inflation updates) and all PCPs began to receive quarterly quality incentive payments for commercial and Medicaid plans and a single yearly payment for Medicare plans. The number of measures increased from seven to 10 (and more measures in later years) with the addition of process diabetes, asthma and heart failure measures (see Table 13 for a description of the measures over time by line of business).

Between 700 and 950 primary care providers participated in a given year. This number expanded over time due to the addition of lines of business in 2012 and the expanding attribution of patients to providers, which is described in detail later. Table 1 presents the total number of participating providers, member-months, and maximum and actual bonus received across all lines of business each year. The average annual bonus was between \$30,000 to \$42,000 from 2012 onward, which is about 10 times the size of the average bonus payment in the preceding program.

Table 1: Bonus pay over time

	Physicians	Member Months	Max Bonus	Actual Bonus
2011	698	6,308 (5,272)	12,616 (10,545)	8,315 (10,008)
2012	938	4,007 (5,274)	45,296 (47,763)	29,565 (36,619)
2013	940	4,075 (4,814)	50,525 (53,487)	42,922 (53,614)
2014	952	4,179 (5,555)	44,687 (45,072)	32,898 (40,049)
2015	983	3,824 (5,770)	40,445 (41,453)	27,608 (34,554)

Notes: Table includes all primary care providers participating in the P4V program in any quarter and any line of business. Bonus amounts represent annual dollar amounts.

The Medicaid and Commercial P4V program structures are similar, whereas the Medicare program structure is different on a number of dimensions. I will first describe the Commercial and Medicaid schemes in detail and then describe how the Medicare program differs.

Commercial and Medicaid Managed Care Pay-for-Value Programs

A key component to many P4V schemes is designing an algorithm to specify a provider's patient panel. The provider is then responsible for all of the patients in their panel. Attribution of patients to physicians for HMO products is straightforward. An enrollee generally chooses a PCP when signing up for a plan. Attribution for PPO products, which covers the largest number of HMSA lives, uses a claims-based algorithm. Each month, a patient is attributed to the PCP who has seen that patient for the majority of PCP visits in the preceding 16 months. If a patient does not have 16 months of claims history or does not have a single visit to a PCP, she will not be attributed. Finally, it is also possible for providers to directly select patients to be in their panel. The attribution algorithm is set such that physician selection overrides patient selection, which overrides the claims base method. One worry about any attribution scheme is the ability of providers to select patients either directly choosing their panel or indirectly by scheduling visits for certain patients and not others. Direct selection of patients by providers occurs less than 2% of the time and observable risk characteristics between patients directly and not directly selected are similar. Additionally patients who do switch to a new provider appear sicker than those who do not switch suggesting that selection is not occurring in order to maximizing one's bonus. Furthermore, the identification strategies described in Chapter 2 estimate short-term responses (responses in the subsequent quarter) and indirect selection should take many months as the claim look back period is 16 months. The descriptive statistics and empirical approaches suggest that direct and indirect patient selection is likely not driving results.[11]

The maximum possible bonus amount for each physician was based on the number of attributed patient-months and an HMSA defined Per-Member-Per-Month (PMPM) amount. This PMPM amount increased over time from $2 in 2011, to $4 in 2012 and 2013, and finally to $4.50 in 2014 and 2015.

Providers received individual bonus payments for each quality measure based on this maximum bonus amount. Quality measures in the program included both process measures - measures where a provider needs to perform a service such as diabetic eye screening or mammogram for a breast cancer screening- and intermediate outcome measures - biometric readings that are often results of process measures such as Diabetic LDL level or HbA1c reading. The measures were based on HEDIS or The Healthcare Effectiveness Data and Information Set specifications. HEDIS is a tool that over 90% of US health plans use to track and compare their quality performance. Insurers submit a mix of information including claims, survey responses and at times medical charts, which are used to generate these measures. Each year, HEDIS publishes measure specifications detailing the collection and aggregation of each measure. An insurer can thus calculate their own internal HEDIS quality measures using these published specifications.

The set of P4V measures evolved over time - adding some measures while dropping others. Many of the new measures were intermediate outcome measures rather than process measures, which require lab and other biometric results rather than simply claims. Additionally, many of these measures require multiple years of claims data to calculate. I will focus on

[11] For a detailed description of patient characteristics for direction selection and patient switchers see Appendix Tables 19 and 20.

a subset of measures, specifically preventative cancer screenings and process diabetes measures, because they are primarily claims based measures and exist in most lines of business over time.[12]

A physician's measure specific bonus payment b_{ijt} for physician i in measure j and quarter t, for Commercial and Medicaid P4V programs was defined as:

$$b_{ijt} = B_j(D_{it}; W_t)F_t(r_{ijt}, r_{ijt-1}; T_{jt}) \tag{1.1}$$

where $B_j(\cdot)$ described the maximum bonus amount for measure j and $F_t(\cdot)$ described the proportion of $B_j(\cdot)$ a provider received. $B_j(\cdot)$ was a portion of B (the maximum bonus amount) and was a function of 1) one's patient panel, D_{it}, and 2) HMSA defined measure weights for all measures, W_t.[13] The maximum bonus amount for each physician was divided up among all measures where a higher B_j went to measures that HMSA decided were more important (HMSA defined weights, W_t) and to measures with more relevant patients (e.g., a physician with a lot of diabetic patients had a higher diabetic B_j than a physician who had more pediatric patients). As an example of the weighting, the HMSA weight for diabetic nephropathy screening was two times the diabetic LDL screening weight and four times the preventative breast cancer screening weight in 2012.[14]

Finally, $F_t(\cdot)$ was defined by where one's current and former total performance, r_{ijt} and r_{ijt-1}, fell in HEDIS's national distribution for the specific measure. Total performance is the sum of quarterly performance from all relevant quarters, $r_{ijt} = \sum_{t-n}^{t} p_{ijt}$, where n was at minimum 3 for diabetic measures and up to 39 for the colorectal cancer screening measure (a one and 10 year period respectively). Quarterly performance, p_{ijt}, was defined as the portion of patients who were screened during quarter t and who were previously not

[12] See Appendix Table 13 for the list of all measures over time for each line of business and Appendix Table 15 for a detailed description of the selected measures.

[13] The specific definition of B_j is detailed in Appendix Section A.3

[14] The list of weights by measure and year for the commercial line of business are described in Appendix Table 14.

screened. Importantly, p_{ijt} was defined by the current set of attributed patients and the patients' screening history independent of the current attributed physician.[15]

Importantly, total performance was compared to a national benchmark. HEDIS collects data from almost all private insurers in the US and annually publishes distributions of each measure by line of business. A provider received an increase in the proportion of bonus pay for the current performance, r_{ijt}, exceeding specific thresholds: the 10th, 25th, 50th, 75th and 90th national percentiles. The provider received an additional increase in the proportion of bonus pay if their previous year's performance, r_{ijt-1}, is below their current performance, r_{ijt}. For example, one received an additional bump if one was in the 50th percentile the preceding year and exceeded the 90th percentile the following year. Figure 3 describes this proportion scheme. Note, there are major improvement bonuses for improving by at least two percentile thresholds. Importantly, $F_t(\cdot)$ introduces a nonlinear element to the payment scheme. Additionally, a major change in this nonlinear element occurs between 2012 and 2013. Figure 4 demonstrates how thresholds evolved over time for two separate measures, breast cancer and diabetic eye screenings. Typically thresholds shifted a small amount each year and did not consistently increase.[16]

Medicare Advantage

Bonus pay in the Medicare program was calculated as:

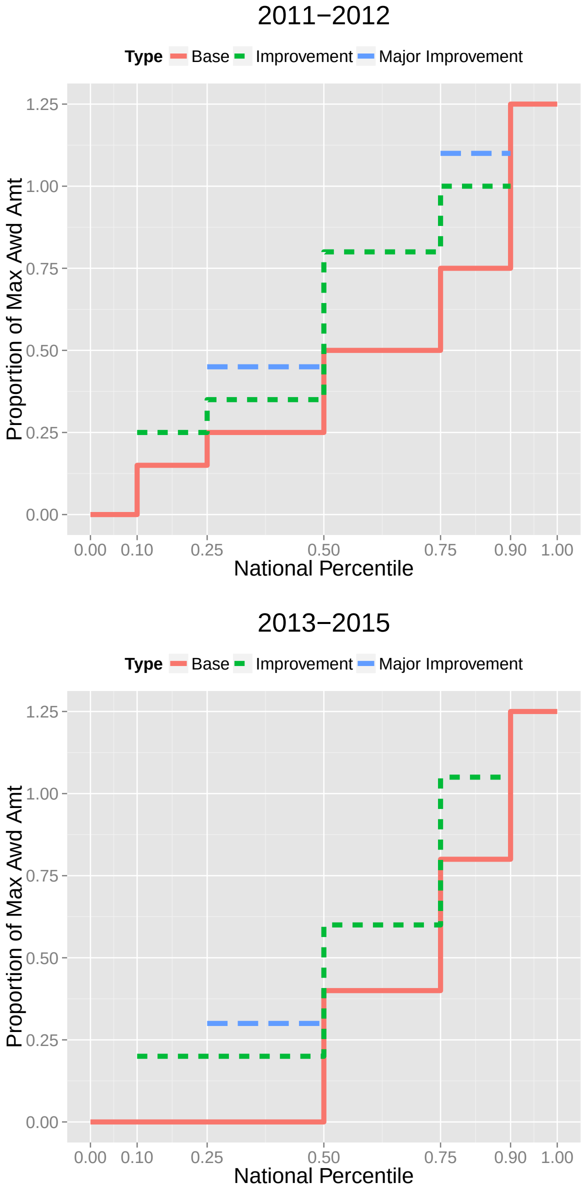
$$\begin{aligned} b_{ijt} &= B^M(d_{ijt}; PMPM^M_{jt}) * F_t(r_{ijt}, r_{ijt-1}; T_{jt}) \\ b_{ijt} &= d_{ijt} * PMPM^M_{jt} * F_t(r_{ijt}, r_{ijt-1}; T_{jt}) \end{aligned} \tag{1.2}$$

The maximum bonus amount for a measure, B^M, was simply the number of relevant pa-

[15] This implies that a patient could have had a screening completed at a time when they were not attributed to their current physician and this screening still counts toward p_{ijt}. Alternatively, a physician could have screened in the current or previous period patients who are not currently attributed to that physician and all of these screenings would not count towards p_{ijt}.

[16] See Apenndix Tables 16 and 17 for a full description of thresholds over time for each measure.

Figure 3: Proportion of Maximum Bonus by National Percentile

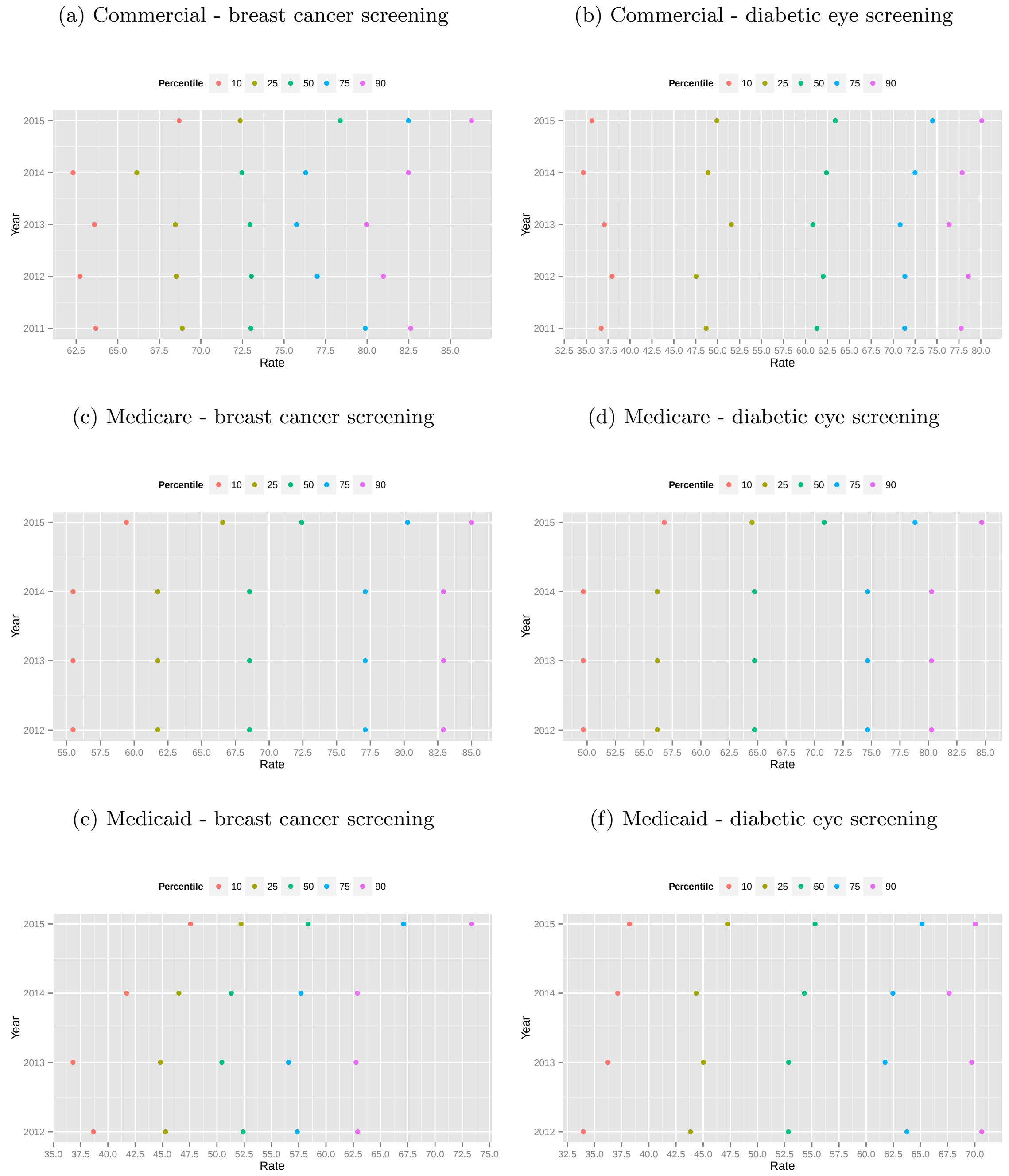


Notes: Figure plots the proportion of the maximum bonus amount received against the national percentile. One receives a higher bonus for improving performance relative to the prior year's performance, "Improvement" and "Major Improvement" (See text for details).

tients, d_{ijt} (e.g., number of diabetic patients) multiplied by the HMSA set $PMPM_{jt}^{M}$.[17] The $F_t(\cdot)$ function follows the same proportion scheme as the Commercial and Medicaid program, but uses CMS's star rating system thresholds instead of HEDIS percentiles. The

[17] Described in the Appendix Table 13.

Figure 4: Thresholds over time for Breast Cancer and Diabetic Eye Screening by LOB



(a) Commercial - breast cancer screening

(b) Commercial - diabetic eye screening

(c) Medicare - breast cancer screening

(d) Medicare - diabetic eye screening

(e) Medicaid - breast cancer screening

(f) Medicaid - diabetic eye screening

Notes: Each figure plots on the x-axis the 10th, 25th, 50th, 75th, and 90th percentile for a single line of business and measure, either breast cancer screening and diabetic eye screening, against year on the y-axis.

CMS star rating thresholds were updated once during this study period rather than every single year. Figure 4 demonstrates how thresholds for two measures evolved over time for all lines of business including Medicare Advantage.[18] Finally, as noted before, payouts for the Medicare bonus system occured once a year rather than quarterly.

1.5. Data

1.5.1. Data files

The data elements for this study include the claims from the universe of HMSA members between 2011 and 2015. The claims data includes medical, lab and pharmacy claims, member enrollment files with age and sex, a provider file with practice name and zip-code. I also have quarterly provider bonus amounts by measure, which includes the provider's attributed member-months and quality measure performance. This end of the quarter quality performance snap-shot aids in the construction of quality measures from the claims data. Additionally, when a quality measure cannot be constructed via claims, I know the final quality measure rate each year.

For patient level risk, I am using Elixhauser Comorbidity Indicators. This is a publicly available algorithm through the Agency for Healthcare Research and Quality's (AHRQ) HCUP and uses inpatient, outpatient and pharmacy claims to identify patients with certain comorbid conditions. I construct physician panel level variables for the percent of attributed patients with the various comorbid conditions. The risk variables are constructed at the physician-quarter level as attribution changes quarterly.

I reconstruct the quality measures using medical and pharmacy claims data. Unfortunately, the majority of lab data does not contain sufficient detail to populate most lab based quality measures. As described above, I chose six quality process measures that are predominately derived from claims, exist in most lines of business and exist in the majority of years.[19]

[18] See Appendix Table 18 for a full description of thresholds over time for each measure.

[19] An individual physician quality metric is typically either a process or an outcome measure. Process measures assess whether specific services are provided to a patient such as the receipt of beta blockers after

These measures include three preventative cancer screenings (breast, cervical, and colorectal cancer) and three diabetic process measure (HbA1c testing, nephropathy screening, and annual eye exam). Note primary care physicians can perform two of the diabetic process measures (HbA1c testing and nephropathy screening), but must refer to other providers for most other screenings.

Table 2 describes the final data set including panel size by line of business and the potential and actual bonus pay for all measures and for the six specific measures studied. I only include physicians, dropping the small number of advanced practice nurses and physician assistants. Physicians-measure pairs must exist in all quarters between 2012 and 2015 to be included. There are 8,224 provider-quarter pairs or 514 unique physicians. The vast majority of P4V physicians see commercial patients (88%), slightly fewer physicians see Medicare Advantage patients (66%), and a minority of physicians see Medicaid managed care patients (18%). The average panel size is largest for the commercial program as would be expected. Finally, the six measures focused on in this study represent about one-quarter of a provider's possible and actual bonus pay. Note the potential bonus and actual bonus across all measures is above the full sample values listed in Table 1 when converting to the year level, which is to be expected when focusing on physicians who consistently participate in the program.

1.5.2. Bonus payment and quality performance over time

This section describes bonus and performance changes over time and reviews at a high level whether changes in the bonus program over time appear correlated with performance improvement. The figures present suggestive evidence of the results in the main analysis.

Figure 6 describes the actual bonus, potential bonus and ratio of actual to potential bonus over time for physicians across all measures. The bonus received is similar for Medicare and commercial lines of business with a spike in the Medicare bonus in 2013. On average

a heart attack or annual eye exam for diabetic patients. Outcome measures assess whether a patient fits a specific health state. For example, whether a heart attack patient is readmitted to a hospital within 30 days or a diabetic patient has their HbA1c level under 8.

Table 2: Summary Statistics of Final Data Set (physician-quarter level)

	Mean	SD	Median	N	Unique MDs
Panel Size - Commercial	518	408	447	7248	453
Panel Size - Akamai Advantage	147	111	130	5408	338
Panel Size - QUEST	327	710	166.5	1488	93
Potential Bonus	16,421	12,688	14,355	8224	514
Bonus	12,114	12,059	8,468	8224	514
Potential Bonus - 6 measures	4,218	4,236	3,059	8224	514
Bonus - 6 measures	2,700	3,665	1,084	8224	514

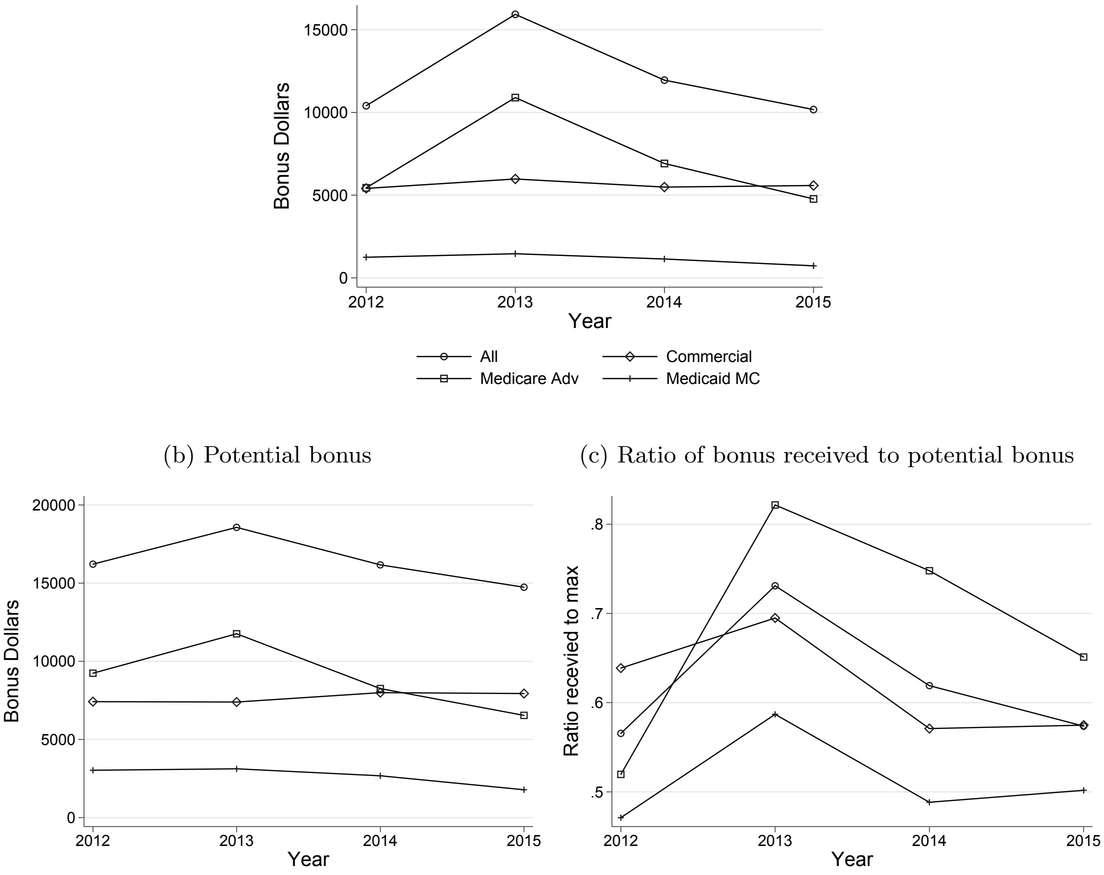
Notes: Observation is at the physician-quarter level. Table includes all primary care physicians consistently participating in the P4V program during 2012 through 2015. Physician-measure-quarter observations are aggregated across all measures and lines of business. The six measures included in the last two row are three preventative cancer screening measures (breast, cervical, and colorectal) and three diabetic screening measures (HbA1c testing, nephropathy screening, and annual eye exam).

these bonus dollars represent 50% and 42% of a physician's total bonus dollars. The Medicaid bonus dollars are much smaller, representing 8%. Commercial bonus dollars appear relatively constant over time whereas the Medicaid and Medicare dollars decline after 2013.

Panels b and c give some insight into why the bonus dollars evolve over time. Panel c plots the ratio of actual to potential bonus (or the $F(\cdot)$ function introduced in Equation 1.1) over time and can be interpreted as the percent of bonus dollars left on the table. Commercial pay appears relatively flat in panel a, however this masks an increase in the potential bonus amount and a decrease in F between 2013 and 2014. Medicare pay spikes in 2013 due to increased bonus pay and increased F. Further Medicare pay declines after 2013 due to declines in pay and F. Similarly, Medicaid pay declines after 2013 due to potenial pay and F.

When conducting the same exercise for the six measures focused on in this study, new patterns emerge. First, Medicare bonus pay is significantly lower than any other line of business largely because the Medicare program only includes four of the six measures and the number of patients in these measures is smaller. Additionally, bonus dollars generally decreased from 2012 to 2015 unlike bonus dollars from all measures in Figure 6 a. Much of

Figure 5: Average bonus dollars over time



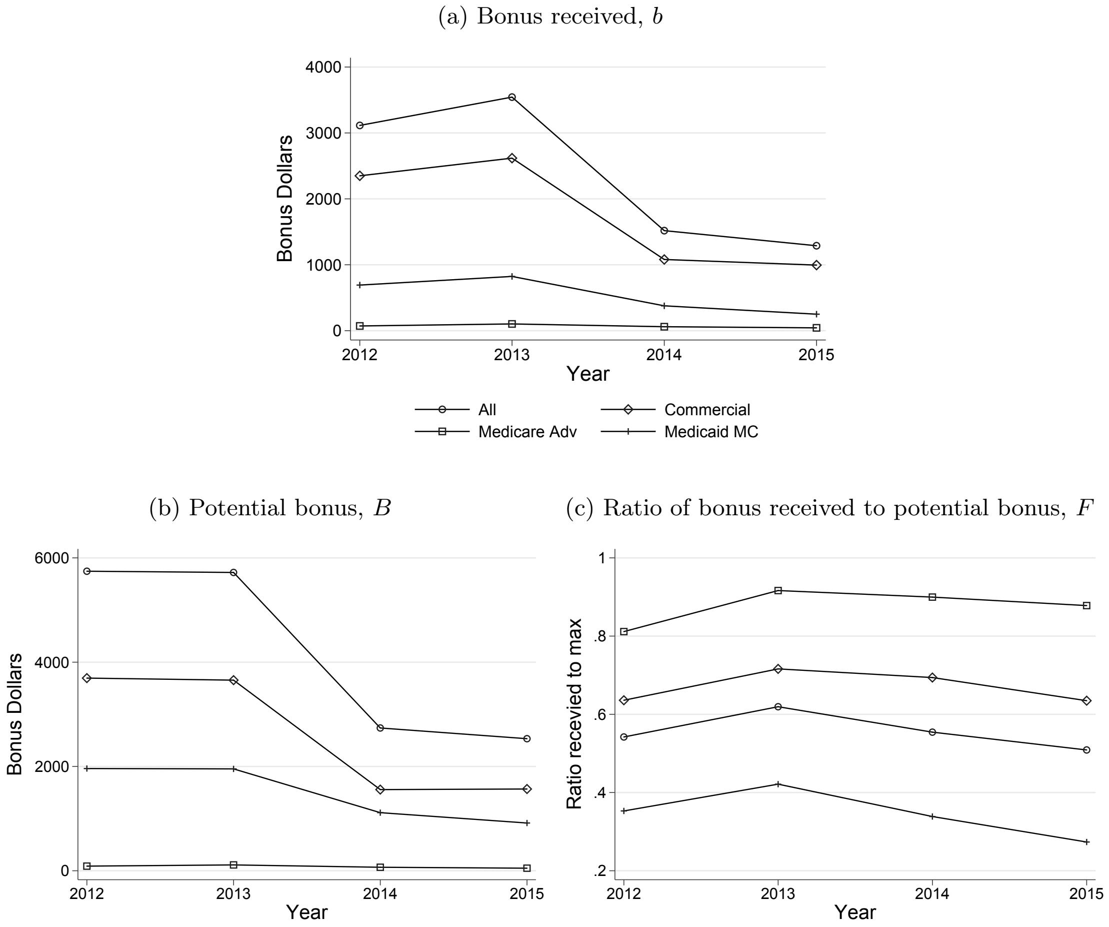
(a) Bonus received

(b) Potential bonus

(c) Ratio of bonus received to potential bonus

Notes: Data comes from insurer generated fourth bonus maximum and bonus received for each year. All physicians who are included in the final data set are included above (see Section 1.5.1). This figure includes bonus dollars from all lines of business and all measures irrespective of which measures are included in the final data set.

Figure 6: Average quarterly bonus dollars over time for six measures


(a) Bonus received, b

(b) Potential bonus, B

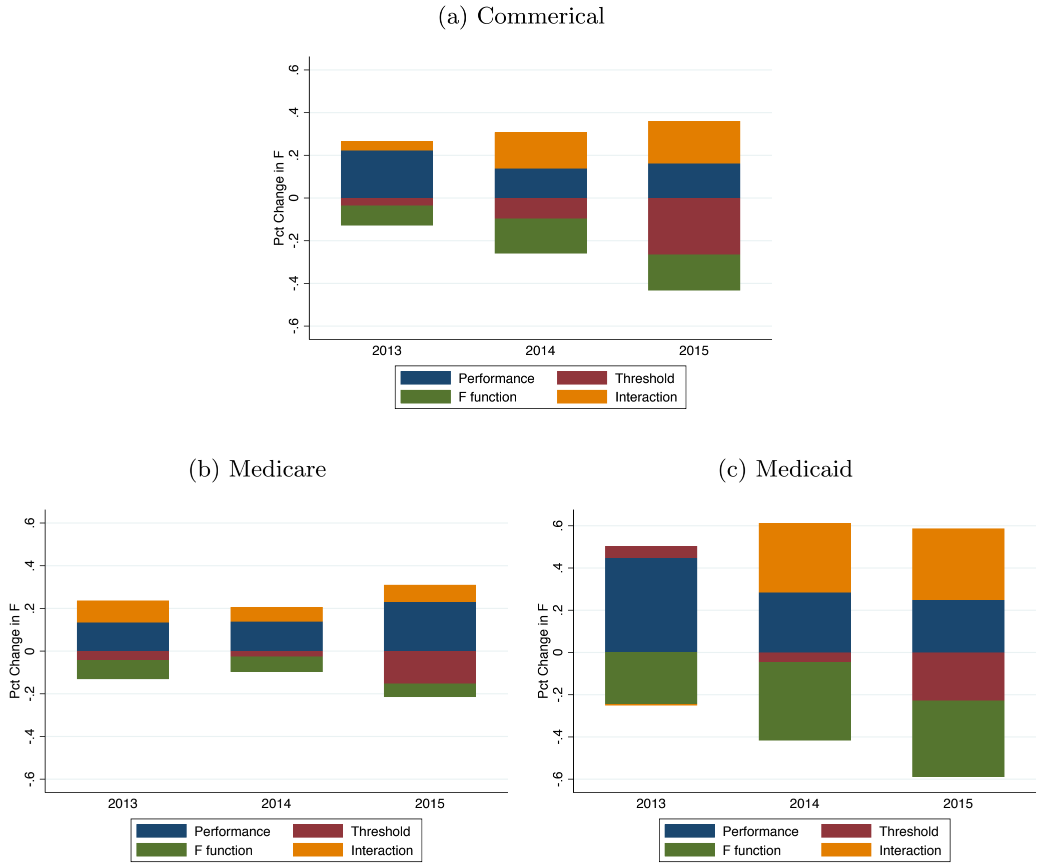
(c) Ratio of bonus received to potential bonus, F

Notes: Data comes from insurer generated fourth quarter bonus maximum and bonus received for each year. All physicians who are included in the final data set are included above (see Section 1.5.1). This figure includes bonus dollars from all lines of business and the six measures focused on in this study.

the decline is due to the potential bonus amount declining which in turn is largely due to the addition of new measures over time. The decrease in potential bonus pay was accompanied with the F measure increasing in 2012 and then declining.

Had only thresholds increased over time and performance remained constant, one would expect the F measure to decrease. Similarly, if performance remained the same, one would expect F to decrease between 2012 and 2013 due to the new piece rate schedule (see Figure 3). Therefore, performance would have to concurrently increase over time in order for the F measure to remain constant as seen in Figure 6 c. Figure 7 precisely decomposes the

Figure 7: Decomposition of changes in the bonus ratio (F)



(a) Commerical

(b) Medicare

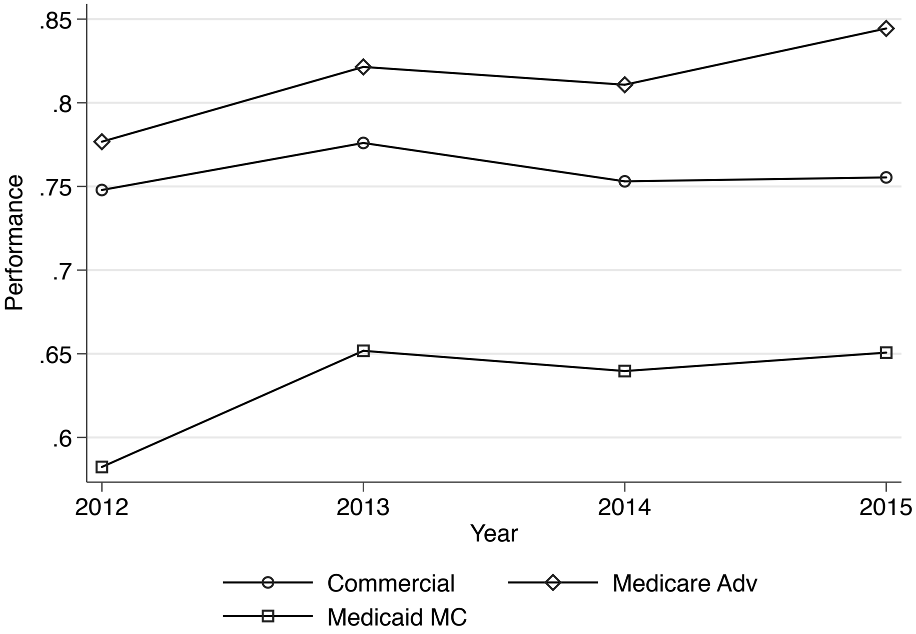
(c) Medicaid

Notes: Each bar size represents the change in F between 2012 and the listed year (F_t/F_{2012} - 1). The bar is colored by the portion of the change that can be accounted for by corresponding variable (see Footnote 19).

changes in F into changes due to actual performance, threshold locations, and the single step function change ($F_{t\in 2011-2012}(\cdot)$ vs $F_{t\in 2013-2015}(\cdot)$). The bar size is the amount of change in F from 2012 to the current year and the bar colors represent the contribution of each factor to that change. More specifically, a bar color can be interpreted as the change in F had only one factor changed.[20] As expected the threshold and step function contribute negatively to the overall change in F. Performance increases prevent F from significantly declining over time. Somewhat surprisingly, this exercise demonstrates that thresholds do not increase significantly over time except for perhaps the final year.

[20] For example the bar size for actual performance is $\hat{b}(r_t, \tau_{2012}, F_{2012}/\hat{b}(r_{2012}, \tau_{2012}, F_{2012}) - 1$.

Figure 8: Performance over time for six measures



Notes: All physicians in final data set included as long as performance represented a relevant panel size above 10.

Finally, Figure 8 presents the actual performance over time or the percent of recommended completed screenings. As expected from the previous exercise, performance increases over time albeit minimally for the commercial program. On average commercial performance improved from 74.8% to 75.5%, Medicare improved 77.7% to 84.4%, and Medicaid improved 58.2% to 65.1%.[21] The exercises demonstrate that bonus pay could not have incentivized performance increases over time and instead perhaps disincentivized improvement because bonus pay declined during the sample period. Alternatively, threshold increases over time could have incentivized improvement. This dissertation identifies the portion of the performance change that can be attributed to changes in the marginal bonus amounts and threshold locations.

[21] See Appendix Table 17 for performance over time at the measure level.

CHAPTER 2 : Response to nonlinear incentives

2.1. Introduction

This chapter estimates physician average response to changes in relevant performance pay contract features. To motivate the empirical analysis, I develop a model to identify the payment contract characteristics that affect a physician's choice of effort. The model includes effort from the current and preceding period as the HMSA bonus amount is based on multiple periods of performance. The current period's effort depends on a physician's distance from a threshold at the beginning of a quarter and the marginal bonus payment for surpassing the threshold. Increasing both of these features is expected to increase effort for physicians close enough to the threshold. Following the theoretical model, I estimate the impact of the two contract features on a physician's choice of effort for the set of six process measures described in Section 1.5.

I identify two natural experiments that occurred between 2011 and 2015 which represent plausibly exogenous changes in, respectively, a physician's distance from a threshold and bonus amount for specific measures. The two natural experiments include an increase in the breast cancer screening threshold in 2015 and a decrease in the diabetic nephropathy screening payment level in 2014. I use a difference-in-difference framework to estimate the differential change in quarterly performance for measures with the threshold or bonus pay change relative to measures without the change. In order to account for the possibility that physician-measures in different parts of the performance pay structure respond differentially to threshold changes, I match observations based on location in the pay structure. I also match observations based on performance trends over time. As a robustness check in an additional specification, I ensure physicians exist in either the treated or control groups and only include observations where the treated measure represents a large bonus. The difference-in-difference estimation strategy identifies the response for only two measures and leverages a single source of variation in the contract features.

Next, I directly estimate the responsiveness to changes in distance from a threshold and marginal bonus pay for all measures. Estimation of these parameters suffers from a variety of biases, including patient selection, additional unobserved physician characteristics and a mechanical relationship. Physician and physician-measure fixed effects are included in some specifications to conservatively remove bias from patient selection and other unobserved physician characteristics. Separately, two new instruments are proposed that leverage plausibly exogenous changes in a patient's performance measurement status. Specifically, many quality measures were captured over a period of one or more years. When a patient receives a screening or visit, the patient counts positively towards that physician's quality measurement for a number of quarters. The patient must be screened once again after a set number of quarters. I consider the quarter when the visit lapses as plausibly exogenous, particularly for measures collected over many years. The second instrument leverages patients aging into measure definitions, which are set by the US Preventative Task Force. These instruments capture plausibly exogenous variation across time within a physician's panel of the physician-measure location in the payment schedule and marginal bonus pay.

This chapter proceeds as follows: Section 2.2 develops a theoretical model to motivate the empirical analysis; Section 2.3 details the natural experiment methods and results; Section 2.4.3 describes the direct estimation strategy and results; and Section 2.5 provides a discussion of the chapter.

2.2. Theory

I introduce a basic contract with a nonlinearity similar to one found in the Hawaii context. I take the principal's (or insurer's) contract as given and find the effort that maximizes the agent's (or physician's) utility. The purpose of this modeling exercise is to determine what features of the contract impact an agent's choice of effort and perform some comparative statics.

First, I assume that a physician's wage only includes the bonus payment. Implicitly, I am removing the traditional fee-for-service pay structure where a physician receives one set fee for each service provided. I discuss implications of the simplification at the end of the section. I define a physician's wage w in time period t as:

$$w_t = b * \mathbb{1}(x_t + x_{t-1} > \tau)$$

where b is a bonus payment that a physician receives after some set number of their patients, τ, meet a quality metric. The number of patients meeting a quality metric include those meeting the metric in the current period, x_t, and those meeting the metric last period or the number of "banked" patients, x_{t-1}. The number of services provided x in time period t is a function of a provider's effort, e, and some error, ϵ.

$$x_t = e_t + \epsilon_t, \text{ with } \epsilon \sim N(0, \sigma^2)$$

In the context of contracting on quality, there is uncertainty around how a provider's effort translates into patients meeting a measure. Through conversations with the insurer about physician response, I view effort as a physician directing their front line staff to either increase the amount of contact with non-compliant patients (e.g., calling and emailing about visits) or having the front line staff increase the time of specific patient visits. For each patient receiving an additional phone call or longer office visit, the likelihood of meeting a measure increases. Furthermore, process and intermediate outcomes measures likely have different levels of uncertainty with higher uncertainty for intermediate outcome measures such as blood pressure control where an additional visit or longer visit does not directly translate into an additional patient meeting the measure.

I assume that the utility function includes only the wage and some cost function, which are linearly separable.

$$U(e_t) = u(w(e_t; x_{t-1})) - f(e_t; x_{t-1})$$

The cost function depends on a given level of x_{t-1} as well as effort e_t. Cost is increasing in both inputs as cost increases for every additional patient seen irrespective of the time period. I also assume that $u(w)$ is concave and $f(e_t; x_{t-1})$ is convex in both e_t and x_{t-1} and both are continuously differentiable.[1] Note, these assumptions do not ensure a unique solution as the nonlinearity introduces non-concavity.

Expected utility is therefore:

$$\begin{aligned} E[U] &= Pr(x_t > \tau - x_{t-1})U(x_t > \tau - x_{t-1}) + Pr(x_t \leq \tau - x_{t-1})U(x_t \leq \tau - x_{t-1}) \\ &= \Phi\Big(\frac{e_t - \tau + x_{t-1}}{\sigma}\Big)u(b) + \Big(1 - \Phi\Big(\frac{e_t - \tau + x_{t-1}}{\sigma}\Big)\Big)u(0) - f(e_t; x_{t-1}) \end{aligned}$$

Taking the first order conditions results in:

$$\begin{aligned} \phi\Big(\frac{e_t - \tau + x_{t-1}}{\sigma}\Big)u(b) - \phi\Big(\frac{e_t - \tau + x_{t-1}}{\sigma}\Big)u(0) &= f'(e_t; x_{t-1}) \\ \phi\Big(\frac{e_t - \tau + x_{t-1}}{\sigma}\Big)[u(b) - u(0)] &= f'(e_t; x_{t-1}) \qquad (2.1) \end{aligned}$$

The expected marginal benefit of increasing effort by one unit is $\phi(e_t - \tau + x_{t-1})[u(b) - u(0)]$. The marginal benefit is a product of 1) the change in the probability of surpassing the threshold τ for an additional unit of effort given the level of effort, e_t, and the number of patients meeting a measure the preceding period, x_{t-1}, and 2) the difference in utility between receiving a wage of b and 0, $u(b) - u(0)$.

Figure 9a depicts the marginal benefit curve, a normal PDF scaled by $u(b) - u(0)$ centered at an effort level of $\tau - x_{t-1}$, and a marginal cost curve, here assumed to be linear. The optimal level of effort is e_t^*. Note that e_t^l is not an optimal level of effort. The marginal benefit curve is above the marginal cost curve for effort directly above e^l, which implies that e^l is a saddle point. Figures 9b and 9c demonstrate that if the threshold is small enough, the agent will put forth little effort. Alternatively, if the threshold is large enough, an agent will put forth no effort. Finally, Figures 9d and 9e demonstrate low levels of effort in the current period e_t could be due to a low τ or high x_{t-1} and similarly, high levels of

[1] This assumption is traditional in the literature. The marginal cost of effort weakly increases and the marginal benefit of effort weakly decreases.

e_t could be to due a high τ or low x_{t-1}.[2] The HMSA payment schedule includes multiple thresholds, which is not captured in this simple model. However, comparing 9b and d could also be thought of as comparing two observations with equal $\tau - x_{t-1}$ and two different τ's. Observations lower in the performance schedule (or closest to lower τ's) are predicted to exert more effort as the marginal cost is lower.

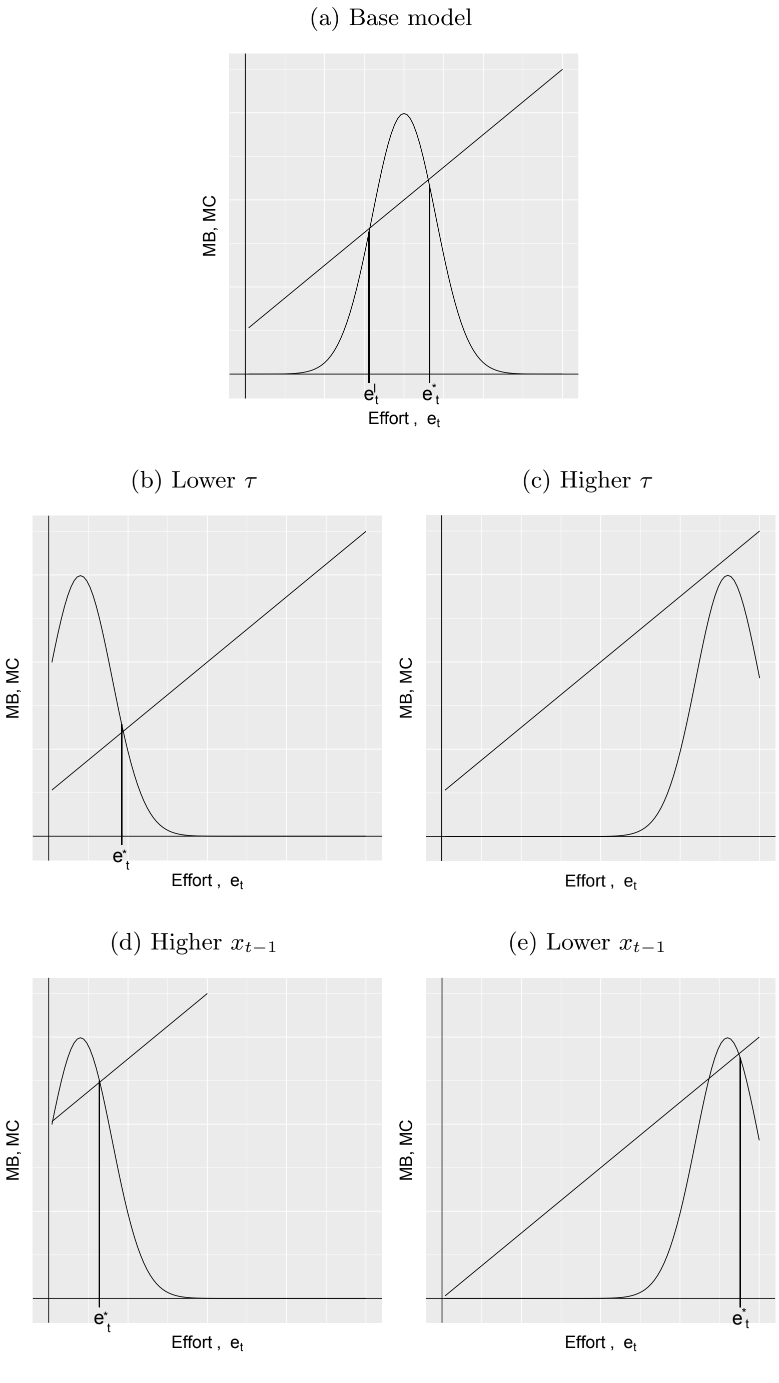
Next, I perform some comparative statics for a number of contract features - b, τ and σ. Note that all of these features affect only the MB curve. Figure 10a depicts an increase in the bonus payment b and Figure 10b depicts an increase in τ. In both instances the optimal effort levels shifts up. Figure 10c depicts an increase in σ. Here the MB curve is now relatively flatter and is below the original MB curve near $\tau - x_{t-1}$ and above the original MB curve far away from $\tau - x_{t-1}$. In this instance, optimal effort increases however if the MC function was flatter or shifted down (lower MC), optimal effort would have decreased. There is no clear change in optimal effort for a change in σ. One additional feature to explore is whether effort increases differentially for equal $\tau - x_{t-1}$'s with two different τ's. For example, does this model predict that an increase in τ will result in a larger change in effort for lower τ's (i.e., differential responsiveness to τ_{10} relative to τ_{90})? With the assumed linear MC curve, the change in e_t would be higher for an increase in b for lower τ's and e_t would be equal across τ's. With different MC curves, these results may not hold.

As noted previously, this model does not include a physician's fee-for-service schedule. Removing the fee schedule greatly simplifies the first order conditions. Further, assuming the income effect is zero, the marginal benefit function would simply have an additional constant, which would shift up the scaled PDFs in all figures.[3] Additionally, this model focuses on a single measure in isolation and therefore it does not consider spillovers or

[2] The optimal e_t is different between Figures 9b and d and similarly between Figures 9c and 9e. This is because cost is a function of e_t and x_{t-1}. A higher x_{t-1} shifts the MC up and a lower x_{t-1} shifts the MC down. Here, I simply want to demonstrate both factors can shift e_t and therefore must be taken into account.

[3] This model does not explicitly include or exclude income effects. The assumption of a concave $u(w)$ function allows either to exist. Income and substitution effects should be more fully explored if future model iterations include a fee-schedule with services outside of those rewarded in the bonus program or include multiple types of bonus measures.

Figure 9: Theoretical model's marginal benefit and marginal cost curves



(a) Base model

(b) Lower τ

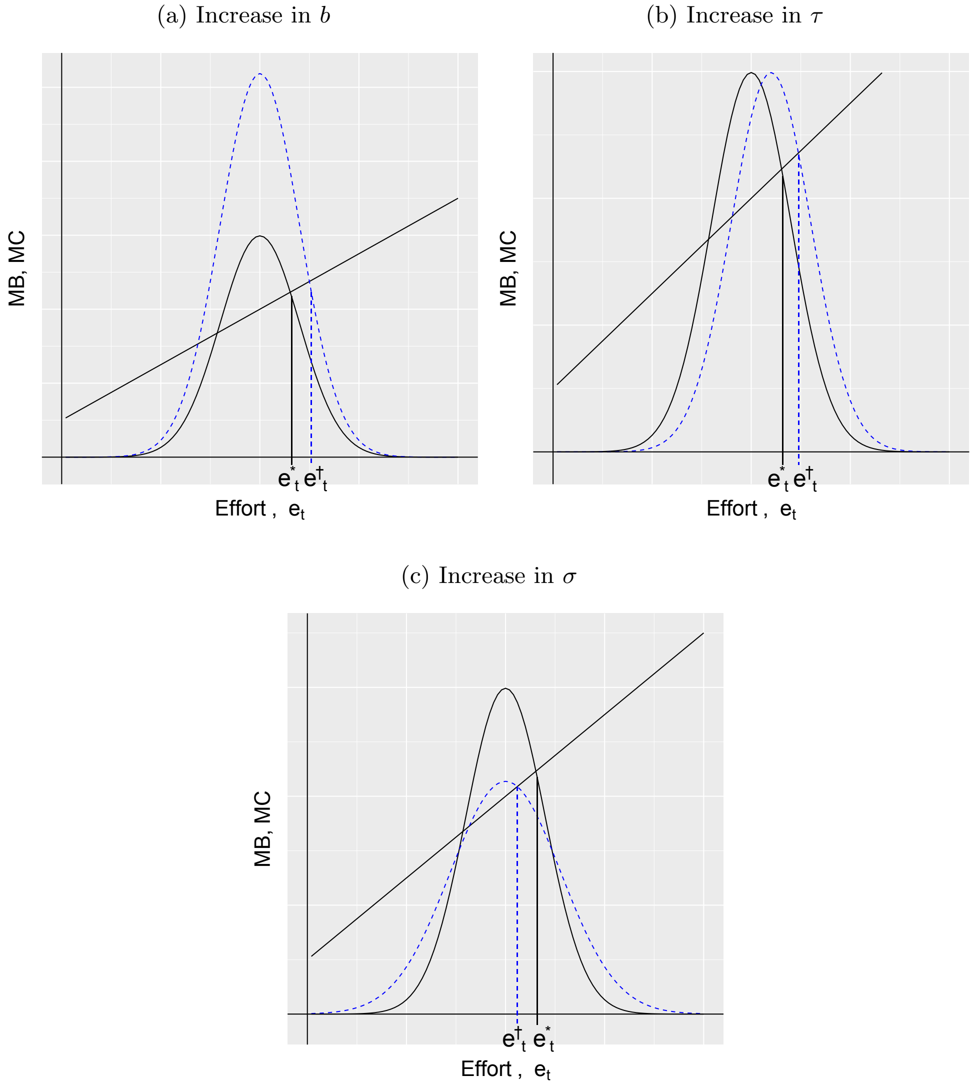
(c) Higher τ

(d) Higher x_{t-1}

(e) Lower x_{t-1}

Note: Figures above plot the marginal benefit (MB) and marginal cost (MC) curves from Equation 2.1. MB is a normal probability density function centered at $\tau - x_{t-1}$ and scaled by $u(l) - u(0)$. MC is assumed to be linear.

Figure 10: Theoretical model's comparative statics


(a) Increase in b

(b) Increase in τ

(c) Increase in σ

Note: Figures above plot the marginal benefit (MB) and marginal cost (MC) curves from Equation 2.1 with increases in various contract features that happen to solely affect MB. Figure 10a increases the bonus payment, b scaling up MB. Figure 10b increases the threshold location, τ. Figure 10c increases the uncertainty of effort translating into an outcome, σ.

multi-tasking, which could be added in the future. Finally, the model is static and does not incorporate learning overtime. For example, sigma could decrease overtime as physicians learn how to best encourage patients to visit. While learning is an interesting and relevant extension, adding the additional dimension is non-trivial and not relevant for the current empirics.

Overall, this model demonstrates that an agent's choice of effort depends on their distance from the threshold, $\tau - x_{t-1}$, size of the bonus, b and output in the previous period, x_{t-1}. Further, increases in l and local increases in τ increase an agent's choice of optimal effort for agents who are close enough to τ.

2.3. Programmatic change approach

Many programmatic changes occurred over the course of the P4V program. The key is to find programmatic changes that occurred uniquely in a single time period and therefore can be attributed to the single change rather than a host of program modifications. Additionally, control groups must exist within the P4V program.

2.3.1. Threshold shift

Each year, thresholds for every measure changed based on the national HEDIS distribution. Over the course of the P4V program, these thresholds starkly changed only once in 2015 for the breast cancer measure. Across all lines of business, all thresholds shifted up approximately 6 percentage points for the breast cancer measure while all other measure thresholds changed less than one percentage point (see Figure 11). The stark change occurred because the breast cancer screening measure definition was redefined in 2013 to include women ages 52-74. Previously women ages 42 - 69 were included. The change in measure definition occurred in 2013 for HMSA and national HEDIS collection efforts, however, the change in contract percentiles did not occur until 2015 as percentiles are two years lagged.[4] This policy change sets up a traditional difference-in-difference specification comparing quarterly

[4] For example in 2015, the available 2014 HEDIS scores were used to construct percentiles. The 2014 HEDIS scores were constructed using 2013 data.

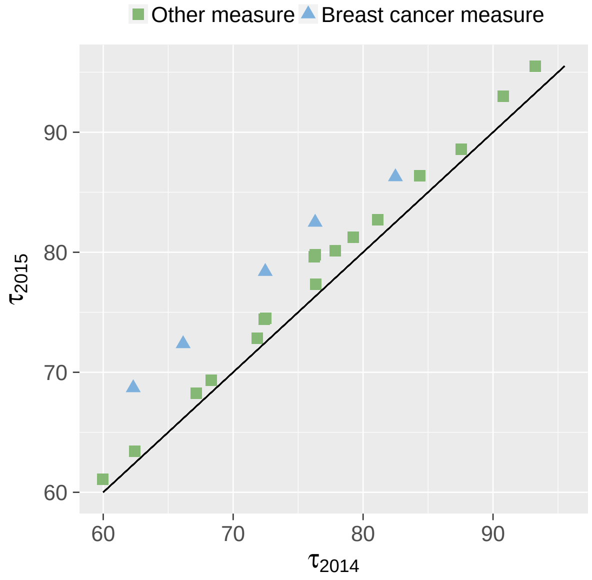
performance for breast cancer measures to quarterly performance for all other measures over time. The empirical specification is:

$$p_{ijt} = \kappa_j + \delta_t + \sum \gamma_t \mathbf{1}(\text{breast cancer}_j)\text{x}\mathbf{1}(\text{quarter}_t) + \lambda X_{it} + \epsilon_{ijt} \tag{2.2}$$

where p_{ijt} is the proportion of patients who are screened or attain a clinical outcome in physician i's panel for measure j during quarter t. Note that line of business subscripts are suppressed for ease of interpretation. Panel risk characteristics X_{it} are included to control for any shifts in panel composition over time. The measure fixed effects, κ_j, control for any time invariant performance differences across measures and quarter fixed effects, δ_t, control for any overall time-varying performance changes. Line of business fixed effects are also included to control for any time invariant performance differences across the lines of business. The coefficients of interest are γ_t which represent the quarterly performance in the breast cancer measure relative to performance in all other measures in quarter t. The threshold changes were implemented in the first quarter of 2015, therefore the fixed effect and interaction dummies representing the preceding time period are left out (2014Q4) and all γ_t coefficients are relative to this period. Observations from all lines of business are included as long as quarterly performance is based on over 10 patients. The regressions is clustered at the physician level.

In order to account for the possibility that physician-measures in different parts of the performance pay structure respond differentially to thresholds changes, I match physician-measures based on location in the pay structure. The experiment I have in mind compares physician-measure pairs that are similar distances away from the same threshold (e.g., 50th percentile threshold), and one observation experiences a shock to their distance. The location variables I construct include 1) the closest threshold at the beginning of a period (or the closest threshold to one's total performance conditional on doing nothing this period, $\hat{r}_{ijt}(p_{ijt} = 0)$) and 2) the percentage point distance from that threshold. Additionally, I match observations that have have similar trends in quarterly performance over time.

Figure 11: Shift of thresholds between 2014 and 2015 for all measures



Note: The black line is a 45° line. Thresholds include the 10th, 25th, 50th, 75th and 90th thresholds for the following measures - breast cancer, cervical cancer, colorectal cancer, diabetic eye and diabetic nephropathy screenings. Thresholds below 60 percent were dropped for ease of interpretation. This data selection dropped the 10th and 25th thresholds for diabetes eye screening and 10th threshold for colorectal cancer screening.

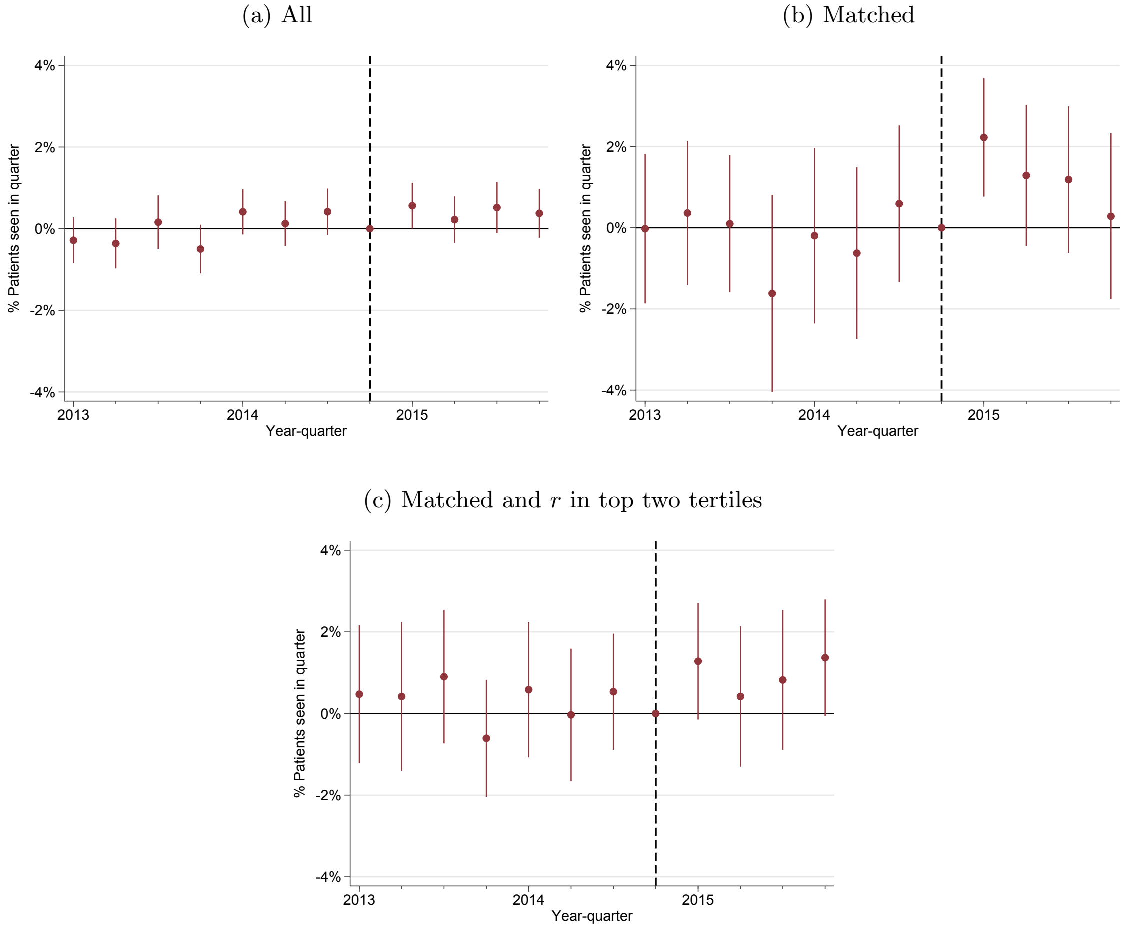
Specifically, I match using the k-nearest neighbor matching algorithm with the two location variables for the end of the pre-period (2014Q4) as well as the trend of quarterly performance during the pre-period. The purpose of this matching is to identify a treatment and control group with common support. Thus the matching exercise simply drops observations if their performance is outside of the treatment or control group's support and provides weighting to better match the distribution of performance across the two groups.

An additional concern is positive or negative correlation between the error terms within a physician. For example, an increase in payment for breast cancer could incentivize a physician to increase effort for all measures (a positive correlation). Alternatively, a physician could have a limited amount of effort to give each quarter and increasing payment for breast cancer could increase effort for breast cancer screening and decrease effort for other measures. To account for potential correlations among errors, I run one additional specification that 1) only includes physician-measure observations in the treatment group where the breast cancer measure represents a large portion of the physician's potential bonus and 2) only includes physician-measure observations in the control group where the breast cancer measure represents a small portion of the physician's potential bonus. The purpose of this exercise is to identify a control group least likely to be affected by the breast cancer threshold changes, and therefore minimize bias from positively or negatively correlated error terms. I identify the importance of the breast cancer measure to a physician using the ratio of potential bonus from breast cancer to potential bonus from all measures. Physicians with ratios in the top two ratio tertiles are included in the control group and physicians with ratios in the bottom tertile are included in the treatment group.[5] I then use the same matching procedure described earlier.

Results are presented in Figure 12 for multiple samples of the data (see Appendix Table 25 for precise point estimates). All physician-measure pairs are included in panel a to estimate the average effect for all physicians. Panel b only includes matched observations and panel c includes the more robust matched observations. The γ_t coefficients prior to 2014Q4 are insignificant or marginally significantly different from 0 for all panels largely satisfying the parallel trends assumption. The coefficients for 2015Q1 are generally positive and highly significant. On average, quarterly breast cancer performance increased 0.57 percentage points the period after the threshold change relative to quarterly performance for all other measures, representing a 12% single period increase (panel a). When focusing on only matched observations, the magnitude increases to 2.0 percentage points, representing a 37% increase (panel b). Finally, the coefficient in the more conservative panel c is marginally significant ($p < 0.05$) and has a magnitude of 1.3 percentage points, which is in between the first two panels. The smaller magnitude in the more conservative approach suggests the errors within a physician are negatively correlated, however the point estimates in the two exercises are not statistically different from one another.

[5] Specifically, I take the average proportion of potential bonus pay for the breast cancer measure prior to the threshold change. The tertile cutoffs are defined within the breast cancer measure and not across all measures.

Figure 12: Effect of threshold shift on performance, breast cancer measure case study


(a) All

(b) Matched

(c) Matched and r in top two tertiles

Notes: Figures plots the γ_t coefficients from estimating Equation 2.2. The specification plotted in Figure 12a includes all observations; the specification plotted in Figure 12b includes all physician-breast cancer measure pairs and their matched controls; and the specification plotted in Figure 12c includes only physician-breast cancer observations where the average potential bonus is in the top two tertiles and their matched controls (see text for details). All regressions include quarter fixed effects and time varying physician panel risk controls. In addition to sample restrictions described above, the physician-measure level observations had to represent at least 10 patients. Standard errors are clustered at the physician level.

Overall, a relative increase in threshold location by 5 percentage points resulted in quarterly performance improvement of 1.3 to 2 percentage points. This exercise suggests a change in threshold location affects quarterly performance for a single period following the change in location. This single period change is expected as bonus payment is based on total performance. To shift total performance, a single period performance increase would increase one's $\hat{r}$ for a number of periods.

2.3.2. Marginal bonus decrease

Total possible bonus pay increased between 2012 through 2015 due to modest increases in the per member per month scaling factor. Total possible bonus pay for individual measures decreased in 2013 due to an influx of new measures accompanied by a relatively small increase in total possible bonus pay. Unfortunately no good control group exists to estimate whether these changes affected performance. One measure specific change with a comparable control group occurred at the beginning of 2014. Prior to 2014, HMSA emphasized the diabetic nephropathy screening measure, giving it a weight four times that of most other measures. In 2014, the weight decreased with HMSA placing no additional weight relative to most other measures. The influx of new measures decreased possible bonus pay for preventative cancer and diabetic measures by $228 on average, while the influx of new measures and the change of weighting decreased possible bonus pay for nephropathy by $600 on average. This change once again sets up a difference-in-difference estimation strategy:

$$p_{ijt} = \kappa_j + \delta_t + \sum \gamma_t \mathbf{1}(\text{diabetic nephropathy}_j)\text{x}\mathbf{1}(\text{quarter}_t) + \lambda X_{it} + \epsilon_{ijt} \tag{2.3}$$

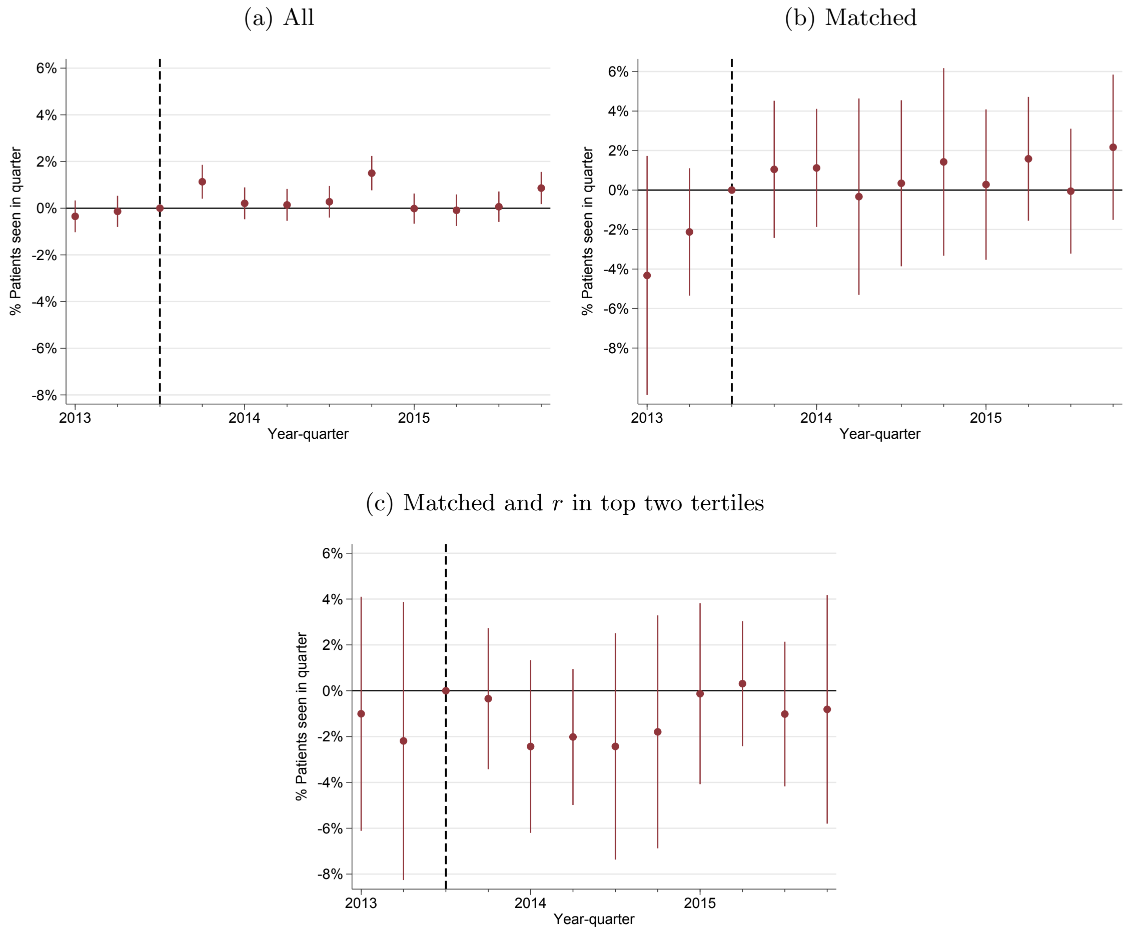
where the empirical strategy is the same as in Equation 2.2 except the "treated" measure is now diabetic nephropathy and the γ_t coefficients are all relative to the third quarter of 2013, the quarter of the price change announcement. I again use matching to account for the possibility that physician-measures in different parts of the performance pay structure respond differentially to payment changes. Similar to the matching algorithm used above, I use the k-nearest neighbor matching algorithm with the two location variables for the end of the pre-period (2013Q3) as well as the trend of quarterly performance during the pre-period. I also again run the more robust specification where the treated observations must have the diabetic nephropathy measure account for a large portion of their potential bonus and require all physician observations to be in either the treated or control group.[6]

Results are presented in Figure 13 (see Appendix Table 26 for precise point estimates). The γ_t coefficients for 2013Q1 and 2013Q2 are not significant in all specifications largely satisfying the parallel trends assumption, however panel b has perhaps a positive pre-trend. The specification with all observations (panel a) appears to have a seasonality effect where nephropathy performance increases relative to all other measure performances in the final quarter of the year. This seasonality effect is corrected when using matched observations to identify controls in panels b and c. However, the standard errors are significantly larger in the matched specifications and no post period γ_t coefficient is significant. The large standard deviations partially reflect a regression with a much smaller set of observations. Performance on the diabetic nephropathy measure is on average much higher than performance on other measures resulting in many dropped observations during matching. The number of observations in the diabetic nephropathy matched analysis is about half the size as the previous breast cancer analysis.

Overall, quarterly performance for nephropathy did not significantly change relative to matched controls after the relative decrease in bonus pay size for the nephropathy measure. When running the specification using year rather than quarter controls (only years 2013 and 2014 included), γ_{2014} remained insignificant. In particular, using the standard errors to bound the response, I cannot detect a differential response of 1.8 percentage points or less. Alternatively, for a 40% decline in bonus pay, the response, if present, must be less than 30%. Overall, this is a very noisy estimate and I cannot rule out a economically and clinically meaningful response.

[6] I take the average proportion of potential bonus pay for the diabetic nephropathy measure prior to the bonus amount change. The tertile cutoffs are defined within the diabetic nephropathy measure and not across all measures.

Figure 13: Effect of bonus shift on performance, diabetic nephropathy measure case study



(a) All

(b) Matched

(c) Matched and r in top two tertiles

Notes: Figures plots the γ_t coefficients from estimating Equation 2.3. The specification plotted in Figure 13a includes all observations; the specification plotted in Figure 13b includes all physician-diabetic nephropathy measure pairs and their matched controls; and the specification plotted in Figure 13c includes only physician-nephropathy observations where the average potential bonus is in the top two tertiles and their matched controls (see text for details). All regressions include quarter fixed effects and time varying physician panel risk controls. In addition to sample restrictions described above, the physician-measure level observations had to represent at least 10 patients. Standard errors are clustered at the physician level.

2.4. Direct estimation

2.4.1. Empirical approach

From theory, the choice of effort in a nonlinear payment structure depends on one's distance from the payment threshold and the size of the bonus payment. In the Hawaii context, a naive regression to recover the relevant parameters would be:

$$p_{ijt} = \nu_0 + \nu_1 d_{ijt} + \nu_2 m_{ijt} + p_{ijt-1} + \lambda \mathbf{X}_{it} + \zeta X_{ijt} + \delta_t + \eta_j + \epsilon_{ijt} \tag{2.4}$$

where the dependent variable, p_{ijt}, is the proportion of patients who are screened or attain a clinical outcome in provider i's panel for measure j during time period t. Recall the numerator in p_{ijt} only includes patients who have not yet received a screening. The main variables of interest are d_{ijt} and m_{ijt} defined as i) the distance between τ and one's performance at the beginning of the quarter and ii) the difference between the bonus received at τ and the predicted bonus received based on performance at the beginning of the quarter:

$$\begin{aligned} d_{ijt} &= \tau_{jt} - \hat{r}_{ijt}(p_{ijt} = 0) \\ m_{ijt} &= \hat{b}_{ijt}(p_{ijt} > d_{ijt}) - \hat{b}_{ijt}(p_{ijt} = 0) \end{aligned}$$

Distance d_{ijt} and marginal bonus m_{ijt} map directly to the model in Section 2.2 as $\tau - x_{t-1}$ and $u(\hat{b}(x_t = \tau - x_{t-1})) - u(\hat{b}(x_t = 0))$ respectively. For interpretability, τ is a single threshold so that an increase in d_{ijt} and s_{ijt} always implies a larger p_{ijt} is necessary to attain the threshold. The naive regression also includes year and quarter fixed effects, measure fixed effects, line of business fixed effects and a provider's panel level risk variables at the quarter level. Panel risk variables are a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbid condition. Lagged quarterly performance and the number of relevant patients for a measure are also added.[7] The

[7] Lagged quarterly performance helps control for a mechanical bias between the variables of interest and performance (see below). Logged measure panel size attempts to control for the measure maximum bonus amount.

question this regression attempts to answer is how will the proportion of patients meeting a measure this quarter, a signal of provider effort, differ between two similar physicians when faced with different distances and marginal bonuses.

Recall from Section 1.4.2 that physicians have different distances and marginal bonuses due to exogenous changes over time (annual changes to threshold locations, increase of per-member-per-month amount and a large pay structure change in 2012) and due to variation across physicians that may be correlated with the error term - previous effort and panel composition (e.g., number of total patients and distribution of patient types).

The main sources of bias in the above regression include a mechanical bias, patient selection, and unobserved physician characteristics. A mechanical bias arises because larger distance and the corresponding marginal bonus values imply a physician has a greater proportion of their panel who need to be screened. This larger set of potential patients implies the marginal cost to seeing a patient is lower than a physician who has fewer patients who need to be screened. The bias could lead to inflated ν coefficients. Patient selection exists because patients are not randomly sorted across physicians. Physicians with a patient panel more likely to visit and follow physician recommendations could have a smaller distance, marginal bonus and higher performance. Finally, there may be unobservable physician characteristics outside of patient selection. Physicians who are unobservably higher "quality" could have smaller distance, marginal bonus and higher performance. Both patient selection and physician unobserved characteristics would bias the ν coefficients downward.

To account for these various biases, I run specifications with physician and physician-measure fixed effects. Specifications with fixed effects conservatively account for unobserved physician characteristics. Additionally, concerns about patient selection will be accounted for in the fixed effect specifications conditional on patient selection not changing in response to the contract. Importantly, the current specification estimates physician response the quarter after a distance or marginal bonus change. Therefore, short term patient selection changes is unlikely to occur. The mechanical bias will not be corrected in the FE specification.

Additionally, I construct two instruments that serve as shocks each quarter to a physician's distance and marginal bonus: the change in "banked" patients, β and the increase in age relevant patients, α. The IV specifications do not at present include physician fixed effects due to the strength of the instrument, which will be discussed in section 2.4.3. The goal of the instruments is to take into account all of the biases listed above.

2.4.2. OLS and fixed effects specifications

As noted previously, fixed effects at the physician and physician-measure level conservatively account for unobserved physician characteristics and should account for patient selection.[8] One would expect the ν coefficients in the fixed effects specifications to be larger than the OLS specifications since the fixed effects correct for negative biases. It is less clear which set of coefficients, the physician or physician-measure level fixed effects, should be larger. If there is a positive correlation within physicians across measures in their performance, the physician fixed effect coefficients should be larger than the physician-measure fixed effects. Alternatively, if the correlation is negative, the physician fixed effects coefficients should be smaller. A negative correlation implies that physicians have a limited amount of effort each period to expend and increasing effort along one measure dimension decreases effort along other measure dimensions.

Table 3 presents OLS and physician and physician-measure level fixed effects results for the estimating Equation 2.4. Recall that observations are at the physician-measure-quarter level. As noted earlier, each specification defines distance and marginal bonus with respect to a single τ percentile. The column labels describe the relevant τ. Note as the τ percentile decreases, the sample size decreases because fewer observations have positive marginal bonus values. The variables of interest, distance and marginal bonus, are structured such that they are approximately normal - distance is in percentage points and marginal bonus is logged.

[8] Recall Equation 2.4 estimates physician response the quarter after a contract change. Therefore, changes in patient selection must occur the subsequent quarter, which is unlikely to occur.

Regressions include all providers who participate in all four years of the program, have at least 10 relevant patients for a measure, and are cluster at the physician level.

The OLS results suggest that as distance increased by one percentage point, the proportion of patients receiving a preventative screening in that quarter increased by about 0.04 percentage points. The coefficient on logged marginal bonus is significant for all specifications. As the logged marginal bonus increased by ten percent or about a 0.1 standard deviation, the proportion of patients receiving a preventative screening in that quarter changes by -0.03 to 0.05 percentage points. A decrease in quarterly performance in response to an increase in marginal bonus is surprising, however the magnitude is perhaps not clinically meaningful. With average quarterly performance of 5%, the coefficients corresponds to an extremely small average elasticity between -0.001 and 0.001.

As anticipated the physician and physician-measure fixed effects coefficients for distance are greater than the OLS coefficients. The coefficients generally double in size with physician fixed effects and triple in size with physician-measure fixed effects. As the distance increased by one percentage point, the proportion of patients receiving a preventative screening in that quarter increased by about 0.1 percentage points. The coefficients on logged marginal bonus are not consistently larger in the fixed effects specifications, but the difference between coefficients across specifications are often not statistically meaningful. Further, the average elasticity implied by all coefficients on logged marginal bonus are below 0.001 and therefore not economically meaningful. Finally, the smaller distance coefficients in the physician fixed effects specification compared to the physician-measure specification suggests a negative correlation between quarterly performance and distance. The fixed effects results generally align with expectations, suggest that physicians are somewhat responsive to their distance from a threshold, and have a limited amount of effort each period to expend across all measures.

Table 3: Effect of distance and marginal bonus on performance, OLS and FE

	Quarterly performance								
		τ_{90}			τ_{75}			τ_{50}	
	OLS	FE	FE	OLS	FE	FE	OLS	FE	FE
Distance to τ (pct)	0.0421***	0.0820***	0.143***	0.0430***	0.0844***	0.141***	0.0403***	0.0819***	0.128***
	(0.00372)	(0.00391)	(0.00551)	(0.00394)	(0.00419)	(0.00572)	(0.00433)	(0.00461)	(0.00618)
Ln marginal bonus	0.00479***	0.00505***	0.00361***	0.00153*	0.00248***	0.00274***	-0.00301***	-0.00193*	-0.00186
	(0.000551)	(0.000567)	(0.000631)	(0.000663)	(0.000696)	(0.000779)	(0.000864)	(0.000920)	(0.00100)
Performance, 1 qtr lag	0.0767***	0.0519***	-0.0149	0.0836***	0.0563***	-0.0152	0.0875***	0.0594***	-0.0170
	(0.00946)	(0.00901)	(0.00871)	(0.0101)	(0.00963)	(0.00935)	(0.0122)	(0.0118)	(0.0114)
Ln relevant panel size	-0.00401***	-0.00565***	-0.00176	-0.00102	-0.00321**	-0.00106	0.00327***	0.00137	0.00353*
	(0.000722)	(0.00102)	(0.00130)	(0.000816)	(0.00111)	(0.00143)	(0.000983)	(0.00132)	(0.00160)
Observations	37780	37780	37780	33626	33626	33626	26745	26745	26745
R^2	0.402	0.439	0.516	0.410	0.649	0.803	0.413	0.599	0.767
Physician FE		x			x			x	
Physician-measure FE			x			x			x

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Regressions include only physician-measure-quarter observations that represent more than 10 patients. Controls include lagged quarterly performance, logged number of patients relevant for a measure, line of business, measure fixed effects, quarter fixed effects, and year fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Standard errors are clustered at the physician level.

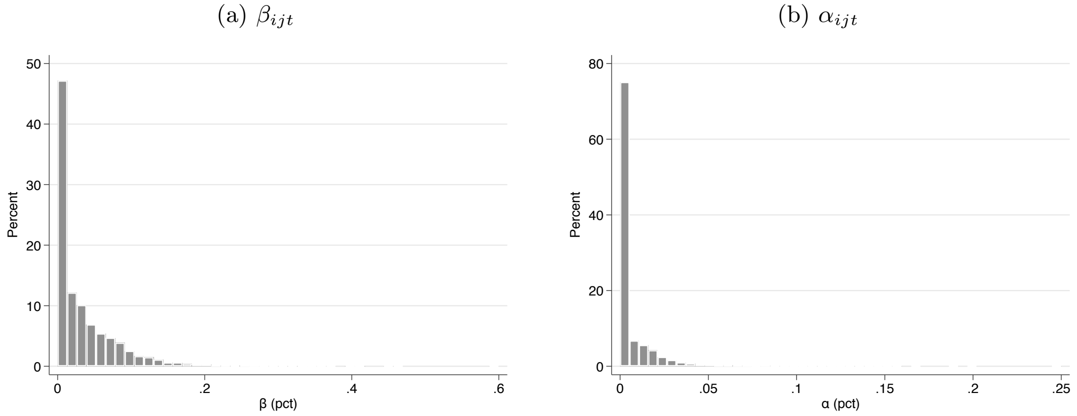
2.4.3. Instruments

The two instruments take advantage of the moving time and age window used to define total performance. A patient must have a screening or outcome met once a year or once every multiple years. At the beginning of each quarter a number of patients who previously satisfied a measure due to the screening or lab test date, no longer satisfy that measure and are therefore no longer “banked”. Similarly, only patients of a certain age are relevant for each measure. At the beginning of each quarter, the number of patients who are relevant to a measure will change as patients age into the age requirements.

The instrument β is defined as the difference between the performance in the previous period and the predicted performance in the current period conditional on the provider doing nothing and conditional on the current quarter's panel composition ($\beta_{ijt} = \hat{r}_{ijt-1}(p_{it-1}, D_{ijt}) - \hat{r}_{ijt}(p_{ijt} = 0, D_{ijt})$). The instrument α is defined as the proportion of patients in the panel in period t who just aged into the measure requirements. Importantly, both of these measures are defined as changes using a provider's current panel to ensure variation is only due to changes in “banked” patients or aging and not panel composition as provider panels slightly shift composition over time particularly at the beginning of a year due to insurance churn (individuals drop HMSA coverage).

The goal of these instruments is to pick up variation in distance and marginal bonus that is unrelated to 1) where a physician is in the payment scheme (mechanical bias), 2) unobserved physician characteristics and 3) patient panel composition. Ideally, physicians would be randomly assigned distances and marginal bonus conditional on the other controls. The instruments serve to bring in this quasi-randomness by acting as shocks to a provider's distance and marginal bonus. For example, 2 years, 3 months ago a patient received a mammography and now her breast cancer screening has lapsed - that provider has one less patient “banked” than he did at the end of the previous measurement period and therefore distance has a positive shock. Furthermore, the farther away a person is from the threshold, the larger the marginal bonus. Note that a provider will always have a non-negative shock

Figure 14: Distribution of instruments


(a) β_{ijt}

(b) α_{ijt}

Note: Histograms of the two instruments for all physician-measure observations that represent at least 10 relevant patients.

to their distance and marginal bonus so the shocks differ in magnitude rather than sign. Similarly, a 58 year old female patient not captured last period by the breast cancer measure will be captured in the current period. For this to be a positive shock to distance, the patient cannot have received screened prior to measure inclusion. Recommendations from the USPTF and HEDIS have age specifications for strong clinical reasons thus I argue that newly age relevant patients will not have met the quality measure. The size of a measure bonus depends on the number of relevant patients as well as the distance from the threshold thus positive α's are also positive shocks to marginal bonus. Figure 14 plots the distribution of β and α for the breast cancer commercial measure. Sixty-six percent of the β mass and thirty-eight percent of the α mass is above 0. Further, the β and α are both defined as a percent of one's patient panel and thus the support of each instrument can be directly compared. Figure 14 also demonstrates that the variation in β is significantly larger than the variation in α.

The instrumented ν_1 and ν_2 coefficients represent the local average treatment effect (LATE). A strength of this instrument is that shocks occur to all observations at different points in time. A limitation is that certain types of observations mechanically experience larger variations in these shocks. Intuitively, observations with smaller relevant panel sizes experience larger variation in both instruments because instruments are constructed as a percent of the relevant panel size. To examine the extent of these mechanical relationships, I look at the mean and standard deviation of the instrument by lagged panel size quantiles. Importantly, I use a lagged variable because the instruments directly affect the values in time t. In Table 4, the standard deviation declines as the relevant panel size grows as expected due to the construction of β and α. It is less obvious how the mean values for β and α should change. Reassuringly, the mean instrument values are relatively consistent for the first four quintiles, but dramatically decreases for β and increases for α in the largest quintile. The changes in the mean values suggest that the largest panels are correlated with a younger and healthier panel (larger α values) and a higher $\hat{r}$ (larger β values). However, the table also demonstrates that the LATE is largely identified by observations with small to medium sized relevant panels who do not have these correlations.[9]

Table 4: Instrument value by relevant panel size quantile, 1 quarter lagged

Relevant Panel Size Quantile	β	α
1st Quantile	0.038 (0.062)	0.004 (0.020)
2nd Quantile	0.037 (0.050)	0.004 (0.015)
3rd Quantile	0.034 (0.040)	0.005 (0.013)
4th Quantile	0.034 (0.037)	0.006 (0.010)
5th Quantile	0.019 (0.026)	0.008 (0.008)

Notes: All physician-measure-quarter observations are included that have an average relevant panel size above 10.

First stage

Tables 5 and 6 present the first stage results regressing distance and marginal bonus, respectively, on both instruments and all controls. The controls and other regression set up is identical to the specifications described in Section 2.4.2.

[9] Additionally, I look at the balance of the instrument along patient risk characteristics. I find that along observable risk characteristics, patients are well balanced across β. However, I do find that a healthier panel is more likely to experience a high α. This suggests that the variation α leverages is in panels where it is easier to get a newly age relevant patient screened and could positively bias the results. These patient risk

Table 5: First stage, Marginal bonus for surpassing τ

	Ln marginal bonus for surpassing τ				
	τ_{90}	τ_{75}	τ_{50}	τ_{25}	τ_{10}
β_{ijt} (pct)	1.971^{***}	0.975^{***}	0.0213	0.252	0.0185
	(0.0911)	(0.0900)	(0.0833)	(0.203)	(0.160)
α_{ijt} (pct)	0.763^{**}	1.011^{***}	0.948^{***}	0.670^{*}	0.614^{*}
	(0.294)	(0.269)	(0.253)	(0.334)	(0.289)
Quarterly performance, 1 qtr lag	-0.463^{***}	-0.372^{***}	-0.134	-0.272^{*}	-0.271^{**}
	(0.0709)	(0.0742)	(0.0748)	(0.133)	(0.0970)
Ln relevant panel size	0.918^{***}	0.925^{***}	0.944^{***}	0.914^{***}	0.949^{***}
	(0.00709)	(0.00670)	(0.00623)	(0.00855)	(0.00779)
Observations	37779	33625	26744	14247	10197
R^2	0.870	0.892	0.920	0.894	0.949

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Regressions include only physician-measure-quarter observations that represent more than 10 patients. Controls include lagged quarterly performance, logged number of patients relevant for a measure, line of business, measure fixed effects, quarter fixed effects, and year fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Standard errors are clustered at the physician level.

Table 6: First stage, Distance to τ

	Distance to τ				
	τ_{90}	τ_{75}	τ_{50}	τ_{25}	τ_{10}
β_{ijt} (pct)	0.391^{***}	0.297^{***}	0.228^{***}	0.287^{***}	0.316^{***}
	(0.0198)	(0.0202)	(0.0217)	(0.0438)	(0.0528)
α_{ijt} (pct)	-0.0520	0.0680	0.125	-0.0464	0.00342
	(0.0776)	(0.0756)	(0.0743)	(0.0795)	(0.0766)
Performance, 1 qtr lag	-0.252^{***}	-0.270^{***}	-0.276^{***}	-0.394^{***}	-0.387^{***}
	(0.0162)	(0.0168)	(0.0185)	(0.0315)	(0.0354)
Ln relevant panel size	-0.0350^{***}	-0.0349^{***}	-0.0347^{***}	-0.0353^{***}	-0.0321^{***}
	(0.00217)	(0.00214)	(0.00215)	(0.00236)	(0.00239)
Observations	37779	33625	26744	14247	10197
R^2	0.468	0.445	0.413	0.357	0.285

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Regressions include only physician-measure-quarter observations that represent more than 10 patients. Controls include lagged quarterly performance, logged number of patients relevant for a measure, line of business, measure fixed effects, quarter fixed effects, and year fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Standard errors are clustered at the physician level.

Coefficients on β presented in Table 5, the first stage results for marginal bonus, are not consistent across different specifications (e.g., defining distance based on different thresholds). The β coefficient mechanically declines as the τ percentile decreases. This occurs because β affects marginal bonus by shifting one's location in the pay schedule at the beginning of the quarter to a lower τ percentile and as one moves down the distribution, there are fewer lower τ percentiles. A one percentage point increase in β is associated with between a 1 to 2 percent marginal bonus increase for specifications with τ_{75} and τ_{90}, respectively. The α coefficients remain relatively constant across specifications. A one percentage point increase in α is associated with between a 0.6 and 1.0 percent increase in marginal bonus. The α coefficients are highly significant for τ_{90}, τ_{75}, and τ_{50}. The consistency is expected as α affects total bonus and is independent of where one is in the payment schedule. Lagged quarterly performance is generally negatively associated with logged marginal bonus as observations with better previous performance are in higher parts of the pay schedule at the beginning of the quarter (i.e., larger $\hat{r}$) leading to smaller marginal bonus. Finally, logged relevant panel size is highly correlated with logged marginal bonus with all coefficients above a 0.9. As relevant panel size increases, one's potential bonus increases so this high correlation is expected.

Coefficients presented in Table 6, the first stage results on distance, are relatively stable across all specifications. As the percent of "unbanked" patients, β, increases by one percentage point, the distance to τ increases by between 0.23 and 0.39 percentage points. All β coefficients are highly significant. The consistency across specification is expected as an increase in β increases distance from each threshold equally. The percent of new patients, α does not significantly impact distance in any specification. Coefficients for lagged quarterly performance and logged relevant panel size are negative demonstrating that providers who performed better last quarter and providers with larger panel sizes are in higher parts of the pay schedule at the beginning of the quarter (i.e., larger $\hat{r}$) and therefore closer to the threshold.

balance exercises are detailed in Appendix Section A.5

2SLS results

Table 7 presents the main 2SLS results for the estimating Equation 2.4, quarterly performance regressed on distance to τ, logged marginal bonus for surpassing τ, and all controls. The sample constructions are the same as those presented in the first stage results. The OLS results are discussed in Section 2.4.2 and included for comparison. Reviewing the main sources of bias, physician unobservables and patient selection are expected to negatively bias the OLS distance and marginal bonus estimates and the mechanical bias is expected to positively bias the OLS estimates. Therefore, it is unclear whether the instrumented coefficients should be higher or lower than the OLS coefficients.

The 2SLS results include tests of under identification, weak identification and the Anderson-Rubin test of whether or not the endogenous regressors are jointly equal to zero, which is robust to weak instruments. With two endogenous regressors and two instruments, values around 7 are acceptable for the under and weak identification test statistics. Specifications for τ_{90}, τ_{75}, and τ_{50} satisfy these tests, but specifications with the lowest τ's do not. As the first stage results suggest, the instruments do not identify marginal bonus in specifications with the lowest τ's. The null that the endogenous regressors are jointly equal to zero is strongly rejected for all of the specifications.

Focusing on the specifications for τ_{90}, τ_{75}, and τ_{50}, the coefficients on distance grow in magnitude, while the coefficients on marginal bonus are no longer significant relative to the OLS results. The magnitude of the coefficients span from 0.5 to 1.2 and this lower bound, which represents the majority of observations, is in the range of the results from Section 2.3.1. The consistently significant distance coefficients demonstrate that quarterly performance will increase with an increase in the threshold. Mirroring the OLS results, the sign on the marginal bonus coefficients is at times negative, however no coefficients are significant as the standard errors are now relatively large. The size of the standard errors between 0.050 and 0.110 suggest I cannot detect an elasticity that is less than between

Table 7: Effect of distance and marginal bonus on performance, 2SLS

	Quarterly performance									
	τ_{90}		τ_{75}		τ_{50}		τ_{25}		τ_{10}	
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
Distance to τ	0.0421***	0.508*	0.0430***	1.152**	0.0403***	1.275***	0.0233***	0.778***	0.0212***	0.712***
	(0.00372)	(0.255)	(0.00394)	(0.372)	(0.00433)	(0.125)	(0.00447)	(0.145)	(0.00541)	(0.141)
Ln marginal bonus	0.00479***	0.0528	0.00153*	-0.0552	-0.00301***	-0.139	-0.000884	-0.00398	-0.00325	-0.0681
	(0.000551)	(0.0504)	(0.000663)	(0.110)	(0.000864)	(0.104)	(0.000849)	(0.0911)	(0.00174)	(0.104)
Qtr performance, 1 qtr lag	0.0767***	0.215***	0.0836***	0.359***	0.0875***	0.406***	0.104***	0.399***	0.0923***	0.340***
	(0.00946)	(0.0438)	(0.0101)	(0.0656)	(0.0122)	(0.0469)	(0.0175)	(0.0594)	(0.0171)	(0.0680)
Ln relevant panel size	-0.00401***	-0.0317	-0.00102	0.0901	0.00327***	0.174	0.00168	0.0310	0.00436*	0.0879
	(0.000722)	(0.0551)	(0.000816)	(0.114)	(0.000983)	(0.0993)	(0.000924)	(0.0857)	(0.00171)	(0.0992)
Observations	37779	37779	33625	33625	26744	26744	14247	14247	10197	10197
Under-identification Test										
Kleibergen-Paap stat		6.966		6.224		12.753		4.643		4.384
Weak-identification Test										
Cragg-Donald stat		7.791		6.399		11.761		4.007		5.911
Test for end x's equal to 0										
Anderson-Rubin stat		340.369		265.093		193.530		47.899		30.953

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Regressions include only physician-measure-quarter observations that represent more than 10 patients. Controls include lagged quarterly performance, logged number of patients relevant for a measure, line of business, measure fixed effects, quarter fixed effects, and year fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Standard errors are clustered at the physician level.

0.01 and 0.02. These elasticities are an order of magnitude larger than those in the OLS specification, however they are still relatively low. To increase quarterly performance by one percent, total pay out for these measures, which was on average $7.5 million, would have to increase between 50 to 100%.

2.5. Discussion

This chapter uses multiple estimation strategies to identify physician responses to nonlinear contract characteristics. Motivated by theory, I estimate the average physician response to a physician's distance from the threshold and marginal bonus amount.

In the first estimation strategy, I identify two natural experiments and apply a difference-in-difference framework. I find that a relative breast cancer threshold increase of 5 percentage points led to a 1 to 2 percentage point improvement in performance in the subsequent quarter relative to the performance of matched observations. I find no average differential response in quarterly performance after a 40% relative decrease in nephropathy bonus pay relative to the performance of the matched controls. While the budget of the HMSA performance pay scheme did not change during the studied time period, the decrease in nephropathy pay represents a transfer of over one million dollars to bonus pay for other measures with no average impact on nephropathy performance.

In the second set of estimation strategies I directly estimate physician responses to changes in the contract features. I use various fixed effects specifications and construct instruments to correct for numerous sources of bias present in the direct estimation regression. The fixed effects specifications conservatively control for negative sources of bias and as expected, result in coefficients that are of similar or larger magnitude than the OLS coefficients. The instruments - the change in "banked" patients, β and the increase in age relevant patients, α - aim to control for both the negative and positive sources of bias, but do so in a less conservative manner than the fixed effects specifications. It is therefore unclear whether the distance and marginal bonus coefficients should be larger or smaller than those in the

OLS and fixed effects. Importantly, the instrumented coefficients largely leverage variation in small to medium sized physician panels and potentially panels with healthier patients. The LATE is therefore quite different than the average treatment effect in the OLS and fixed effects specifications.[10][11] The instrumented results demonstrate that physicians are responsive to changes in distance to a threshold, an increase in thresholds by one percentage point resulted in quarterly performance increases of between 0.5 - 1.2 percentage points. This lower bound is of similar magnitude to results from the breast cancer difference-in-difference analysis. Also, I find that on physicians are not responsive to marginal bonus payment in the quarter after a payment change, again similar to the difference-in-difference nephropathy results. Using the standard errors to bound this finding, I cannot detect an increase in quarterly performance of 0.1 percentage points following a one percentage point increase in pay. These results demonstrate that the size of the bonus payment will have little effect on provider effort on average. Small increases in thresholds improve performance without increasing cost.

This chapter has a number of limitations. One lingering question is the interaction between these two contract features. In 2012, the payment schedule was restructured (see Figure 3) such that the step size and distance drastically increased for physicians in the lower portion of the distribution. It was not possible to directly estimate physician response to this striking and concurrent change in contract features. The second set of limitations pertain to the instruments. Variation in both instruments was driven by physicians with smaller relevant panel sizes. The number of "unbanked" or age relevant patients did not perfectly scale up with physicians relevant panel sizes - therefore higher instrument values were typically physicians with smaller relevant panel sizes. The IV specification therefore estimated the local average treatment effect for this part of the physician panel size distribution, rather the

[10] The larger 2SLS distance coefficients relative to the OLS specification suggest the negative biases are larger than the positive mechanical bias. Additionally, the larger 2SLS distance coefficients relative to the fixed effects specification suggest the fixed effects wipe out a significant amount of variation across physicians that the is used by the 2SLS estimates.

[11] When the fixed effects specifications are subsetted by panel size, the coefficients grow in magnitude (results not shown). This finding is reassuring based on the larger 2SLS vs fixed effects distance coefficients. The difference between coefficients is partially driven by differences in average vs LATE effects.

full distribution. Additionally, the second instrument α was marginally strong and did not always well identify marginal bonus. This was due to the relatively small variation in this instrument - fewer individuals aged into measures than became “unbanked”. Specifications with all measures are robust to weak and under identification. Finally, the IV specification relied on variation in a single period to drive changes in that quarter’s performance. The current specification does not capture responses that take place over a longer period of time.

This section can be extended along a number of dimensions. A simple extension would be to add forthcoming data on intermediate outcome measures. One could examine the substitutability or complementarity between these measures and their related process measures. Separately, one particularly attractive feature of this setting is the detailed and known contract design. In the future, more structure could be developed around a physician’s decision of effort and that model can be directly taken to the data. For example, adding a second period to the current model does not produce any interesting dynamic results. The model could be extended multiple periods and include multiple thresholds.

CHAPTER 3 : Mechanisms

3.1. Introduction

This chapter explores various mechanisms for the responses found in the preceding chapter and provides insight into how the effects are generated. The first exercise explores responses by expected high and low performing physicians. The distribution of responses by physician performance level are important to policy makers and insurers as these actors typically cite improving performance for the lowest quality physicians to be a primary goal of performance pay programs.

The next exercise focuses on the timing of physician effort. Previous work finds agent effort across time to vary in order to take advantage of the nonlinear pay schedule. Potential negative impacts of varying effort across time periods is discussed in health care as well as in other settings. I run various specifications at the month level to estimate differential responses across time.

A final exercise explores the types of patients who are screened due to the performance pay contract. I estimate whether the additional patients screened due to the performance pay contract are on average more or less risky and more or less costly. These final two exercises provide more detail on how physicians respond to the nonlinear performance pay contract.

3.2. Physician type: High and low performers

A natural extension to the average response findings in Chapter 2 is to determine whether there are certain types of physicians who are more or less responsive to changes in the contract features. Unfortunately the theoretical model developed in Section 2.2 does not provide any insights into different physician responses unless more assumptions are placed on the marginal cost function or physician types are directly added to the model. The variable explored in this section is a physician's location in the contract structure or $\hat{r}$. Alternatively, this can be interpreted as expected high and low performers. Differential responses for observations in higher and lower performance pay structure locations have important policy implications as incentivizing low performing providers has been cited as an objective in P4V programs and some previous programs have found larger responses to the introduction of a scheme for lower performing providers (e.g., Greene et al., 2015). To evaluate whether high and low performers have different responses to changes in the contract features, I repeat the main analysis in Chapter 2 subsetting the data into expected high and low performers.

Programmatic change approach

I revisit the two case studies in Section 2.3 that applied a difference-in-difference framework to changes in threshold locations and bonus amounts for individual measures. For the breast cancer screening (or threshold shift) case study I define observations to be in high and low performance pay locations based on $E[\hat{r}_{ij2014}(p_{ij2014} = 0)]$ being above or below the 2014 50th percentile, τ_{2014}. Similarly, I define observations using $E[\hat{r}_{ij2013}(p_{ij2013} = 0)]$ and the 2013 50th percentile, τ_{2013}, for the diabetic nephropathy (or bonus amount shift) case study. Henceforth, I will describe these two types of physicians as low and high performing physicians.

Figure 15 plots the γ_t coefficients, which represent the quarterly performance for the breast cancer measure relative to performance in all other measures in quarter t. All γ_t coefficients are relative to the period prior to the change. Additionally, Figure 15 plots two specifications of equation 2.2. The top two panels include all matched pairs and the bottom panels include breast cancer observations that represents a large portion of a physician's bonus. Recall matching accounts for the possibility that physician-measures in different parts of the performance pay structure respond differentially to thresholds changes. And, the inclusion of breast-cancer observations with a large portion of a physician's bonus accounts for concerns of positively or negatively correlated errors (for further details see Section 2.3.1).

The γ_t coefficients are not statistically significant in the quarters prior to the change in thresholds for the all panels. However the high performing physicians have noisy pre-trends, particularly in panel b, and large confidence intervals. The quarter after a 5 percentage point relative threshold increase, low performing physicians increased their quarterly performance by 2 percentage points relative to all other measure performance. This change in performance is marginally significant in panel c. These results mirrors the overall results. No γ_t coefficients are significant in any quarter for the high performing physicians, however it is important to note that the coefficients in the quarter after the change in thresholds (2015Q1) are around 2 percentage points. These results suggest that low performing physicians may be more responsive to changes in threshold locations relative to high performing physicians, however the large confidence intervals and imprecise pre-trends for the high performing physicians limit the confidence of this conclusion.

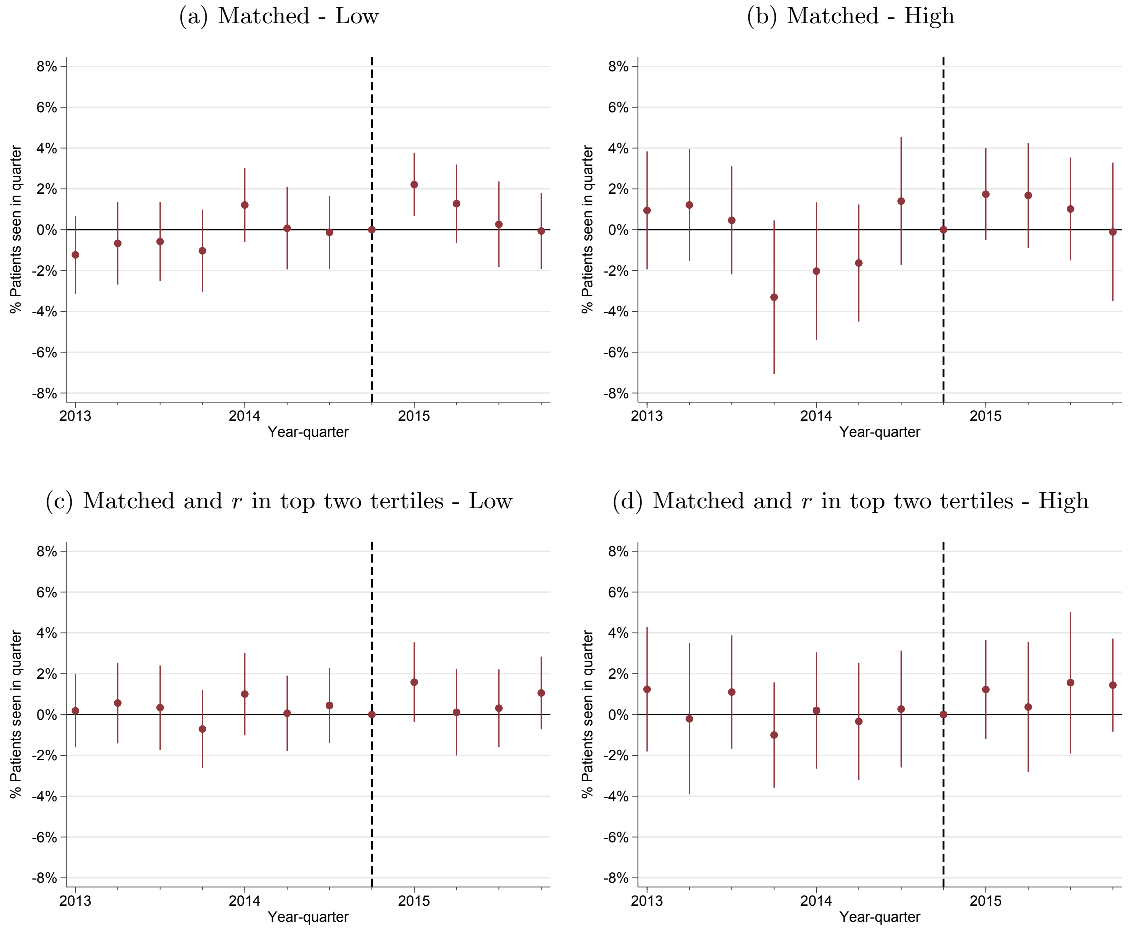
Figure 16 repeats the analysis in Figure 15 for the diabetic nephropathy case study. The γ_t coefficients for the two quarters prior to the diabetic payment change are not statistically significant, but have somewhat worrisome pre-trends in panels a and b. Across all panels, no coefficients are statistically significant, however the confidence intervals are quite large. These results where no differential change in quarterly performance was detected for high or low performers are parallel to the overall results.

Direct estimation

I revisit the instrumented version of the direct estimation specification (Equation 2.4) and again specify observations as being low and high performers. To categorize observations, I define observations based on $\hat{r}$ and the corresponding 50th percentile threshold, τ_{50}. Table 8 presents results that run the τ_{90} specification on observations with total performance at the beginning of the quarter (or $\hat{r}_{ijt}(p_{ijt} = 0)$) below and above τ_{50} respectively.[1] Both specifications satisfy the weak- and under-identifying tests. Low performers ($\hat{r} < \tau_{50}$) are more responsive to distance from threshold than high performers. In fact, the total response to distance in Table 7 appears driven by these observations as the other distance

[1] Recall, each specification uses a single threshold to define all distance and marginal bonus values.

Figure 15: Effect of threshold shift on performance by high and low type, breast cancer measure case study



(a) Matched - Low

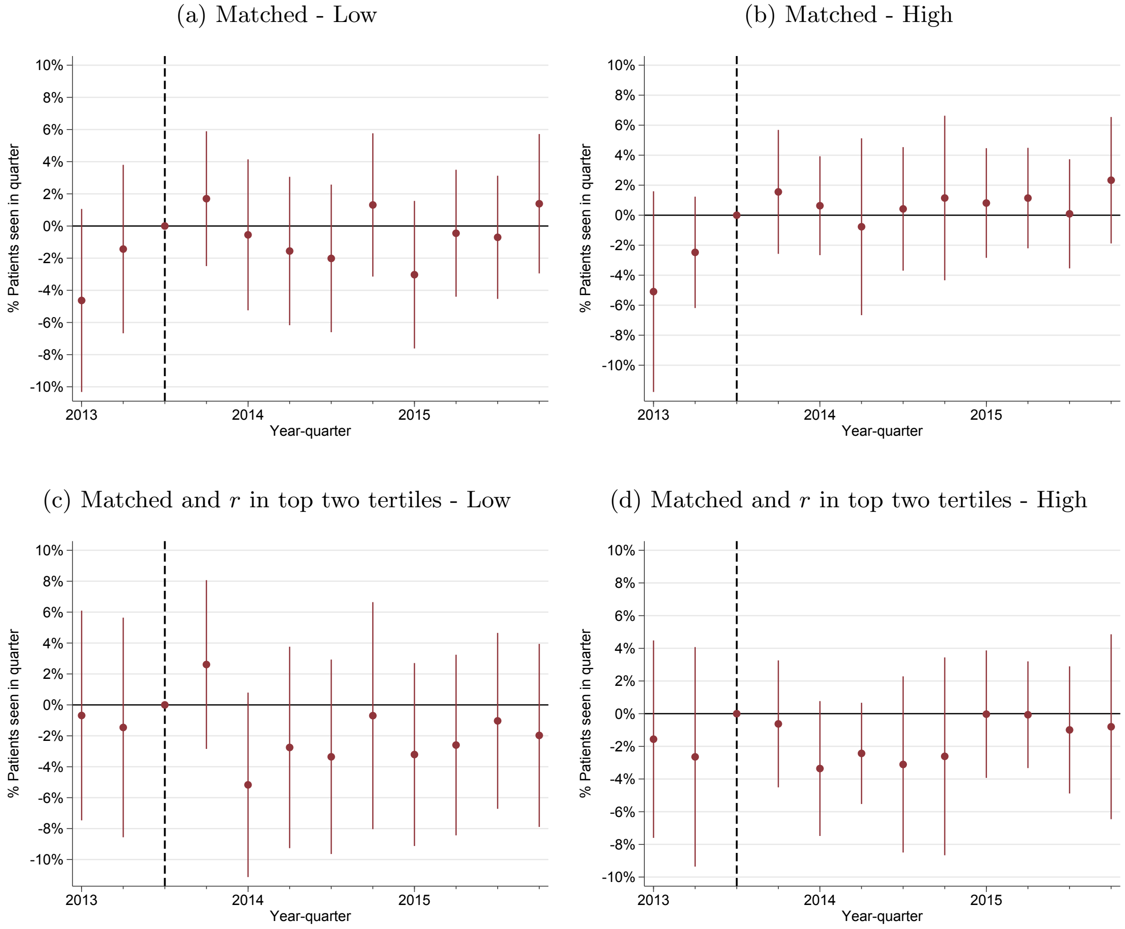
(b) Matched - High

(c) Matched and r in top two tertiles - Low

(d) Matched and r in top two tertiles - High

Notes: Figures plots the γ_t coefficients from estimating Equation 2.2. The specifications plotted in Figure 15a and b include physician-breast cancer measure pairs close to thresholds and their matched controls. The specifications plotted in Figure 15c and d include physician-breast cancer observations where the average potential bonus is in the top two tertiles and their matched controls (see text for details). All regressions include quarter fixed effects and time varying physician panel risk controls. In addition to sample restrictions described above, the physician-measure level observations had to represent at least 10 patients. Standard errors are clustered at the physician level.

Figure 16: Effect of bonus size decrease on performance by high and low type, diabetic nephropathy measure case study



(a) Matched - Low

(b) Matched - High

(c) Matched and r in top two tertiles - Low

(d) Matched and r in top two tertiles - High

Notes: Figures plots the γ_t coefficients from estimating Equation 2.3. The specifications plotted in Figure 16a and b include physician-breast cancer measure pairs close to thresholds and their matched controls. The specifications plotted in Figure 16c and d include physician-breast cancer observations where the average potential bonus is in the top two tertiles and their matched controls (see text for details). All regressions include quarter fixed effects and time varying physician panel risk controls. In addition to sample restrictions described above, the physician-measure level observations had to represent at least 10 patients. Standard errors are clustered at the physician level.

coefficient is not significant. Observations with larger $\hat{r}$'s are responsive to step size, a ten percent increase in the marginal bonus leads to a 0.014 percentage point increase in quarterly performance. However, the implied average elasticity remains relatively low at 0.03. These results suggest that low performing physicians are more responsive to distance from a threshold than high performing physicians and that high performing physicians are weakly responsive to the marginal bonus.

Table 8: Effect of distance and marginal on performance for τ_{90} by location in pay structure, 2SLS

	Quarterly performance	
	$\hat{r}_{ijt} < \tau_{jt,50}$	$\hat{r}_{ijt} \geq \tau_{jt,50}$
Distance to τ_{90}	1.169***	-0.330
	(0.177)	(0.331)
Ln marginal bonus	0.146	0.137***
	(0.0925)	(0.0251)
Quarterly performance, 1 qtr lag	0.400***	0.0958***
	(0.0492)	(0.0193)
Ln relevant panel size	-0.0994	-0.135***
	(0.0921)	(0.0251)
Observations	26744	11035
Under-identification Test		
Kleibergen-Paap stat	17.981	34.879
Weak-identification Test		
Cragg-Donald stat	21.524	28.888
Test for end x's equal to 0		
Anderson-Rubin stat	193.530	97.822

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Regressions include only physician-measure-quarter observations that represent more than 10 patients. Controls include lagged quarterly performance, logged number of patients relevant for a measure, line of business, measure fixed effects, quarter fixed effects, and year fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Standard errors are clustered at the physician level.

3.3. Physician timing of effort

Next, I explore whether the contract features affect the timing of physician effort. In other settings with nonlinear contracts, agents maximize pay by altering the timing of their effort, which negatively impacts the firm (Oyer, 1998; Larkin, 2014). In a health care setting, manipulating when a certain type of patient comes in for a visit may or may not impact important health outcomes. Key to whether the timing of effort affects important outcomes is to identify the types of patients who are crowded out by the preventative screening visits. For example, if annual physicals are pushed to the beginning of a quarter rather than smoothly administered overtime, other health outcomes are likely not impacted. Alternatively, if preventative screening visits crowd out same day or higher acuity patients who then must go to the Emergency Department, other important health outcomes are likely negatively impacted.

To examine whether physician effort changes overtime, I run the same direct estimation specification considered in Section 2.4.3 at the month level. The original outcome is the percent of patients screened during the quarter. Now the outcome is the percent of patients screened during the month. All other variables are the same including distance from a threshold, d_{ijt} and marginal bonus for surpassing a threshold, m_{ijt}, which are defined by performance at the beginning of the quarter.

$$p_{ijm} = \nu_0 + \nu_1 d_{ijt} + \nu_2 m_{ijt} + p_{ijt-1} + \lambda \mathbf{X}_{it} + \zeta X_{ijt} + \delta_t + \eta_j + \epsilon_{ijm} \tag{3.1}$$

Table 9 presents the distance and marginal bonus coefficients for the instrumented version of the above equation where each specification includes only one month of each quarter. Additionally, the regressions are run using the 90th and 50th percentile thresholds to define the distance and marginal bonus variables. The controls and standard errors are defined as in previous specifications.

Similar to the overall instrumented results, the specifications using the 90th percentile threshold marginally pass the under and weak identification tests with values around 7. The specifications using the 50th percentile threshold have values over 7. The coefficients on marginal bonus are generally not significant. The coefficients on distance are significant, but smaller than the coefficients in the main specification at the quarter level in Table 7. In the quarter level specifications, the 90th percentile distance coefficient was 0.5 and the 50th percentile distance coefficient was 1.3. The month level coefficients are about one third of the quarter level coefficients, which is to be expected. Finally, the coefficients on distance are relatively consistent across months with a small positive trend that is not significant.[2][3] The consistency of the distance coefficients across months demonstrates that physicians do not manipulate effort across months.

Table 9: Effect of distance and marginal bonus on performance by month, 2SLS

	Monthly performance					
		τ_{90}			τ_{50}	
	Month 1	Month 2	Month 3	Month 1	Month 2	Month 3
Distance to τ (pct)	0.0294	0.200*	0.235*	0.347***	0.442***	0.450***
	(0.0791)	(0.0988)	(0.107)	(0.0472)	(0.0543)	(0.0541)
Ln marginal bonus	0.0358*	0.0143	0.00685	-0.0112	-0.0579	-0.0638
	(0.0155)	(0.0195)	(0.0213)	(0.0309)	(0.0373)	(0.0388)
Observations	37779	37779	37779	26744	26744	26744
Under-identification Test						
Kleibergen-Paap stat	6.966	6.966	6.966	12.753	12.753	12.753
Weak-identification Test						
Cragg-Donald stat	7.791	7.791	7.791	11.761	11.761	11.761
Test for end x's equal to 0						
Anderson-Rubin stat	85.106	114.782	122.684	51.874	59.468	76.908

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Regressions include only physician-measure-month observations that represent more than 10 patients. Controls include lagged quarterly performance, logged number of patients relevant for a measure, line of business, measure fixed effects, quarter fixed effects, and year fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Standard errors are clustered at the physician level.

3.4. Patient characteristics

The final section focuses on the types of patients who are screened due to the performance pay contract. Based on the model developed in Section 2.2, one would expect less costly

[2] When the observations are stacked and interactions between month and distance and month and marginal bonus are included, the month interactions with distance are not significant in both the 90th and 50th percentile thresholds.

[3] Fixed effects at the physician and physician-measure level are included in the Appendix (see Tables 29 and 30). The results are very similar with consistent distance coefficients across months that are smaller than the quarter level results. Marginal bonus coefficients are also significant and consistent, but very small.

patients to be seen first or alternatively, the more difficult and least compliant patients are screened later. While, it is not clear how to define difficult or compliant patients, patient risk and spending may be correlated with patient cost. Patient risk could be associated with compliance as patients with worse health outcomes are more likely to regularly visit a physician. However, the opposite could also be true - patients have worse health outcomes because they do not receive regular care. A similar argument could be made about the association of patient health spending and compliance - high spending suggests that patients seek care more frequently or that patients seek care infrequently, but when they do, care is expensive due to the extended wait. Alternatively, risk and spending are perhaps a signal of marginal benefit. For risk, one might think that higher risk patients would most benefit from preventative screenings. Similarly, with spending, a high spending patient may suggest a more comorbid patient who should be screened.

I estimate characteristics of the additional patients screened due to the performance pay contract using the original direct estimation specification, Equation 2.4. I replace the dependent variable with the average risk or spending of the patients a physician screens in the quarter. This construction implies that when a physician screens no patients, the dependent variable is missing. The construction does not directly estimate characteristics of the marginal patient, but does directly reveal whether the marginal patients are increasing or decreasing in risk and spending. In particular, the distance coefficient represents the change in average risk or spending of patients screened when distance is increased by one percentage point. From the original analysis, increasing distance increases the patients screened so the coefficient on distance can further be interpreted as the change in average risk for screening more patients.

Two patient risk scores are included - the Elixhauser comorbidity mortality score and an insurer generated ERG risk score. Both risk scores are calculated using claims data from the prior year. The Elixhauser risk score is constructed to predict in-hospital mortality while the ERG risk score is constructed to predict annual health care utilization.[4] The spending variable is current year total spending.

Tables 10 through 12 present the results using instrumented and fixed effects specifications. Fixed effects specifications are included because the coefficients are more precise than 2SLS coefficients in Chapter 2 and may provide suggestive evidence of a relationship that the 2SLS specification cannot identify.[5] Additionally, results using the 90th and 50th percentiles are included. Note the sample sizes are slightly different from the original specifications in Section 2.4.3 because observations were dropped when no patients were screened and not all observations had an associated ERG score. The differences in sample size also drive the differences in the first stage tests across Tables 10 through 12 and in Section 2.4.3.

Results of the Elixhauser mortality risk score, Table 10, demonstrate there is generally no average difference in risk score for changes in distance and marginal bonus amount.[6]

Across instrumented specifications in Table 11, distance and marginal bonus amount do not affect the average ERG score.[7] However, both fixed effects specifications have significant distance and marginal bonus coefficients. Note that the instrumented specifications do not pass the under and weak identification test. Also, recall that both fixed effects specifications are conservatively accounting for two negative biases and do not account for an important positive bias. Therefore the fixed effects results are suggestive and not conclusive. In the fixed effects specifications, the additional patients seen due to performance pay contracts are higher in risk. A one percentage point increase in patients screened increases the average ERG risk of all patients screened between 3 and 6% .[8] The coefficients on marginal

[4] ERG was originally developed by Optum and optimized from Medicare claims data. From other work using HMSA claims data, ERG risk scores track well with total spending.

[5] Recall, the fixed effect specifications do not correct for the positive mechanical bias and could upwardly bias these results.

[6] The Elixhauser mortality score at the patient level is zero inflated with very long and thin tales. Taking the average at the physician level for patients who are screened creates a more compact, but similar distribution. The average risk score is 0.52 with a standard deviation of 4.5 suggesting it will be difficult to find significant changes in the average patient seen.

[7] ERG score has a similar distribution at the patient level as total spending. Additionally, since this variable is averaged at the physician level, there are few 0 values. To make this dependent variable approximately normal, I use a log specification.

[8] This increase in risk combines the point estimates on distance in the original specification from Table

Table 10: Effect of distance and marginal bonus on Elixhauser mortality risk score

	Risk score					
		τ_{90}			τ_{50}	
	IV	FE	FE	IV	FE	FE
Distance to τ (pct)	-14.81	-0.0956	0.372	-7.656	0.224	0.951*
	(12.25)	(0.282)	(0.397)	(4.735)	(0.344)	(0.467)
Ln marginal bonus	1.808	-0.0114	-0.0622	6.938	-0.0488	-0.206*
	(2.409)	(0.0548)	(0.0590)	(4.967)	(0.0926)	(0.0999)
Observations	30931	30932	30932	21982	21983	21983
Under-identification Test						
Kleibergen-Paap stat	7.102	-	-	5.559	-	-
Weak-identification Test						
Cragg-Donald stat	8.283	-	-	4.984	-	-
Physician fixed effects		x			x	
Physician-measure fixed effects			x			x

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Regressions include only physician-measure-quarter observations that represent more than 10 patients. Controls include lagged quarterly performance, logged number of patients relevant for a measure, line of business, measure fixed effects, quarter fixed effects, and year fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Standard errors are clustered at the physician level.

bonus also generally suggest a positive increase in average risk, however the point estimates imply unreasonable average risk score increases when combined with quarterly performance estimates.

Finally, Table 12 presents results for total spending.[9] As in previous results, coefficients for the instrumented versions are not significant. The under and weak identification tests are marginally acceptable for the 90th percentile threshold, but not for the 50th percentile threshold. Coefficients on marginal bonus are not significant for any specification. Coefficients on distance are significant and positive for both fixed effects specifications and suggest a one percentage point increase in patients screened increases the average screened

3 (dependent variable is percent of patients screened) and the current Table 11. For example, the 90th percentile threshold with the physician fixed effects specification has point estimates of 0.082 and 0.746. To find the average ERG risk increase for a 1 percentage point increase in patients screened I divide 0.746 by 0.082.

[9] Total spending is logged to make the dependent variable approximately normal. Since all patients have to visit a physician and total spending is averaged across a physician, there are ultimately very few 0's and therefore very few dropped observations.

Table 11: Effect of distance and marginal bonus on logged ERG score

	Ln ERG score					
		τ_{90}			τ_{50}	
	IV	FE	FE	IV	FE	FE
Distance to τ (pct)	3.250	0.497***	0.513***	-0.104	0.399***	0.414***
	(2.879)	(0.0577)	(0.0799)	(0.828)	(0.0745)	(0.100)
Ln marginal bonus	-0.523	-0.0289*	-0.0310*	-0.826	-0.0716***	-0.0738***
	(0.540)	(0.0113)	(0.0125)	(0.968)	(0.0171)	(0.0192)
Observations	27387	27387	27387	19080	19080	19080
Under-identification Test						
Kleibergen-Paap stat	5.065	-	-	4.980	-	-
Weak-identification Test						
Cragg-Donald stat	5.696	-	-	3.849	-	-
Physician fixed effects		x			x	
Physician-measure fixed effects			x			x

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Regressions include only physician-measure-quarter observations that represent more than 10 patients. Controls include lagged quarterly performance, logged number of patients relevant for a measure, line of business, measure fixed effects, quarter fixed effects, and year fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Standard errors are clustered at the physician level.

patient spending between 4 and 9%.[10]

3.5. Discussion

This chapter explores mechanisms for the average response found in the previous chapter. One dimension of physician heterogeneity is included - high and low physician performance - and I find evidence of low performing physicians responding more to changes in distance from a threshold. Additionally, I find high performing physicians responsive to the marginal bonus amount, albeit a small response. I did not find evidence that physicians altered their response to the contract among months in a quarter. This result alleviates a potential concern of the performance contract crowding out certain patient visits. Finally, I find evidence that the marginal patients screened due to the performance pay contract are higher risk and higher spending patients relative to those already screened. This results suggests that the marginal screening benefit is increasing. The last result is not supported

[10] See Footnote 7

Table 12: Effect of distance and marginal bonus on logged total spending for screened patients

	Ln patient total spending					
		τ_{90}			τ_{50}	
	IV	FE	FE	IV	FE	FE
Distance to τ (pct)	-1.622	0.746***	0.535***	-0.551	0.762***	0.506***
	(2.603)	(0.0781)	(0.0967)	(1.006)	(0.0970)	(0.118)
Ln marginal bonus	0.362	-0.0160	0.00730	0.658	-0.00493	-0.0135
	(0.512)	(0.0144)	(0.0154)	(0.991)	(0.0219)	(0.0246)
Observations	30931	30932	30932	21982	21983	21983
Under-identification Test						
Kleibergen-Paap stat	6.937	-	-	5.483	-	-
Weak-identification Test						
Cragg-Donald stat	8.283	-	-	4.984	-	-
Physician fixed effects		x			x	
Physician-measure fixed effects			x			x

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Regressions include only physician-measure-quarter observations that represent more than 10 patients. Controls include lagged quarterly performance, logged number of patients relevant for a measure, line of business, measure fixed effects, quarter fixed effects, and year fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Standard errors are clustered at the physician level.

with the most robust specification and should be interpreted as suggestive.

CHAPTER 4 : Conclusion

This paper estimates physician responses to nonlinear contract features. First, I develop a theoretical model to identify specific contract components that affect a physician's choice of effort - distance from a threshold and marginal bonus amount. In the main empirical analysis, I use difference-in-difference and instrumental variable approaches to estimate physician response to both of these contract components. Physician level fixed-effects are also considered. Results from the preferred specification of both approaches demonstrate that a one percentage point increase in a threshold location increases quarterly performance by 0.3 to 0.5 percentage points in the subsequent period. I did not detect a response to a large decrease in marginal bonus for a single measures using the difference-in-difference specification. I also did not detect an average response to changes in marginal bonus using the instrumental variable specification.

In the final chapter, I focus on mechanisms for the average response. I find heterogeneous responses to changes in both features. Physicians with expected lower performance are more responsive to changes in distance in both specifications. Physicians with expected high performance are responsive to changes in marginal bonus in the instrumental variable specification, however the responsiveness is low. I do not find evidence of physicians exerting more effort towards the end of a quarter alleviating concerns of negative access spillovers. I also find suggestive evidence of the marginal screening benefit increasing as the additional patients screened due to the contract are potentially higher risk and higher cost patients.

Physician performance improved slightly from 74.8% to 75.5% between 2012 and 2015.[1] The corresponding average threshold change was between -0.4 and 0.5 percentage points.[2] A back of the envelop calculation suggests the change in thresholds changed performance

[1] Physician performance is defined for the commercial line of business and for the six measures focused on in the dissertation.

[2] Thresholds changes average across all six measures and are weighted by physician (i.e., accounts for the prevalence of the measures - more physicians had a breast cancer measure than diabetic measure). The range comes from the inclusion of all thresholds (10th, 25th, 50th, 75th, and 90th) and inclusion of only the last three thresholds.

between -0.2 and 0.3 percentage points. Thresholds locations in the program studied are tied to national averages and did not increase significantly between 2012 and 2015. Had the thresholds instead increased overtime, results suggest performance would have increased more.

Overall, this dissertation demonstrates that physicians are weakly responsive to changes in nonlinear contract features particularly changes to the marginal bonus. This is of particular importance because the health care field is moving towards more linear contracts including Medicare's new Merit Based Incentive Payment Scheme (MIPS) as nonlinear contracts are often assumed to be suboptimal. However, the main feature a linear scheme has to manipulate is the marginal bonus amount.

Economic models find nonlinear schemes to be suboptimal under three main assumptions: 1) effort is continuous, 2) agents are sophisticated and 3) more complex contract structures are costly to construct (Mirrlees, 1971). The results that physicians are responsive to threshold locations and not marginal bonus amounts suggest the second assumption may be violated. A growing body of literature in health and many other areas have found agents to be inattentive and more responsive to salient attributes in goods, taxes, and finances (Handel and Kolstad, 2015; Chetty et al., 2009; Malmendier and Nagel, 2011). In this setting, physicians may interpret thresholds to be more salient and changes in the marginal bonus amount to be undetectable. Note, I only estimate responses to changes in a contract feature and not responses to the introduction to a new contract. Additionally, all responses are within the support of my data and represent local changes. Perhaps larger increases in the marginal bonus would elicit a greater response. Nonetheless the size of the incentives (bonus amount) in the contract studied are large relative to many other in the literature making this new evidence on the responsiveness to various nonlinear contract features compelling.

APPENDIX

A.1. P4V Measures

Table 13: Quality Measure Names Over Tover timeime by Line of Business

	Commercial					QUEST				Akamai Advantage			
	2011	2012	2013	2014	2015	2012	2013	2014	2015	2012	2013	2014	2015
Preventative Services - Breast cancer screening	x	x	x	x	x	x	x	x	x	x	x	x	x
Preventative Services - Cervical cancer screening	x	x	x	x	x	x	x	x	x				
Preventative Services - Colorectal cancer screening	x	x	x	x	x	x	x	x	x	x	x	x	x
Preventative Services - Chlamydia screening for women	x	x	x	x	x	x	x	x	x				
Preventative Services - BMI assessment				x	x			x	x			x††	x††
Preventative Services - Advance care planning				x	x							x†	x†
Diabetes - eye exam	x	x	x	x	x	x	x	x	x	x	x	x	x
Diabetes - LDL-C screening	x	x	x			x	x						
Diabetes - HbA1C testing	x	x	x			x	x						
Diabetes - Nephropathy	x	x	x	x	x	x	x	x	x	x	x	x	x
Diabetes - Blood pressure control <140/90				x	x			x	x			x†	x†
Diabetes - HbA1C poor control (>9% or not measured)				x	x							x†	x†
Diabetes - Med adherence, oral diabetes med				x	x			x	x	x*	x*	x†	x†
Diabetes - HbA1c control (<8%)								x	x				
Diabetes - HbA1C <9%										x*	x*		
Diabetes - LDL-C <100mg/mL										x*	x*		
Diabetes - Comprehensive Diabetes Treatment										x	x	x	x
Asthma - Appropriate medication	x	x	x	x	x	x	x	x	x				
Asthma - Spirometry testing for COPD		x	x			x	x						
Asthma - Avoidance of antibiotic treatment for bronchitis		x	x	x	x	x	x	x	x				
Heart Disease - LDL-C screenings	x	x	x			x	x			x	x		
Heart Disease - ACE or ARB		x	x	x	x	x	x	x	x	x*	x*		
Heart Disease - Annual monitoring for members on diuretics		x	x	x	x	x	x	x	x				
Heart Disease - Controlling blood pressure				x	x			x	x	x*	x*	x†	x†
Heart Disease - Medication adherence for cholesterol (statins)				x	x			x	x	x*	x*	x†	x†
Heart Disease - Med adherence for hypertension				x	x			x	x			x†	x†
Peds: Preventive - Chlamydia screening		x	x	x	x	x	x	x	x				
Peds: Preventive - Well-child first 15 months		x	x	x	x	x	x	x	x				
Peds: Preventive - Well-child in 3rd, 4th, 5th and 6th years		x	x	x	x	x	x	x	x				
Peds: Preventive - BMI assessment				x	x			x	x				
Peds: Respiratory - Appropriate testing for pharyngitis		x	x	x	x	x	x	x	x				
Peds: Respiratory - Appropriate treatment for URI		x	x	x	x	x	x	x	x				
Peds: Respiratory - Appropriate medications for asthma		x	x	x	x	x	x	x	x				
Peds: Immunization		x	x	x	x	x	x	x	x				
Adolescent: Immunization		x	x	x	x	x	x	x	x				
Review of Chronic Conditions										x**	x	x‡	x‡‡
Per Member Per Month Dollar Amount	$2	$2	$4	$4.50	$4.50	$3	$3	$3	$3	$2	$2	1$	$1

* PMPM = $4, ** PMPM = $8, † PMPM = $2, †† PMPM = $0.25, ‡ = $5, ‡‡ = $6.50

Table 14: Commercial HMSA weights for measures over time

Measure Name	2011	2012	2013	2014	2015
Preventative screening - Breast cancer	1	1	1	1	1
Preventative screening - Cervical cancer	1	1	1	1	1
Preventative screening - Colorectal cancer	1	1	1	1	1
Preventative screening - Chlamydia	1	1	1	1	1
Diabetes care - eye exam	1	1	1	1	1
Diabetes care - LDL-C screening	1	1	1		
Diabetes care - HbA1C testing	2	2	2	2	
Diabetes care - medical attention for nephropathy	4	4	4	1	1
Diabetes care - Blood pressure control				2	2
Diabetes Care HbA1c Poor Control (>9%)					2
Asthma care - Use of appropriate medication	3	3	3	3	3
Cardiovascular Condition - LDL-C screening	1	1	1		
Acute Bronchitis - Avoidance of antibiotics		1	1	1	1
Heart Disease Care - ACE/ARB		1	1	1	1
Heart Disease Care - diuretics		1	1	1	1
Medication adherence - Oral diabetes medication				3	3
Medication adherence - Hypertension medications				3	3
Medication adherence - Statins				3	3
COPD - Spirometry testing		1	1		
Advance care planning				2	2
Body Mass Index				0.25	0.15
Hypertension - Blood Pressure control (<140/90)				2	2
Well-child visits in first 15 months of life	3	3	3	3	3
Well-child visits in the 3-6 years of life	2	2	2	2	2
Childhood immunization status	4	4	4	4	4
Immunizations for adolescents		1	1	1	1
Appropriate testing for children with pharyngitis	2	2	2	2	2
Appropriate treatment for children with URI	1	1	1	1	1

Table 15: Description of Measures and Any Changes over time

Diabetes |Eye Exam	Measure Description	Percentage of diabetes patients 18—75 years of age who received a dilated eye exam, seven standard field stereoscopic photos with interpretation by an ophthalmologist or optometrist, or imaging validated to match diagnosis from these photos during the measurement period. A negative dilated eye exam (negative for retinopathy) in the prior measurement period also meets criteria for the eye exam indicator.
	Years	2011-2015
	Modifications	None
Diabetes |HbA1c Testing	Measure Description	Percentage of patients with diabetes 18—75 years of age who receive one or more HbA1c test(s) per measurement period.
	Years	2011-2013
	Modifications	None
Diabetes |Nephropathy	Measure Description	Percentage of diabetes patients 18—75 years of age with at least one test for microalbumin during the measurement period or evidence of medical attention for existing nephropathy (diagnosis of nephropathy or documentation of microalbuminuria or albuminuria).
	Years	2011-2015
	Modifications	None

Breast Cancer Screening	Measure Description	The percentage of women 52–69 years of age as of the end of the measurement period who had one or more mammograms to screen for breast cancer during the measurement period or the 12 months prior to the measurement period.
	Years	2011-2015
	Modifications	2011-2013 included women ages 42—69, 2014-2015 included women ages 52-74 and used 15 month look-back period rather than 12 month
Cervical Cancer Screening	Measure Description	The percentage of women 24–64 years of age who were screened for cervical cancer using cervical cytology, which must be performed every three years.
	Years	2011-2015
	Modifications	2011-2013 included women ages 21—64, 2014-2015 could instead use the criteria: Women age 30–64 who had cervical cytology and human papillomavirus (HPV) co-testing performed every five years
Colorectal Cancer Screening	Measure Description	Percentage of adults 51–75 years of age who had appropriate screening for colorectal cancer. Either 1) Fecal occult blood test (FOBT) during the measurement period. 2) Flexible sigmoidoscopy during the measurement period or the four prior measurement periods. 3) Colonoscopy during the measurement period or the nine prior measurement periods.
	Years	2011-2015
	Modifications	None

Table 16: Commercial HEDIS National Percentiles for all Measures

Measure name	Year	10th Ptl	25th Ptl	50th Ptl	75th Ptl	90th Ptl
Breast Cancer Screening	2011	63.68	68.89	73.01	79.89	82.62
Breast Cancer Screening	2012	62.73	68.52	73.04	77.00	80.98
Breast Cancer Screening	2013	63.60	68.47	72.95	75.76	79.97
Breast Cancer Screening	2014	62.31	66.15	72.47	76.31	82.49
Breast Cancer Screening	2015	68.70	72.37	78.39	82.49	86.28
Cervical Cancer Screening	2011	69.09	72.84	77.89	81.08	85.45
Cervical Cancer Screening	2012	70.27	72.79	77.24	79.92	86.15
Cervical Cancer Screening	2013	68.83	72.93	77.17	80.70	86.20
Cervical Cancer Screening	2014	68.33	71.84	76.34	79.25	84.36
Cervical Cancer Screening	2015	69.33	72.84	77.34	81.25	86.36
Colorectal Cancer Screening	2011	48.91	56.95	62.96	69.06	74.20
Colorectal Cancer Screening	2012	49.07	57.31	63.95	70.30	75.08
Colorectal Cancer Screening	2013	51.09	59.25	66.08	71.86	74.06
Colorectal Cancer Screening	2014	49.56	59.95	67.16	72.41	76.24
Colorectal Cancer Screening	2015	50.56	61.07	68.27	74.41	79.61
Diabetic Eye Screening	2011	36.71	48.69	61.31	71.31	77.78
Diabetic Eye Screening	2012	37.96	47.54	62.04	71.34	78.59
Diabetic Eye Screening	2013	37.08	51.54	60.85	70.80	76.40
Diabetic Eye Screening	2014	34.67	48.91	62.41	72.51	77.87
Diabetic Eye Screening	2015	35.67	49.91	63.41	74.51	80.11
Diabetic Nephropathy Screening	2011	75.91	80.49	84.81	90.01	92.53
Diabetic Nephropathy Screening	2012	74.95	80.26	85.63	89.78	92.7
Diabetic Nephropathy Screening	2013	77.65	81.00	87.34	90.51	93.19
Diabetic Nephropathy Screening	2014	76.33	81.11	87.56	90.79	93.25
Diabetic Nephropathy Screening	2015	79.80	82.72	88.56	93.00	95.51
Diabetic HbA1c Screening	2011	84.84	86.77	90.31	93.06	94.09
Diabetic HbA1c Screening	2012	84.84	87.58	90.30	93.41	94.24
Diabetic HbA1c Screening	2013	86.31	88.08	90.88	94.16	95.40
Diabetic HbA1c Screening	2014	.	.	.	.	.
Diabetic HbA1c Screening	2015	.	.	.	.	.

Table 17: Medicaid Managed Care HEDIS National Percentiles for all Measures

Measure name	Year	10th Ptl	25th Ptl	50th Ptl	75th Ptl	90th Ptl
Breast Cancer Screening	2012	38.66	45.29	52.40	57.37	62.92
Breast Cancer Screening	2013	36.80	44.82	50.46	56.58	62.76
Breast Cancer Screening	2014	41.72	46.51	51.32	57.71	62.88
Breast Cancer Screening	2015	47.59	52.21	58.37	67.12	73.34
Cervical Cancer Screening	2012	53.04	64.04	69.72	74.24	78.65
Cervical Cancer Screening	2013	51.85	61.81	69.09	73.24	78.51
Cervical Cancer Screening	2014	47.22	59.15	66.42	71.95	76.64
Cervical Cancer Screening	2015	48.22	60.15	67.42	73.95	78.64
Colorectal Cancer Screening	2012	49.07	57.31	63.95	70.30	75.08
Colorectal Cancer Screening	2013	51.09	59.25	66.08	71.86	74.06
Colorectal Cancer Screening	2014	49.56	59.95	67.16	72.41	76.24
Colorectal Cancer Screening	2015	50.56	60.95	68.16	74.41	78.24
Diabetic Eye Screening	2012	33.97	43.82	52.85	63.75	70.64
Diabetic Eye Screening	2013	36.25	45.03	52.88	61.75	69.72
Diabetic Eye Screening	2014	37.14	44.37	54.31	62.46	67.64
Diabetic Eye Screening	2015	38.23	47.25	55.31	65.14	70.04
Diabetic HbA1c Screening	2012	73.58	77.59	82.19	87.09	90.84
Diabetic HbA1c Screening	2013	74.90	78.54	82.38	87.01	91.13
Diabetic HbA1c Screening	2014	.	.	.	.	.
Diabetic HbA1c Screening	2015	.	.	.	.	.
Diabetic Nephropathy Screening	2012	68.12	73.90	78.48	82.48	86.86
Diabetic Nephropathy Screening	2013	68.43	73.48	78.70	83.03	86.93
Diabetic Nephropathy Screening	2014	69.76	75.00	79.23	82.73	85.84
Diabetic Nephropathy Screening	2015	72.43	76.67	81.05	85.11	88.86

Table 18: Medicare Advantage CMS Percentiles for all Measures

Measure name	Year	10th Ptl	25th Ptl	50th Ptl	75th Ptl	90th Ptl
Breast Cancer Screening	2012	55.47	61.76	68.56	77.13	82.92
Breast Cancer Screening	2013	55.47	61.76	68.56	77.13	82.92
Breast Cancer Screening	2014	55.47	61.76	68.56	77.13	82.92
Breast Cancer Screening	2015	59.42	66.56	72.41	80.27	85.00
Cervical Cancer Screening	2012	.	.	.	.	.
Cervical Cancer Screening	2013	.	.	.	.	.
Cervical Cancer Screening	2014	.	.	.	.	.
Cervical Cancer Screening	2015	.	.	.	.	.
Colorectal Cancer Screening	2012	40.05	48.66	56.94	70.70	77.56
Colorectal Cancer Screening	2013	40.05	48.66	56.94	70.70	77.56
Colorectal Cancer Screening	2014	40.05	48.66	56.94	70.70	77.56
Colorectal Cancer Screening	2015	51.00	57.84	66.45	73.53	79.86
Diabetic Eye Screening	2012	49.67	56.19	64.72	74.66	80.28
Diabetic Eye Screening	2013	49.67	56.19	64.72	74.66	80.28
Diabetic Eye Screening	2014	49.67	56.19	64.72	74.66	80.28
Diabetic Eye Screening	2015	56.79	64.50	70.84	78.83	84.69
Diabetic HbA1c Screening	2012	.	.	.	.	.
Diabetic HbA1c Screening	2013	.	.	.	.	.
Diabetic HbA1c Screening	2014	.	.	.	.	.
Diabetic HbA1c Screening	2015	.	.	.	.	.
Diabetic Nephropathy Screening	2012	84.67	86.81	89.09	92.56	94.92
Diabetic Nephropathy Screening	2013	84.67	86.81	89.09	92.56	94.92
Diabetic Nephropathy Screening	2014	84.67	86.81	89.09	92.56	94.92
Diabetic Nephropathy Screening	2015	87.43	90.05	92.31	95.92	98.11

A.2. Selection

Table 19: Comparing Risk of Provider Attribution to Other Attribution

	Provider Attrib	Attrib Other	Z score
AIDS	**0.0004**	**0.001**	**-2.677**
Alcohol abuse	0.011	0.010	1.300
Anemia deficiency	0.080	0.078	0.794
Rhumatoid arthritis	0.019	0.014	5.245
Blood loss anemia	0.008	0.008	-0.881
CHF	0.015	0.015	0.556
Chronic pulmonary disease	0.106	0.101	2.196
Coagulation deficiency	0.014	0.011	2.906
Depression	0.046	0.030	9.773
Diabetes w/o complications	0.111	0.099	4.732
Diabetes w complications	0.038	0.023	10.109
Drug abuse	**0.011**	**0.01**3	**-2.128**
Hypertension	0.280	0.250	7.943
Hypothyroidism	0.084	0.058	11.811
Liver disease	0.028	0.025	2.292
Lymphoma	0.002	0.002	0.967
Fluid and electrolyte disorder	0.044	0.042	1.087
Metastatic cancer	0.006	0.005	1.722
Other neurological	0.025	0.023	1.744
Obesity	0.139	0.076	22.543
Paralysis	0.005	0.005	0.954
Peripheral vascular disesase	0.026	0.021	3.851
Psychoses	0.018	0.014	3.851
Pulmonary circulation disorder	0.005	0.004	0.635
Renal failure	0.035	0.032	1.716
Tumor	0.025	0.023	2.263
Ulcer	0.001	0.001	2.633
Valvular disease	0.033	0.025	5.562
Weightloss	0.020	0.018	1.933

Table 20: Comparing Risk of Never Switchers to Switchers

	Never Switch	Switch	Z score
Aids	0.001	0.001	-0.788
Alcohol abuse	0.009	0.011	-6.705
Deficiency anemias	**0.081**	**0.069**	**16.718**
Rheumatoid arthritis	0.014	0.015	-3.130
Blood loss anemia	0.008	0.009	-1.517
CHF	**0.015**	**0.013**	**8.564**
Chronic pulmonary disease	0.101	0.103	-2.964
Coagulation deficiency	0.011	0.011	1.480
Depression	0.028	0.036	-17.421
Diabetes w/o complications	**0.105**	**0.084**	**26.054**
Diabetes w complications	0.023	0.025	-5.395
Drug abuse	0.012	0.015	-11.600
Hypertension	**0.263**	**0.213**	**41.535**
Hypothyroidism	0.057	0.063	-8.238
Liver disease	0.026	0.025	2.414
Lymphoma	0.002	0.002	-0.950
Fluid and electrolyte disorders	**0.043**	**0.039**	**7.376**
Metastic cancer	0.005	0.004	2.026
Other neurological	0.023	0.023	0.410
Obesity	0.065	0.114	-69.717
Paralysis	0.005	0.005	0.098
Peripheral vascular disorder	**0.021**	**0.020**	**3.255**
Psychoses	0.013	0.017	-13.365
Pulmonary circulation disorder	0.005	0.004	1.342
Renal failure	**0.034**	**0.028**	**13.055**
Tumor	**0.024**	**0.019**	**11.465**
Ulcer	0.001	0.001	-4.088
Valvular disorder	**0.026**	**0.023**	**7.152**
Weight loss	**0.019**	**0.017**	**5.585**

A.3. Definition of B_i

$B_j(\cdot)$ is defined as:

$$\begin{aligned} B_j(D;W) &= \frac{d_j w_j}{\sum_{j\in J} d_j w_j} B \\ &= \frac{\sum_{j\in J} \delta_i(z_j) w_i}{\sum_{i\in I} \sum_{j\in J} \delta_j(z_i) w_i} B \end{aligned} \tag{A.1}$$

where J is the set of all measures, I is the set of all attributed patients, and B is the maximum bonus amount (note $B = \sum_J B_j$). The variable d_j is defined as the sum over all attributed patients of an indicator function that determines whether a patient's attributes, z_i, make the patient applicable for the measure. For example, $\delta_j(z_i)$ for the breast cancer screening measure is one for patients who are female, between the ages of 52 and 65 and who have not had two mastectomies or a bilateral mastectomy. Finally, w_j is the HMSA measure specific weight. As an example, the HMSA weight for diabetic nephropathy screening is two times the diabetic LDL screening weight and four times the preventative breast cancer screening weight. The list of weights by measure and year for the commercial line of business are described in Table 14.

A.4. Data Construction

The reconstruction of these measures is nontrivial and often imperfect. First, a number of measures require more than one year of claims data. This implies the construction of measures in earlier time periods are not as complete as the later years. For this reason, I restrict my analysis to 2012 through 2015. Furthermore, it is not possible for some of these measures to ever be fully accurate. For example, one acceptable form of screening for colorectal cancer should occur once a decade. The five year claim window simply cannot identify all acceptable screenings. Second, the quality measure definitions change slightly year to year for two measures. For example, Breast Cancer Screening required a mammogram every 24 months for women between the ages of 42 and 69 in 2011. In 2014 this changed to requiring a mammogram every 27 months for a slightly smaller group of women, women between the ages of 52 and 74. For consistency, I chose the narrowest definition over time for the measures so the same types of patients are included each year (see Appendix Table 15 for full measure descriptions and any changes over time).

After applying the HEDIS logic to the 2011 through 2015 claims data, I am able to match the final quarter quality measure rates to HMSA's internally calculated rates. Additionally, I generate bonus payment based on the claims derived r_t, r_{t-1},and D for the six measures. In order to calculate the bonus payment, I must rely on HMSA's internally calculated d_i's for the measures I do not calculate. Tables 21 and 22 describe the correlation between the claim and HMSA generated $r_{i,t}$ and $b_{i,t}$ by year and measure for provider-measure pairs that are above the first quartile of d_i. I do this to decrease measurement error as one would expect estimates for $r_{i,t}$ and $b_{i,t}$ to vary most for providers with a small panel size. Note that the correlation for Diabetic HbA1c Screening is missing in 2014 and 2015 because it was no longer a P4V measure. Additionally, Diabetic LDL Screening is missing from all years. The lab data was not as complete as expected so I have not been able to replicate this measure. For now, this remains for completeness.

Table 21: Correlation between estimated and HMSA generated r_t

(denominator above Q1)

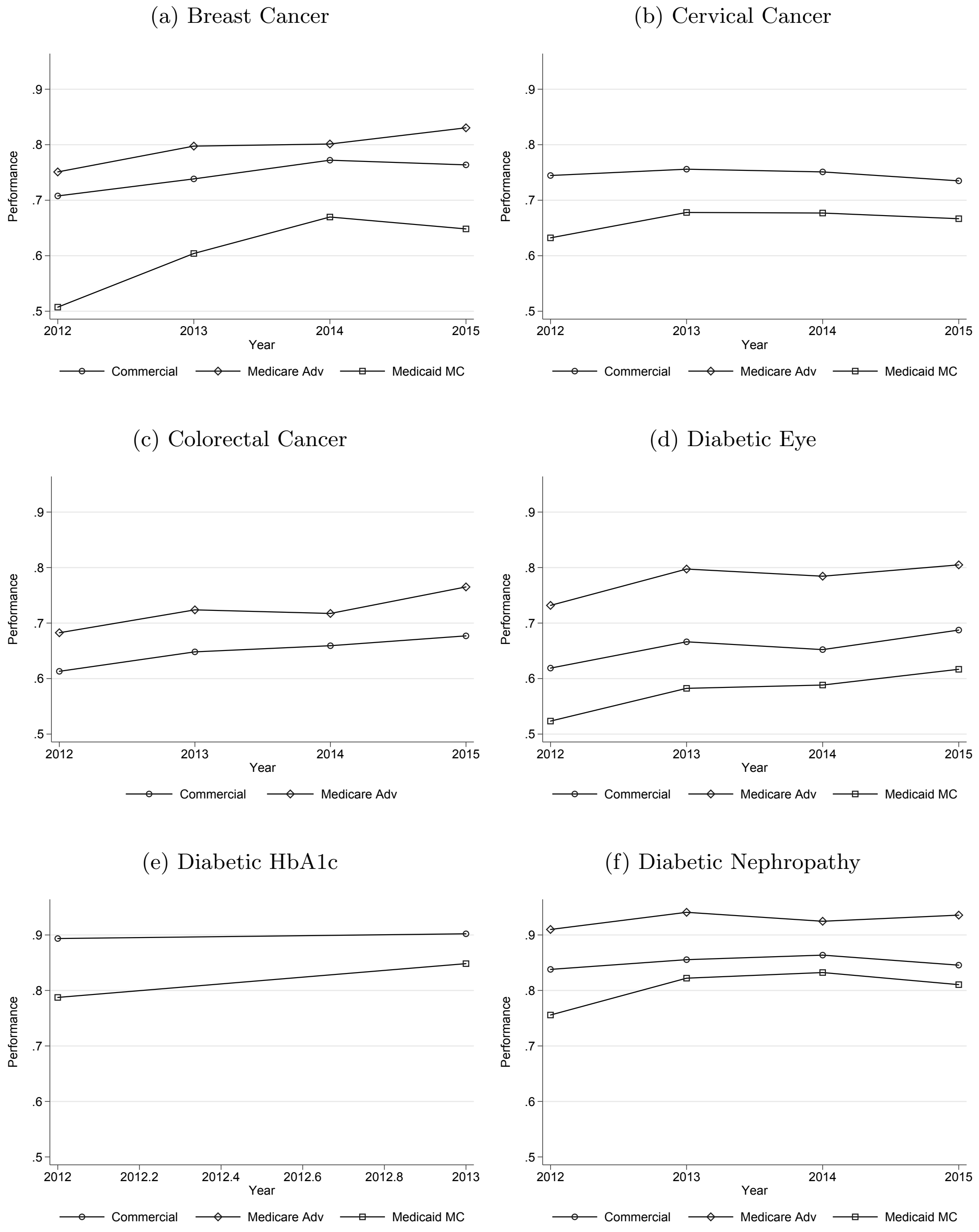
	2011	2012	2013	2014	2015
Diabetes |Eye Exam	0.788	0.836	0.844	0.848	0.825
Diabetes |HbA1C Testing	0.834	0.862	0.868		
Diabetes |Nephropathy	0.766	0.807	0.792	0.822	0.649
Preventive Screening |Breast Cancer	0.843	0.916	0.928	0.983	0.981
Preventive Screening |Cervical Cancer	0.719	0.914	0.955	0.923	0.919
Preventive Screening |Colorectal Cancer	0.567	0.592	0.672	0.608	0.637

Table 22: Correlation between HMSA generated and estimated b

(denominator above Q1)

	2011	2012	2013	2014	2015
Diabetes |Eye Exam	0.784	0.885	0.887	0.858	0.844
Diabetes |HbA1C Testing	0.789	0.858	0.776		
Diabetes |Nephropathy	0.688	0.908	0.883	0.872	0.640
Preventive Screening |Breast Cancer	0.700	0.905	0.912	0.963	0.955
Preventive Screening |Cervical Cancer	0.138	0.736	0.891	0.852	0.817
Preventive Screening |Colorectal Cancer	0.206	0.356	0.436	0.226	0.202

Figure 17: Performance over time for six measures



(a) Breast Cancer

(b) Cervical Cancer

(c) Colorectal Cancer

(d) Diabetic Eye

(e) Diabetic HbA1c

(f) Diabetic Nephropathy

Notes: Note Colorectal Cancer screening does not include Medicare Managed Care product.

A.5. Instrument balance tables

While it is not possible to directly test whether the instruments satisfy the exclusion restriction, one can examine whether the instruments effectively randomize observations along observable dimensions. Tables 23 and 24 describe physician contract and patient risk panel characteristics across quartiles of the instruments. I generated instrument quartiles for each year and measure quartile bin. Column one through four are mean values for the various characteristics within the quartile and column five tests the difference between the first and fourth quartile values.

In Table 23, the average quartile values for β ranges from a 1.2 to 8.7 percent gain of "banked" patients. Only seven of the 29 comorbidity indicators are statistically different between the first and fourth quartiles and all of these differences are under one percentage point.This demonstrates a β is relatively well balanced across patient panel characteristics.

In Table 24, values for α range from 0.3 to 3.3 percent of new patients. Unlike the balance for β values, the majority of comorbidity characteristics are statistically different across top and bottom quartiles. Generally, providers with a smaller portion of new patients have more comorbid conditions suggesting that younger patients do not randomly age into all provider panels, rather panels have different age distributions – some providers see younger patients on average. Nonetheless, the majority of the comorbid differences are less than 1 percentage point. With an average panel size of 1,800 patients, these panels differ by fewer than 18 patients on average.

These balance tables help characterize the local average treatment effect (LATE). The patient panel risk characteristics are relatively even across the four β quartiles, but not across the four α quartiles. In particular, observations with larger α values have a healthier population. This could be interpreted as the LATE representing an upper bound as the α generates a shock in panels where it is likely easier to have a patient meet a measure.

Table 23: Comorbidity balance for β (pct)

	1st Quartile	2nd Quartile	3rd Quartile	4th Quartile	1st - 4th Diff.
β (pct)	0.012	0.014	0.054	0.087	-0.075***
Comorbidities:					
AIDS	0.001	0.001	0.001	0.001	0.000
Alcohol abuse	0.008	0.009	0.009	0.009	-0.000*
Deficiency Anemias	0.105	0.101	0.103	0.097	0.008***
Rheumatoid arthritis/collagen vas	0.018	0.018	0.018	0.018	-0.001*
Chronic blood loss anemia	0.008	0.009	0.009	0.008	-0.000
Congestive heart failure	0.018	0.017	0.018	0.018	-0.000
Chronic pulmonary disease	0.102	0.103	0.105	0.103	-0.001
Coagulopthy	0.012	0.012	0.013	0.012	-0.000
Depression	0.030	0.032	0.032	0.032	-0.002***
Diabetes w/o chronic complications	0.147	0.149	0.153	0.148	-0.000
Diabetes w/ chronic complications	0.032	0.031	0.033	0.031	0.001*
Drug abuse	0.009	0.011	0.009	0.010	-0.001*
Hypertension	0.365	0.364	0.378	0.365	0.001
Hypothyroidism	0.083	0.083	0.086	0.082	0.001
Liver disease	0.036	0.036	0.037	0.033	0.003***
Lymphoma	0.002	0.002	0.003	0.002	-0.000***
Fluid and electrolyte disorders	0.049	0.048	0.048	0.046	0.003**
Metastatic cancer	0.004	0.004	0.004	0.004	-0.000
Other neurological disorders	0.024	0.024	0.025	0.024	0.000
Obesity	0.092	0.087	0.099	0.098	-0.007***
Paralysis	0.005	0.005	0.005	0.005	-0.000
Peripheral vascular disease	0.031	0.029	0.031	0.030	0.000
Psychoses	0.014	0.014	0.014	0.014	-0.001*
Pulmonary circulation disease	0.005	0.005	0.005	0.005	-0.000
Renal failure	0.048	0.044	0.048	0.046	0.002*
Solid tumor w/out metastasis	0.029	0.029	0.031	0.030	-0.001*
Peptic ulcer Disease x bleeding	0.001	0.001	0.001	0.001	-0.000*
Weight loss	0.020	0.019	0.020	0.020	0.000
Valvular disease	0.032	0.033	0.036	0.034	-0.002**

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Table reports the mean proportion of patients in a physician's panel with a specific comorbid condition by β quartile. Instruments quartiles are defined for each measure-year-insurer. All physician-measure-insurer-quarter observations are included that have an average relevant panel size above 10. Additionally the mean β value is included for each quartile at the top of the table. The final column presents the difference between the 1st and 4th quartile and indicates whether that difference is significant.

Table 24: Comorbidity balance for α (pct)

	1st Quartile	2nd Quartile	3rd Quartile	4th Quartile	1st - 4th Diff.
α (pct)	0.003	0.002	0.019	0.033	-0.029***
Comorbidities:					
AIDS	0.001	0.001	0.001	0.001	0.000
Alcohol abuse	0.010	0.009	0.009	0.010	0.000
Deficiency Anemias	0.103	0.103	0.094	0.090	0.014***
Rheumatoid arthritis/collagen vas	0.019	0.018	0.017	0.015	0.004***
Chronic blood loss anemia	0.009	0.009	0.009	0.008	0.000
Congestive heart failure	0.018	0.018	0.016	0.016	0.002***
Chronic pulmonary disease	0.104	0.104	0.101	0.099	0.005*
Coagulopthy	0.013	0.012	0.011	0.010	0.003***
Depression	0.033	0.032	0.031	0.032	0.001
Diabetes w/o chronic complications	0.150	0.153	0.136	0.130	0.021***
Diabetes w/ chronic complications	0.034	0.032	0.029	0.027	0.007***
Drug abuse	0.010	0.010	0.010	0.012	-0.002***
Hypertension	0.365	0.374	0.342	0.321	0.044***
Hypothyroidism	0.087	0.084	0.079	0.074	0.014***
Liver disease	0.034	0.037	0.033	0.029	0.005***
Lymphoma	0.002	0.002	0.002	0.002	0.000***
Fluid and electrolyte disorders	0.052	0.049	0.042	0.042	0.010***
Metastatic cancer	0.005	0.004	0.004	0.004	0.001***
Other neurological disorders	0.025	0.024	0.021	0.020	0.005***
Obesity	0.097	0.090	0.087	0.093	0.004
Paralysis	0.005	0.005	0.004	0.004	0.001***
Peripheral vascular disease	0.029	0.031	0.026	0.025	0.004***
Psychoses	0.015	0.014	0.014	0.015	0.000
Pulmonary circulation disease	0.005	0.005	0.004	0.004	0.001***
Renal failure	0.045	0.047	0.041	0.038	0.007***
Solid tumor w/out metastasis	0.032	0.030	0.026	0.023	0.009***
Peptic ulcer Disease x bleeding	0.001	0.001	0.001	0.001	0.000
Weight loss	0.020	0.020	0.018	0.016	0.004***
Valvular disease	0.034	0.034	0.030	0.028	0.006***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Table reports the mean proportion of patients in a physician's panel with a specific comorbid condition by α quartile. Instruments quartiles are defined for each measure-year-insurer. All physician-measure-insurer-quarter observations are included that have an average relevant panel size above 10. Additionally the mean α value is included for each quartile at the top of the table. The final column presents the difference between the 1st and 4th quartile and indicates whether that difference is significant.

A.6. Programmatic change approach additional tables

Table 25: Effect of threshold shift on performance, breast cancer measure case study

	Quarterly Performance				
		Matched			
				High r	
	All		10 - 90 pctl		10 - 90 pctl
γ_{2013Q1}	-0.00284	-0.000237	-0.00180	0.00473	0.00812
	(0.00218)	(0.00713)	(0.00752)	(0.00655)	(0.00678)
γ_{2013Q2}	-0.00362	0.00363	0.00189	0.00417	0.00364
	(0.00238)	(0.00688)	(0.00765)	(0.00707)	(0.00813)
γ_{2013Q3}	0.00160	0.000990	-0.00110	0.00902	0.00734
	(0.00254)	(0.00655)	(0.00686)	(0.00634)	(0.00707)
γ_{2013Q4}	-0.00498*	-0.0162	-0.0217*	-0.00607	-0.00727
	(0.00231)	(0.00940)	(0.0100)	(0.00556)	(0.00661)
γ_{2014Q1}	0.00414	-0.00196	-0.00382	0.00584	0.00765
	(0.00215)	(0.00838)	(0.00873)	(0.00643)	(0.00693)
γ_{2014Q2}	0.00125	-0.00626	-0.00732	-0.000335	0.000792
	(0.00212)	(0.00819)	(0.00841)	(0.00629)	(0.00685)
γ_{2014Q3}	0.00415	0.00592	0.00624	0.00535	0.00421
	(0.00220)	(0.00748)	(0.00729)	(0.00552)	(0.00645)
γ_{2015Q1}	0.00563**	0.0223***	0.0198***	0.0128*	0.0146*
	(0.00218)	(0.00566)	(0.00590)	(0.00553)	(0.00665)
γ_{2015Q2}	0.00221	0.0129	0.0148*	0.00418	0.00319
	(0.00221)	(0.00673)	(0.00721)	(0.00667)	(0.00820)
γ_{2015Q3}	0.00518*	0.0119	0.00652	0.00822	0.00885
	(0.00244)	(0.00700)	(0.00746)	(0.00664)	(0.00775)
γ_{2015Q4}	0.00377	0.00282	-0.000666	0.0137*	0.0125*
	(0.00232)	(0.00793)	(0.00841)	(0.00553)	(0.00599)
Observations	39291	14131	10511	11130	8436
R^2	0.050	0.322	0.374	0.151	0.172
E[p]	0.026	0.049	0.053	0.038	0.039

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Table reports the γ_t coefficients from estimating Equation 2.2. These coefficients are relative to the quarter prior to the threshold increase, 2014Q4. Column 1 estimation includes all observations and columns 2 through 5 estimations include physician-breast cancer measure pairs and their matched controls (see text for details). Further, columns 3 and 5 estimations include physician-breast cancer observations close to thresholds and their matched controls (see text for details). Columns 4 and 5 include only physician-breast cancer measure observations where the average potential bonus is in the top two tertiles and their matched controls (see text for details). All regressions include quarter fixed effects and time varying physician panel risk controls. In addition to sample restrictions described above, the physician-measure level observations had to represent at least 10 patients. Standard errors are clustered at the physician level.

Table 26: Effect of bonus size shift on performance, diabetic nephropathy measure case study

	Quarterly Performance				
		Matched			
				High r	
	All		10 - 90 pctl		10 - 90 pctl
γ_{2013Q1}	-0.00350	-0.0433	-0.0488*	-0.0101	-0.0157
	(0.00264)	(0.0237)	(0.0237)	(0.0189)	(0.0202)
γ_{2013Q2}	-0.00138	-0.0212	-0.0214	-0.0219	-0.0228
	(0.00259)	(0.0125)	(0.0133)	(0.0235)	(0.0230)
γ_{2013Q4}	0.0113***	0.0105	0.0153	-0.00346	-0.00396
	(0.00280)	(0.0134)	(0.0144)	(0.0117)	(0.0131)
γ_{2014Q1}	0.00208	0.0112	0.00212	-0.0243	-0.0379*
	(0.00265)	(0.0116)	(0.0117)	(0.0146)	(0.0149)
γ_{2014Q2}	0.00142	-0.00334	-0.0116	-0.0202	-0.0265*
	(0.00264)	(0.0181)	(0.0184)	(0.0114)	(0.0117)
γ_{2014Q3}	0.00274	0.00343	-0.00299	-0.0243	-0.0315
	(0.00262)	(0.0163)	(0.0155)	(0.0190)	(0.0188)
γ_{2014Q4}	0.0150***	0.0143	0.0105	-0.0179	-0.0227
	(0.00285)	(0.0184)	(0.0187)	(0.0197)	(0.0212)
γ_{2015Q1}	-0.000167	0.00278	-0.00143	-0.00130	-0.00674
	(0.00250)	(0.0147)	(0.0140)	(0.0152)	(0.0148)
γ_{2015Q2}	-0.000890	0.0158	0.00694	0.00309	-0.00332
	(0.00263)	(0.0121)	(0.0117)	(0.0104)	(0.0111)
γ_{2015Q3}	0.000627	-0.000555	-0.00110	-0.0102	-0.00918
	(0.00253)	(0.0124)	(0.0128)	(0.0120)	(0.0130)
γ_{2015Q4}	0.00862**	0.0217	0.0203	-0.00812	-0.0106
	(0.00267)	(0.0144)	(0.0149)	(0.0192)	(0.0198)
Observations	39291	6817	2907	5069	2137
R^2	0.021	0.155	0.261	0.132	0.263
E[p]	0.048	0.055	0.063	0.052	0.061

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Table reports the γ_t coefficients from estimating Equation 2.3. These coefficients are relative to the quarter prior to the threshold increase, 2013Q3. Column 1 estimation includes all observations and columns 2 through 5 estimations include physician-diabetic nephropathy measure pairs and their matched controls (see text for details). Further, columns 3 and 5 estimations include physician-diabetic nephropathy observations close to thresholds and their matched controls (see text for details). Columns 4 and 5 include only physician-diabetic nephropathy measure observations where the average potential bonus is in the top two tertiles and their matched controls (see text for details). All regressions include quarter fixed effects and time varying physician panel risk controls. In addition to sample restrictions described above, the physician-measure level observations had to represent at least 10 patients. Standard errors are clustered at the physician level.

A.7. Mechanisms additional tables

Table 27: Effect of threshold shift on quarterly performance by high and low type, difference-in-difference breast cancer measure case study

	Quarterly Performance			
			High r	
	$E[\hat{r}_{2014}] \geq \tau_{50,2014}$	$E[\hat{r}_{2014}] < \tau_{50,2014}$	$E[\hat{r}_{2014}] \geq \tau_{50,2014}$	$E[\hat{r}_{2014}] < \tau_{50,2014}$
γ_{2013Q1}	0.00946	-0.0123	0.0124	0.00184
	(0.0112)	(0.00739)	(0.0118)	(0.00695)
γ_{2013Q2}	0.0121	-0.00667	-0.00209	0.00564
	(0.0106)	(0.00783)	(0.0143)	(0.00766)
γ_{2013Q3}	0.00460	-0.00581	0.0110	0.00338
	(0.0103)	(0.00752)	(0.0107)	(0.00801)
γ_{2013Q4}	-0.0331*	-0.0103	-0.0100	-0.00709
	(0.0146)	(0.00783)	(0.01000)	(0.00744)
γ_{2014Q1}	-0.0203	0.0121	0.00197	0.0100
	(0.0131)	(0.00704)	(0.0110)	(0.00784)
γ_{2014Q2}	-0.0163	0.000708	-0.00337	0.000653
	(0.0111)	(0.00781)	(0.0112)	(0.00714)
γ_{2014Q3}	0.0140	-0.00126	0.00271	0.00445
	(0.0121)	(0.00692)	(0.0111)	(0.00715)
γ_{2015Q1}	0.0174*	0.0221***	0.0123	0.0159*
	(0.00876)	(0.00601)	(0.00935)	(0.00757)
γ_{2015Q2}	0.0168	0.0128	0.00369	0.00108
	(0.00998)	(0.00744)	(0.0123)	(0.00819)
γ_{2015Q3}	0.0102	0.00263	0.0156	0.00312
	(0.00977)	(0.00815)	(0.0135)	(0.00737
γ_{2015Q4}	-0.00109	-0.000630	0.0144	0.0106
	(0.0132)	(0.00726)	(0.00884)	(0.00690)
Observations	4495	7101	4285	5448
R^2	0.432	0.342	0.213	0.155
E[p]	0.049	0.062	0.038	0.044

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Table reports the γ_t coefficients from estimating Equation 2.2. These coefficients are relative to the quarter prior to the threshold increase, 2014Q4. The specifications in columns 1 and 2 include physician-breast cancer measure pairs close to thresholds and their matched control. The specifications in columns 3 and 4 include only physician-breast cancer observations where the average potential bonus is in the top two tertiles and their matched controls. Further the specifications include data subsetted into breast cancer-physician observations expected to be in higher and lower parts of the performance pay structure (see text for details). All regressions include quarter fixed effects and time varying physician panel risk controls. In addition to sample restrictions described above, the physician-measure level observations had to represent at least 10 patients. Standard errors are clustered at the physician level.

Table 28: Effect of bonus size decrease on quarterly performance by high and low type, difference-in-difference diabetic nephropathy measure case study

	Quarterly Performance			
			High r	
	$E[\hat{r}_{2014}] \geq \tau_{50,2014}$	$E[\hat{r}_{2014}] < \tau_{50,2014}$	$E[\hat{r}_{2014}] \geq \tau_{50,2014}$	$E[\hat{r}_{2014}] < \tau_{50,2014}$
γ_{2013Q1}	-0.0510	-0.0463*	-0.0156	-0.00685
	(0.0262)	(0.0223)	(0.0228)	(0.0252)
γ_{2013Q2}	-0.0248	-0.0143	-0.0264	-0.0146
	(0.0148)	(0.0192)	(0.0261)	(0.0275)
γ_{2013Q4}	0.0155	0.0170	-0.00625	0.0261
	(0.0160)	(0.0168)	(0.0143)	(0.0212)
γ_{2014Q1}	0.00632	-0.00549	-0.0336*	-0.0517*
	(0.0126)	(0.0187)	(0.0158)	(0.0232)
γ_{2014Q2}	-0.00769	-0.0156	-0.0243*	-0.0275
	(0.0210)	(0.0189)	(0.0120)	(0.0252)
γ_{2014Q3}	0.00420	-0.0201	-0.0311	-0.0336
	(0.0158)	(0.0185)	(0.0208)	(0.0242)
γ_{2014Q4}	0.0115	0.0131	-0.0261	-0.00695
	(0.0212)	(0.0182)	(0.0233)	(0.0285)
γ_{2015Q1}	0.00812	-0.0303	-0.000294	-0.0321
	(0.0139)	(0.0185)	(0.0150)	(0.0229)
γ_{2015Q2}	0.0114	-0.00450	-0.000656	-0.0260
	(0.0128)	(0.0159)	(0.0124)	(0.0227)
γ_{2015Q3}	0.000929	-0.00703	-0.00994	-0.0103
	(0.0140)	(0.0164)	(0.0149)	(0.0220)
γ_{2015Q4}	0.0233	0.0139	-0.00801	-0.0197
	(0.0162)	(0.0183)	(0.0219)	(0.0229)
Observations	2368	779	1836	563
R^2	0.271	0.295	0.275	0.337
E[p]	0.062	0.055	0.061	0.051

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Table reports the γ_t coefficients from estimating Equation 2.3. These coefficients are relative to the quarter prior to the threshold increase, 2013Q3. The specifications in columns 1 and 2 include physician-diabetic nephropathy measure pairs close to thresholds and their matched control. The specifications in columns 3 and 4 include only physician-diabetic nephropathy observations where the average potential bonus is in the top two tertiles and their matched controls. Further the specifications include data subsetted into diabetic nephropathy-physician observations expected to be in higher and lower parts of the performance pay structure (see text for details). All regressions include quarter fixed effects and time varying physician panel risk controls. In addition to sample restrictions described above, the physician-measure level observations had to represent at least 10 patients. Standard errors are clustered at the physician level.

Table 29: Effect of distance and marginal bonus on performance by month, physician fixed effects

	Monthly performance					
		τ_{90}			τ_{50}	
	Month 1	Month 2	Month 3	Month 1	Month 2	Month 3
Distance to τ (pct)	0.0241***	0.0255***	0.0267***	0.0241***	0.0250***	0.0263***
	(0.00186)	(0.00200)	(0.00161)	(0.00219)	(0.00239)	(0.00199)
Ln marginal bonus to τ	0.00203***	0.00159***	0.00185***	-0.0000329	-0.000789	0.000321
	(0.000297)	(0.000267)	(0.000281)	(0.000479)	(0.000436)	(0.000438)
Observations	37780	37780	37780	26745	26745	26745
R^2	0.212	0.186	0.177	0.225	0.193	0.190

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Table reports coefficients from estimating Equation 3.1 with physician FE regressions of monthly performance on distance to τ_{jt} (d_{ijt}), logged marginal bonus to τ_{jt} (m_{ijt}), and all controls. The τ_{jt} columns headings represent specifications that use distance and step size constructed with different τ_{jt}'s: 90th and 50th percentiles. Note as τ_{jt} decreases, the sample size decreases as fewer observations have positive step size values. Further, specifications in each column include one month each each quarter (1st, 2nd or 3rd). Regressions include all relevant physician-measure month observations (measures include preventative measures: breast cancer, colorectal cancer and cervical cancer screening; and diabetic measures: HbA1c screening, annual eye exam, and nephropahy screening). Controls include one month lagged quarterly performance, logged number of patients relevant for a measure, line of business, and measure fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Finally, regressions include only physician-measure-month observations that represent more than 10 patients.

Table 30: Effect of distance and marginal bonus on performance by month, physician-measure fixed effects

	Monthly performance					
		τ_{90}			τ_{50}	
	Month 1	Month 2	Month 3	Month 1	Month 2	Month 3
Distance to τ (pct)	0.0454***	0.0461***	0.0452***	0.0414***	0.0398***	0.0402***
	(0.00261)	(0.00281)	(0.00228)	(0.00291)	(0.00307)	(0.00264)
Ln marginal bonus to τ	0.00162***	0.00117***	0.00143***	0.000152	-0.000864	0.000199
	(0.000327)	(0.000286)	(0.000326)	(0.000516)	(0.000453)	(0.000499)
Observations	37780	37780	37780	26745	26745	26745
R^2	0.288	0.264	0.252	0.333	0.299	0.294

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Table reports coefficients from estimating Equation 3.1 with physician-measure FE regressions of monthly performance on distance to τ_{jt} (d_{ijt}), logged marginal bonus to τ_{jt} (m_{ijt}), and all controls. The τ_{jt} columns headings represent specifications that use distance and step size constructed with different τ_{jt}'s: 90th and 50th percentiles. Note as τ_{jt} decreases, the sample size decreases as fewer observations have positive step size values. Further, specifications in each column include one month each each quarter (1st, 2nd or 3rd). Regressions include all relevant physician-measure month observations (measures include preventative measures: breast cancer, colorectal cancer and cervical cancer screening; and diabetic measures: HbA1c screening, annual eye exam, and nephropahy screening). Controls include one month lagged quarterly performance, logged number of patients relevant for a measure, line of business, and measure fixed effects. All regressions also include a set of Elixhauser comborbidity scores representing the percent of a physician's panel in quarter t with the comorbidity. Finally, regressions include only physician-measure-month observations that represent more than 10 patients.

BIBLIOGRAPHY

J. Abaluck, J. Gruber, and A. Swanson. Prescription Drug Use under Medicare Part D: A Linear Model of Nonlinear Budget Sets. Working Paper 20976, National Bureau of Economic Research, Feb. 2015.

K. Baicker, S. Mullainathan, and J. Schwartzstein. Behavioral Hazard in Health Insurance. *The Quarterly Journal of Economics*, 130(4):1623–1667, Nov. 2015.

G. P. Baker. Incentive Contracts and Performance Measurement. *Journal of Political Economy*, 100(3):598–614, June 1992.

N. D. Beaulieu and D. R. Horrigan. Putting Smart Money to Work for Quality Improvement. *Health Services Research*, 40(5p1):1318–1334, Oct. 2005.

Z. C. Brot-Goldberg, A. Chandra, B. R. Handel, and J. T. Kolstad. What does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics. *The Quarterly Journal of Economics*, 132(3):1261–1318, Aug. 2017.

K. M. Brown. The link between pensions and retirement timing: Lessons from California teachers. *Journal of Public Economics*, 98:1–14, Feb. 2013.

L. R. Burns and M. V. Pauly. Transformation of the Health Care Industry: Curb Your Enthusiasm? *The Milbank Quarterly*, 96(1):57–109, Mar. 2018.

Centers for Medicare and Medicaid Services (CMS). Historical NHE Tables, 2015.

J. Y. Chen, N. Kang, D. T. Juarez, K. A. Hodges, and R. S. Chung. Impact of a Pay-for-Performance Program on Low Performing Physicians. *Journal for Healthcare Quality*, 32 (1):13–22, Jan. 2010.

J. Y. Chen, H. Tian, D. T. Juarez, I. Yermilov, R. S. Braithwaite, K. A. Hodges, A. Legorreta, and R. S. Chung. Does Pay for Performance Improve Cardiovascular Care in a Real-World Setting? *American Journal of Medical Quality*, 26(5):340–348, Sept. 2011.

R. Chetty, A. Looney, and K. Kroft. Salience and Taxation: Theory and Evidence. *American Economic Review*, 99(4):1145–1177, Sept. 2009.

R. Chetty, J. N. Friedman, T. Olsen, and L. Pistaferri. Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records. *The Quarterly Journal of Economics*, 126(2):749–804, May 2011.

R. Chetty, J. N. Friedman, and E. Saez. Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of the EITC on Earnings. *American Economic Review*, 103(7):2683–2721, Dec. 2013.

J. Clemens and J. D. Gottlieb. Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health? *American Economic Review*, 104(4):1320–1349, Apr. 2014.

K. Coleman, K. L. Reiter, and D. Fulwiler. The impact of pay-for-performance on diabetes care in a large network of community health centers. *Journal of Health Care for the Poor and Underserved*, 18(4):966–983, Nov. 2007.

R. F. Coulam and G. L. Gaumer. Medicare's prospective payment system: A critical appraisal. *Health Care Financing Review*, 1991(Suppl):45, Mar. 1992.

R. Diamond and P. Persson. The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests. Working Paper 22207, National Bureau of Economic Research, Apr. 2016.

F. Eijkenaar, M. Emmert, M. Scheppach, and O. Schffski. Effects of pay for performance in health care: A systematic review of systematic reviews. *Health Policy*, 110(23):115–130, May 2013.

L. Einav, A. Finkelstein, and P. Schrimpf. The Response of Drug Expenditure to Nonlinear Contract Design: Evidence from Medicare Part D. *The Quarterly Journal of Economics*, 130(2):841–899, May 2015.

L. Einav, A. Finkelstein, and N. Mahoney. Provider Incentives and Healthcare Costs: Evidence from Long-Term Care Hospitals. Working Paper 23100, National Bureau of Economic Research, Jan. 2017a.

L. Einav, A. Finkelstein, and P. Schrimpf. Bunching at the kink: Implications for spending responses to health insurance contracts. *Journal of Public Economics*, 146:27–40, Feb. 2017b.

R. P. Ellis and T. G. McGuire. Supply-Side and Demand-Side Cost Sharing in Health Care. *Journal of Economic Perspectives*, 7(4):135–151, Dec. 1993.

S. Felt-Lisk, G. Gimm, and S. Peterson. Making Pay-For-Performance Work In Medicaid. *Health Affairs*, 26(4):w516–w527, July 2007.

M. Gaynor and P. Gertler. Moral Hazard and Risk Spreading in Partnerships. *RAND Journal of Economics*, 26(4):591–613, 1995.

A. S. Gilmore, Y. Zhao, N. Kang, K. L. Ryskina, A. P. Legorreta, D. A. Taira, and R. S. Chung. Patient Outcomes and Evidence-Based Medicine in a Preferred Provider Organization Setting: A Six-Year Evaluation of a Physician Pay-for-Performance Program. *Health Services Research*, 42(6 Pt 1):2140–2159, Dec. 2007.

J. Greene, J. H. Hibbard, and V. Overton. Large Performance Incentives Had The Greatest Impact On Providers Whose Quality Metrics Were Lowest At Baseline. *Health Affairs*, 34(4):673–680, Apr. 2015.

J. Gruber. The effect of competitive pressure on charity: Hospital responses to price shopping in California. *Journal of Health Economics*, 13(2):183–211, July 1994.

J. Gruber, J. Kim, and D. Mayzlin. Physician fees and procedure intensity: the case of cesarean delivery. *Journal of Health Economics*, 18(4):473–490, Aug. 1999.

B. R. Handel and J. T. Kolstad. Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare. *American Economic Review*, 105(8):2449–2500, Aug. 2015. ISSN 0002-8282. doi: 10.1257/aer.20131126. URL `https://www.aeaweb.org/articles?id=10.1257/aer.20131126`.

P. Harasztosi and A. Lindner. Who pays for the minimum wage? 2015.

B. Holmstrom and P. Milgrom. Multitask PrincipalAgent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization*, 7 (special issue):24–52, Jan. 1991.

M. Jacobson, C. C. Earle, M. Price, and J. P. Newhouse. How Medicares Payment Cuts For Cancer Chemotherapy Drugs Changed Patterns Of Treatment. *Health Affairs*, page 10.1377/hlthaff.2009.0563, June 2010.

I. Larkin. The Cost of High-Powered Incentives: Employee Gaming in Enterprise Software Sales. *Journal of Labor Economics*, 32(2):199–227, Apr. 2014.

C. H. Lemak, T. A. Nahra, G. R. Cohen, N. D. Erb, M. L. Paustian, D. Share, and R. A. Hirth. Michigans Fee-For-Value Physician Incentive Program Reduces Spending And Improves Quality In Primary Care. *Health Affairs*, 34(4):645–652, Apr. 2015.

K. N. Lohr, R. H. Brook, C. J. Kamberg, G. A. Goldberg, A. Leibowitz, J. Keesey, D. Reboussin, and J. P. Newhouse. Use of Medical Care in the Rand Health Insurance Experiment: Diagnosis- and Service-Specific Analyses in a Randomized Controlled Trial. *Medical Care*, 24(9):S1–S87, 1986.

U. Malmendier and S. Nagel. Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?*. *The Quarterly Journal of Economics*, 126(1):373–416, Feb. 2011.

M. McClellan, R. Richards, and M. Japinga. Evidence on Payment Reform: Where are the Gaps?, 2017. URL `https://www.healthaffairs.org/do/10.1377/hblog20170425.059789/full/`.

T. G. McGuire and M. V. Pauly. Physician response to fee changes with multiple payers. *Journal of Health Economics*, 10(4):385–410, Jan. 1991.

J. A. Mirrlees. An Exploration in the Theory of Optimum Income Taxation. *The Review of Economic Studies*, 38(2):175–208, Apr. 1971.

P. Oyer. Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality. *The Quarterly Journal of Economics*, 113(1):149–185, 1998.

M. B. Rosenthal and R. G. Frank. What Is the Empirical Basis for Paying for Quality in Health Care? *Medical Care Research and Review*, 63(2):135–157, Apr. 2006.

M. B. Rosenthal, R. G. Frank, Z. Li, and A. M. Epstein. Early Experience With Pay-for-Performance: From Concept to Practice. *JAMA*, 294(14):1788–1793, Oct. 2005.

M. B. Rosenthal, F. S. de Brantes, A. D. Sinaiko, M. Frankel, R. D. Robbins, and S. Young. Bridges to Excellence–recognizing high-quality care: analysis of physician quality and resource use. *The American Journal of Managed Care*, 14(10):670–677, Oct. 2008.

A. M. Ryan and C. L. Damberg. What can the past of pay-for-performance tell us about the future of Value-Based Purchasing in Medicare? *Healthcare (Amsterdam, Netherlands)*, 1 (1-2):42–49, June 2013.

E. Saez. Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy*, 2(3):180–212, Aug. 2010.

Z. Song, D. G. Safran, B. E. Landon, M. B. Landrum, Y. He, R. E. Mechanic, M. P. Day, and M. E. Chernew. The Alternative Quality Contract, Based On A Global Budget, Lowered Medical Spending And Improved Quality. *Health Affairs*, 31(8):1885–1894, Aug. 2012.

Z. Song, J. Z. Ayanian, J. Wallace, Y. He, T. B. Gibson, and M. E. Chernew. Unintended Consequences of Eliminating Medicare Payments for Consultations. *JAMA Internal Medicine*, 173(1):15–21, Jan. 2013.

Z. Song, J. Wallace, H. T. Neprash, M. R. McKellar, M. E. Chernew, and J. M. McWilliams. Medicare Fee Cuts and Cardiologist-Hospital Integration. *JAMA Internal Medicine*, 175 (7):1229–1231, July 2015.

G. R. Wilensky. Will MACRA Improve Physician Reimbursement? *New England Journal of Medicine*, 378(14):1269–1271, Apr. 2018.

W. C. Yip. Physician response to Medicare fee reductions: changes in the volume of coronary artery bypass graft (CABG) surgeries in the Medicare and private sectors. *Journal of Health Economics*, 17(6):675–699, Dec. 1998.

G. J. Young, J. F. Burgess, and B. White. Pioneering Pay-for-Quality: Lessons from the Rewarding Results Demonstrations. *Health Care Financing Review; Washington*, 29(1): 59–70, 2007a.

G. J. Young, M. Meterko, H. Beckman, E. Baker, B. White, K. M. Sautter, R. Greene, K. Curtin, B. G. Bokhour, D. Berlowitz, and J. F. Burgess. Effects of Paying Physicians Based on their Relative Performance for Quality. *Journal of General Internal Medicine*, 22(6):872–876, June 2007b.