NBER WORKING PAPER SERIES

PHYSICIAN RESPONSE TO PAY-FOR-PERFORMANCE:
EVIDENCE FROM A NATURAL EXPERIMENT

Jinhu Li
Jeremiah Hurley
Philip DeCicca
Gioia Buckley

Working Paper 16909
http://www.nber.org/papers/w16909

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2011

We thank Catherine Deri Armstrong, Linda Dynan, Jasmin Kantarevic, Sisira Sarma, Adam Wagstaff and participants at the 7th World Congress on Health Economics of the International Health Economics Association, the 3rd biennial conference of the American Society of Health Economists, the 44th annual conference of the Canadian Economics Association, the May 2010 annual conference of the Canadian Health Economic Study Group, and the 2010 annual Canadian Association for Health Services and Policy Research conference for helpful comments and advice. This research was funded by the grant from the Ontario Ministry of Health and Long-Term Care to the Centre for Health Economics and Policy Analysis as part of their collaborative research program. We thank Marsha Barnes for initially sponsoring this project. We also thank Ji Li for excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Physician Response to Pay-for-Performance: Evidence from a Natural Experiment
Jinhu Li, Jeremiah Hurley, Philip DeCicca, and Gioia Buckley
NBER Working Paper No. 16909
March 2011
JEL No. I18,J33

ABSTRACT

Explicit financial incentives, especially pay-for-performance (P4P) incentives, have been extensively employed in recent years by health plans and governments in an attempt to improve the quality of health care services. This study exploits a natural experiment in the province of Ontario, Canada to identify empirically the impact of pay-for-performance (P4P) incentives on the provision of targeted primary care services, and whether physicians' responses differ by age, practice size and baseline compliance level. We use an administrative data source which covers the full population of the province of Ontario and nearly all the services provided by practicing primary care physicians in Ontario. With an individual-level data set of physicians, we employ a difference-in-differences approach that controls for both "selection on observables" and "selection on unobservables" that may cause estimation bias in the identification. We also implemented a set of robustness checks to control for confounding from the other contemporary interventions of the primary care reform in Ontario. The results indicate that, while all responses are of modest size, physicians responded to some of the financial incentives but not the others. The differential responses appear related to the cost of responding and the strength of the evidence linking a service with quality. Overall, the results provide a cautionary message regarding the effectiveness of pay-for-performance schemes for increasing quality of care.

Jinhu Li
McMaster University
1280 Main St. West
Hamilton, ON L8S 4M4
lij53@mcmaster.ca

Jeremiah Hurley
McMaster University
1280 Main St. West
Hamilton, ON L8S 4M4
hurley@mcmaster.ca

Philip DeCicca
Department of Economics
422 Kenneth Taylor Hall
McMaster University
Hamilton, ON L8S, 4M4
CANADA
and NBER
decicca@mcmaster.ca

Gioia Buckley
McMaster University
1280 Main St. West
Hamilton, ON L8S 4M4
buckle@mcmaster.ca

1. Introduction

Explicit financial incentives, especially pay-for-performance (P4P) incentives, have been extensively employed and strongly advocated in recent years by health plans and governments in an attempt to improve the quality of health care services. Pay-for-performance is now a concept that is embraced by a lot of policy makers and is deemed as a critical component of health care reforms. A typical P4P program offers financial rewards to health care providers for meeting pre-established targets for the provision of specific health care services. These explicit financial incentives, which are used within different compensation schemes, aim to motivate health care providers to provide high-quality care.

A variety of P4P programs have been established in several countries. In the United States, as of 2005 at least 100 nationwide P4P initiatives had been sponsored by health plans, employer coalitions and the Centers for Medicare and Medicaid Services (CMS) (Baker and Carter 2005). Initially, most of the P4P programs were targeted at primary care physicians affiliated with Health Maintenance Organizations (HMO). Since 2004 there has been significant expansion of P4P programs to specialists and hospitals, which use more sophisticated measures for performance assessment (Rosenthal and Dudley 2007, Baker 2004, Baker and Carter 2005). In the United Kingdom, the British National Health Service (NHS) introduced a pay-for-performance contract for family practitioners in 2004 which linked physician income to performance with respect to 146 quality indicators relating to clinical care for 10 chronic diseases, the organization of care and patient experience (Doran et al. 2006). P4P incentive programs have also been used in Canada, Australia, Haiti and other nations (Frolich et al. 2007).

The rationale for employing P4P incentives to induce desired physician behaviour comes primarily from principal-agent theory and incentive-contract theory. The classic principal-agent and incentive contract theories analyze how pay-for-performance can be used to elicit desired behaviours from individuals in the presence of information asymmetry. The analysis focuses particularly on how the ability to elicit desired behavior is constrained by the noisiness of the performance measures, the extent to which the performance is easily monitored, the ability of agents to handle risk, and the extent to which the desired behavior consists of multiple tasks (Prendergast 1999; Baker 1992; Hart and Holmstrom 1987; Milgrom and Roberts 1992; Stiglitz 1974). The take-away message from these theories is that performance-based contracting can induce agents to improve performance when payment is based on achieving pre-specified performance targets.

In reality though, using P4P programs to motivate health care providers' behaviour is controversial. Advocates believe that P4P can fix many of the long-standing deficiencies in health care system, especially the failure to deliver appropriate and evidence-based care to all patient populations. Years of reforms to general payment mechanisms have had little impact on reducing the deficiencies in health care delivery. This has led to the gradual employment of explicit P4P incentives to link financial gains and losses to quality indicators (Maynard 2008). The belief is that, by making payments at least partly contingent on indicators of high-quality care, P4P programs will induce providers to improve health care quality (Rosenthal and Frank 2006). However, critics argue that P4P programs are not as effective as commonly claimed and often create unintended consequences. Some argue that P4P programs can be very costly because payment used to induce even marginal improvement in quality is often expensive (Christianson et al. 2008; Lewis 2009). Others argue that P4P will induce gaming behaviors by physicians such as strategic coding of patient diagnoses, patient selection and patients-exception reporting (Hutchison 2008; Shen 2003; Richards 2009; Doran et al. 2008; Gravelle et al. 2010). Finally, some P4P programs create unintended consequences such as provider focus on the clinical outcomes subject to incentives to the neglect of other aspects of care (Rosenthal and Frank 2006; Mullen et al. 2010).

Theoretical predictions on physician responses to P4P incentives are ambiguous. Health economics generally models physicians as utility-maximizing service providers who choose their optimal level and mix of services to trade off among income, leisure and other consumption goods (McGuire and Pauly 1991; McGuire 2000). Physician responses to the price increase of the targeted services, generated by individual P4P incentives, are ambiguous because income effect and substitution effect work in opposite directions. Furthermore, there is no consensus about the specific form of physicians' utility function. Besides financial objectives, non-pecuniary factors including medical ethics, professional autonomy and social status, and altruistic concerns about patient outcomes are also argued influence physician utilities (Scott 2001; Eisenberg 1985; Eisenberg 1986). As a result, physicians are less likely to respond to financial incentives when a falling marginal utility of income renders income less attractive in relation to other objectives (McGuire 2000). Moreover, P4P incentives in health care are often embedded within complex compensation systems and provider organizations (Conrad and Christianson 2004; Frolich et al. 2007), where physicians face different incentives from multiple payers and operate in highly regulated settings. The effect of P4P incentives can thus be mitigated by other simultaneous incentives. Therefore, how physicians would respond to P4P incentives remains an empirical issue.

Empirical studies providing good evidence of how performance incentives influence physician delivery of targeted services are scarce. Studies based on Random Controlled Trials (RCTs) have limited generalizability due to the small scale of the experiments. Although the number of observational studies is growing, these empirical studies often suffer from poor study-design. Furthermore, the findings from the existing empirical studies are mixed and inconclusive. Most of them find partial effects of P4P incentives in sense that, physicians respond to some of the incentives but not the others; for the subset of incentives which did improve the performance, the magnitude of the improvement is modest. A few studies find consistent positive effects but others find no effect. We will discuss these studies, and others, in more detail in the following section.

This study exploits a natural experiment in the province of Ontario, Canada to identify empirically the impact of pay-for-performance (P4P) incentives on the provision of targeted primary care services. The P4P scheme rewards family physicians (FPs) and general practitioners (GPs) when they achieve targeted levels of service provision[5]. Primary care reform in Ontario provides a good setting that allows us to employ a difference-in-differences approach to control for potential sources of bias when identifying the effect of P4P incentives on physician behaviour. The policy intervention exposed some, but not all, of the GPs in Ontario to P4P incentives. Therefore, the GPs who were not eligible for the P4P incentives constitute a natural comparison group for our study design. Also, the timing of the P4P implementation allows us to mitigate perfect confounding of other attributes of primary care reform interventions with P4P. The majority of the GPs were exposed to P4P incentives sometime after they participated into the primary care reforms. Using this group of GPs as the treatment group in a difference-in-differences with individual fixed effects method allows us to disentangle the impact of P4P incentives from the effect generated by other policy changes.

We exploit an administrative data source which covers the full population of the province of Ontario and nearly all GPs. The administrative databases include detailed information on services provided that constitutes over 98% of all physician activity. By linking different sources of administrative databases, we can observe the group of physicians who were affected by the incentives and the group of physicians that were not affected by the incentives in both pre- and post- intervention periods. The population-based nature of this data provides us a large sample size, while the rich content of the data allows us to address a variety of potential biases that are caused by "selection on observables" and to partially control for potential bias that are caused by "selection on unobservables".

Furthermore, the universal public insurance and single-payer system in Canada provides an extra advantage for identifying the P4P incentive effects. In multiple-payers settings, such as the U.S., as Robinson notes (Robinson 2001), comprehension and compliance to any payment mechanisms will be undermined when physicians face different incentives from multiple insurers or organizations. Therefore, the estimates of the P4P

[5] For ease of exposition, for the rest of the paper I will refer to both FPs and GPs as "GPs".

incentives from the US studies are expected to be biased towards zero. In Ontario, however, physicians face only a single payer.

This study also examines the heterogeneity of the P4P incentive effects across different physician types and different practice characteristics. We expect that the impact of P4P incentives is heterogeneous because both the benefit of responding to P4P incentives and the cost of responding likely differ across physicians, services and practices. We compare the incentives effects across physician age, across practices that differ in patient population size, and across practices with different baseline levels of service provision.

2. Empirical evidence on physician response to P4P

A large body of empirical studies has examined the effect of financial incentives on physician behavior. There is considerable evidence that physicians respond to the incentives embedded in different payment schemes (McGuire and Pauly 1991; McGuire 2000; Hurley et al. 1990; Yip 1998; Nguyen and Derrick 1997; Hickson et al. 1987; Krasnik et al. 1990; Scott and Shiell 1997). There is less evidence on physician in responses to explicit financial incentives in the form of targeted performance payments intended to guide specific behaviours.

This study focuses on the effect of pay-for-performance incentives on the behaviour of physicians so we focus the review on thirty studies identified by several recent survey papers (Rosenthal and Frank 2006; Christianson et al. 2008; Petersen et al. 2006; Town et al. 2005; Armour et al. 2001) and by our own search of the literature for papers that pertain to physician responses to P4P incentives (See Appendix 1 for the identified empirical studies). Among the thirty studies, eight of them are based on Random Control Trials (RCTs) and twenty-two are based on observational studies.

The RCTs examine the effects of alternative forms of performance incentives, such as bonus, bonus based on capitation payment, bonus with performance feedback, on the provision of targeted services by physicians. In most RCTs, the incentives are mostly targeted on preventive care services, including influenza immunizations, mammograms, Pap smear, colorectal screening and pediatric immunization. The sample sizes are generally small.

The results from the RCTs are mixed. Three studies (Grady et al. 1997; Hillman et al. 1998; Hillman et al. 1999) didn't detect any significant effect of P4P bonus rewards or bonus rewards combined with performance feedback on physician compliance with cancer screening, pediatric immunization and mammography referrals. Two studies (Fairbrother et al. 1999; Fairbrother et al. 2001) found that a bonus or bonus with performance feedback incentives increased documented coverage levels for childhood immunization, but the measured increase was primarily due to better documentation not better immunization practices. A study of bonus payments for smoking cessation clinics (Roski et al. 2003) found a significant improvement in documentation of patient smoking status and in providing advice to quit, but no effect on quitting rates. The other two RCTs showed a significantly positive effect of using bonus payment at the practice or the clinic level: Kouides et al. (1998) showed that a bonus payment for influenza immunization increased rates by 7 percent; Lawrence et al. (2008) found that the clinics with P4P payments had higher level of referral rates on tobacco quitline services than the clinics without payments.

RCTs are often deemed as the "gold standard" to identify the causal effects, but the results from these RCTs often suffer from small sample size problem and cannot easily be generalized or extrapolated. All of these RCTs are based on small scale experiments involving fewer than a hundred physicians or practices. One study (Hillman et al. 1998) involved only 52 physician practices in total. As a result, the effect size might not be statistically identified due to the lack of power. Moreover, the intervention studied by these RCTs can make it impossible to disentangle the pure P4P financial incentives effects from other quality management tools. Among these RCTs, two studies (Hillman et al. 1998; Hillman et al. 1999) bundled the bonus payment with performance feedback regarding compliance level; one study (Grady et al. 1997) bundled financial reward with the provision of education in the form of chart reminder stickers.

The observational studies are mostly based on small to large scale pilot pay-for-quality programs or quality-improvement initiatives adopted by health plans in US, UK and Taiwan. These programs generally covered a broader set of quality indicators than merely preventive care services, such as process and outcome measures for diabetic care, asthma and coronary heart disease and other chronic conditions

Doran et al. (2006) evaluated the effect of the nationwide P4P program introduced by Britain's National Health Service in 2004 for family practitioners. The program linked increases in income to performance with respect to 146 quality indicators covering clinical care for 10 chronic diseases, organization of care, and patient experience. The English family practices attained high levels of achievement meeting the quality indicators, as the median reported achievement was 83.4 percent in the first year of the P4P program (April 2004 through March 2005). But this study is based on a cross-sectional analysis so it only established an association between high level of reported achievement and the P4P contracting, not the real effect of the P4P incentives. As Campbell et al. (2007) noted, because a wide range of initiatives, including limited use of incentive programs, had been introduced in the UK since 1990, the high levels of quality attained after the 2004 contract might just reflect improvements that were already under way.

Campbell et al. (2007; 2009) used a before-after design to examine the effect of 2004 P4P contracting on the quality of care. Both studies measured quality indicators for three chronic conditions --- asthma, coronary heart disease, and type-2 diabetes --- for representative groups of general practitioners. Campbell et al. (2007) measured these quality indictors two times before the P4P contracting (1998 and 2003) and one time after the contracting (2005), and compared the quality score predicted by 1998-2003 trend against the observed quality score in 2005. The results indicate that the introduction of pay for performance was associated with a modest acceleration in improvement for two of these three conditions, diabetes and asthma. Campbell et al. (2009) assessed the same quality indicators at an additional time point of 2007, and extended the previous study by using an interrupted time series analysis. The study found that in 2005 the rate of improvement in quality increased for diabetes care and asthma but remained unchanged for coronary heart disease; by 2007, the rate of improvement for all three conditions had slowed down: as compared with the period before the pay-for-performance scheme was introduced, the improvement rate was unchanged for asthma or diabetes and was reduced for heart disease. Since the P4P contracting is offered to all general practitioners in the U.K., neither study could include a plausible control group against which to compare changes in service provision following the introduction of the incentives. Other studies based on the same pay-for-performance scheme in the U.K. (Millett et al. 2007; Steel et al. 2007;Vaghela et al. 2009) examined the effect of P4P incentives on other quality indicators such as smoking cessation and hypertension outcomes, and found statistically significant increase in these quality indicators after the introduction of this P4P scheme. They suffer from the same problem of identification thus hardly to provide reliable evidence by using only simple before-after analysis.

Evidence of P4P incentives from the U.S. is rapidly growing. Most U.S. studies have been based on small-scale pilot P4P programs adopted by health plans in different states. These studies often suffer from poor study design: some of them only employed simple before-after mean comparison or trend comparison (Levin-Scherz et al. 2006; Young et al. 2007; Cutler et al. 2007; Pearson et al. 2008); others do not provide any comparison group as the counterfactuals (Amundson et al. 2003; Mandel and Kotagal 2007; Chung et al. 2010; Boland et al. 2010; Lester et al. 2010; Coleman et al. 2007). Some of the programs were targeted at health plans or clinics instead of individual physicians, so the lack of individual-level data makes it difficult to draw inference on physician responses to P4P incentives (Felt-Lisk et al. 2007; Gavagan et al. 2010). Furthermore, results are often limited by the small size of these programs. For example, Beaulieu and Horrigan (2005) examined the effect of performance bonus on the improvement of nine measures for diabetic care by using only 21 physicians as the treatment group. So it is difficult to draw reliable inference from this study.

The best evidence to date on the effects of P4P programs are from two observational studies in the U.S. drawn from the P4P initiatives introduced by a large network Health Managed Organization (HMO): PacifiCare Health Plan. The first study (Rosenthal et al. 2005) examined the effect of Quality Incentive Programs (QIP) provided by PacifiCare Health Plan to medical groups in California in 2002 on physician delivery of cervical cancer screening, mammography and haemoglobin A1c test. It used a difference-in-difference design by comparing provider groups in California which were affected by these incentives with provider groups in the Pacific Northwest which were unaffected by the incentives but also contracted with PacifiCare Health Plan. It found that outcomes improved for cervical cancer screening, but did not improve for mammography and the haemoglobin A1c test. The second study (Mullen et al. 2009) built on the first paper and examined the effect of QIP incentives along with another larger P4P program by the Integrated Healthcare Association (IHA). It also concluded that the P4P incentive effects are mixed. In line with the previous study, the analysis found evidence of a positive effect only for cervical cancer screening, but not for mammography, haemoglobin A1c test and asthma medication. Overall, the study concluded that the pay-for-performance scheme resulted in neither a major improvement in quality nor a notable disruption in care (which some hypothesized would be a negative side-effect).

The findings from these empirical studies suggest that the evidence of physician responses to P4P incentives is mixed and inconclusive. Physicians respond to some P4P incentives but not the others. In general, physicians' response to these financial incentives is of modest size with no evidence of ultimate health improvement for the patients.

3. Ontario's Natural Experiment

This study draws on primary care reform interventions in Ontario, Canada as a natural experiment of P4P incentive payments to address the following questions: 1) Does P4P stimulate the delivery of targeted health care services by GPs? 2) Are P4P incentives effects heterogeneous across physician and practice characteristics? Primary care reform in Ontario provided a set of performance-based incentives to some of the primary care physicians in Ontario but not to the others. This produces natural treatment and comparison groups by which to identify the effect of P4P incentives on physician behavior. The ten-year study period (fiscal years 1998/1999-2007/2008) covers years prior to the provision of the performance-based incentives and those after the implementation. At the beginning of the study period in April 1998, all but a few hundred primary care physicians in Ontario were in the traditional fee-for-service practice; at the end of the study period, more than half of these GPs converted to one or more of the primary care reform models that included P4P inventive.

3.1 Background: Primary Care Reform

Over the last two decades, the province of Ontario, Canada has launched a series of primary care renewal (PCR) models to improve the quality of primary health care. The PCR models are intended to improve quality by: 1) providing P4P incentives to stimulate the delivery of targeted health care services; 2) converting from traditional fee-for-service payment to a blended payment method; 3) integrating primary care physicians, nurses and other professionals into more collaborative, multidisciplinary teams (Wilson 2006).

The Ontario Ministry of Health and Long-Term Care (MOHLTC) introduced the different PCR models at different points of time for different purposes. This study focuses on four PCR models: the Family Health Network model (FHN), the Family Health Group model (FHG), the Comprehensive Care Model (CCM) and the Family Health Organization (FHO). The earliest model introduced among these four PCR models is the FHN, which existed as early as 2002, requires a group practice with at least 3 GP, and is funded through a blended system of capitation for "core" services provided to rostered patients and fee-for-service for both non-rostered patients and for "non-core" services excluded from the basket of capitated services. FHGs were introduced in 2003, also required a group of 3 or more GPs but the basic payment scheme is a enhanced fee-for-service formula, which consists of the traditional fee-for-service payment for usual care, plus some capitation payments for comprehensive care services provided to rostered patients. The CCM model was introduced in 2005, can include only a solo GP, and is funded through fee-for-service. It is the most similar to traditional FFS practice. The FHO model was introduced in 2006, like FHNs, and FHGs requires a group of at least 3 GPs, and is funded through a blend of capitation payment and fee-for-service payment for non-rostered patients and for "non-core" services. FHOs and FHNs are similar in the funding scheme but different in size or rostering regulation. There is no size regulation in patient roster size for the FHO model, but for FHN practices the required minimum roster size is 2,400 patients for a group of 3 GPs while a financial penalty applies if the average roster size is greater than 2400 patients/GP in the practice. Unlike traditional fee-for-service practice, all of the above four PCR models offer enrolment to their patients (optional for FHGs, required for FHNs, CCMs and FHOs), provide comprehensive care, impose requirements on GPs to provide a minimum of after-hours care.

3.2 Pay-for-performance Incentives

Ontario initially introduced elements of pay-for-performance in primary care in 1999 to some small-scale pilot PCR models, and expanded it within primary care in 2004. The 2004

Physician Services Agreement included a large number of incentives targeting various aspects of the organization of PCR practices and the care delivered by physicians in those practices. Further, as discussed below, the specific incentives and dates of eligibility differ across the various PCR models.

We focus on a set of P4P financial incentives for five preventive care services (referred to as the Service Enhancement Payments for Preventive Care): Pap smears, mammograms, flu shot for seniors, toddler immunizations, and colorectal cancer screening; and on special payments for services in six areas of care of particular interest to the MOHLTC: payments for obstetrical deliveries, hospital services, palliative care, office procedures, prenatal care, and home visits. Table 1 lists the details of the five performance-based incentives for preventive care services and the six special payments for designated sets of services.

3.2.1 P4P incentives for Preventive Care

The P4P incentives for the five preventive care services include two components: a contact payment and the cumulative preventive care bonus payment. The contact payment rewards PCR practices for contacting patients to schedule an appointment to receive a targeted preventive service. Specifically, the PCR practice receives a contact payment of $6.86 for each eligible patient in the target population that it contacts and for which it provides the Ministry the required documentation. The cumulative preventive care bonus payment rewards PCR practices for achieving high rates of coverage for the targeted preventive services in the physician's practice populations.

Physicians receive cumulative bonus payment for each service on March 31 each year based on the proportion of its physicians' eligible and rostered patients who received the targeted service over a specified period of time prior to March 31. Physicians receive a specified amount of money if the proportion reaches a pre-specified coverage threshold, and the payment grows as the proportion exceeds higher thresholds. For example, if 60% of a physician's rostered female patients in the age of 35 to 69 received a Pap smear for cervical cancer screening during the previous 30 months as of March 31, a physician is rewarded 220 dollars. If 65% of the eligible patient population received a pap smear, a physician receives 440 dollars. The physician is compensated with 660 dollars, 1,320 dollars and 2,200 dollars for coverage rates of 70%, 75% and 80%, respectively. It should be noted that, the bonus payment is only based on the proportion of a physician's rostered and eligible patients who received the service in the defined time period; the physician with whom the patient is rostered on March 31 need not have provided this service. For example, if a physician provided a pap smear to a patient on February 1 and that patient changed physicians on March 1, the patient's receipt of the Pap smear would count toward the second physician's bonus calculation on March 31.

It should also be noted that, although the payment is based on the performance of individual physicians, whether the payment is made directly to individual physician varies across the four PCR models. The payment is made to the physician's PCR practice for GPs in a FHN; how the practice uses the funds received is determined by the practice.[6] Physicians in FHGs, CCMs and FHOs receive the payment directly; it does not go to the PCR practice.

3.2.2 P4P Special Payments

The special payments are structured differently. In each case, a physician received a fixed payment if the targeted service was delivered to a minimum absolute level of service provision during the preceding fiscal year, where that minimum is defined in terms of number of services, dollar value of services, number of patients, or a combination of these factors. For each incentive there is also only a single threshold level: if it is reached, the physician receives the special payment; if it is not reached, the physician does not receive the payment. For example, if five or more obstetrical services[7] were delivered to five or more patients in a fiscal year, a physician receives a fixed payment of 3,200 dollars (with an increase to 5,000 dollars since October 2007). Unlike the preventive care bonuses, the services had to be provided by the physician. Moreover, for all six designated services, the payments were made directly to the physician.

3.3 Eligible Physicians

Not all GPs in Ontario were eligible for these financial incentives. In general, these financial incentives were offered only to physicians practicing in a PCR practice. Therefore, physicians who remained in fee-for-service practices were never eligible to receive these P4P incentives. Only physicians who converted from traditional fee-for-service to PCR models were eligible

[6] Beginning in 2006, if there is unanimous agreement among the physicians in a FHN practice, the practice could request that the payments be made directly to its individual physicians rather than the FHN. We have no information on the number of practices that have exercised this option. For ease of exposition, for all incentives we refer to "whether a physician receives a payment" even in those instances when the payment was made to the practice rather than the physician.

[7] Specific services eligible to count toward this special payment include: vaginal delivery, attendance at labor and delivery, Caesarean section, attendance at labor when patient transferred to another centre for delivery, etc.

for some or all of the P4P incentives. Furthermore, eligibility of these P4P incentives differs by PCR models. As a result, physicians were eligible for a P4P incentive only after they converted to one of the PCR models and only after the P4P incentives were in effect for the specific PCR model they joined. During the study period of 1999-2008, the P4P incentives were provided at different time points to the four PCR models. Table 2 presents the eligibility timing for the 11 targeted services by PCR models types.

As only some physicians in Ontario were entitled to these P4P incentives, this policy intervention serves as a natural experiment that we can exploit to identify the casual effect of P4P incentives. Since we can observe the practice activities of almost every GP in Ontario over 10 years (1999-2008) and because this period spans the introduction of P4P incentives implementation, we can assess the impact of P4P incentives within a difference-in-differences framework by comparing the responses of the GPs exposed to the P4P incentives against those not exposed to the P4P incentives.

Of course, the natural experiment formed by this intervention poses some difficulties for the identification. First, physicians are not randomly assigned to the PCR models. This will lead to selection bias if we use simple difference-in-differences mean comparison on the responses from eligible GPs against ineligible GPs. Moreover, the PCR model practices are different from the traditional fee-for-service practice in various aspects. Table 3 lists the main differences among each of the four PCR models in the aspects of general payment scheme, practice composition, after-hour services and patient enrolment requirement. Traditional FFS GPs receive only FFS payments, while all PCR model GPs receive a blend of capitation payment and FFS payments, with different proportions of these two components. Unlike the traditional FFS practices, most of the PCR models require GPs to work in group practice (the only exception is CCMs that allows solo practice). Also PCR model GPs have to provide extended services, nurse-staffed telephone health advisory services and on-call services. Lastly, patient enrolment is required in these PCR models except for FHGs but not for FFS GPs. As a result, the identification of the P4P incentive effects may be confounded by differences between the traditional fee-for-service practices and the PCR model practices..

In spite of these problems, the implementation of the performance-based incentives in Ontario still allows us identify empirically the P4P incentive effects using several identification strategies to mitigate selection bias and control for confounding effects. As described in the method section below, eligibility for the incentive payments is not perfectly confounded with joining a PCR: some GPs joined a PCR model before they became eligible for bonus payments (unaware that they would later become eligible for such payments). This enables the evaluation to distinguish the effects of the incentive payments from the effect of joining a new practice model. Furthermore, variation in general payment scheme and practice setting among the four PCR models themselves provides us an opportunity to disentangle the effect of P4P incentives from that of other primary care reform features.

4. Data

4.1 Data Sources

The study draws primarily on four administrative databases of the Ontario Ministry of Health and Long-Term Care (MOHLTC), linked by patient encrypted health number and physician encrypted number. *OHIP Claims Database* provided information on all OHIP-funded services received by each resident of Ontario each month of the study period; the *Registered Persons Database* provided basic information on each OHIP beneficiary; the *Corporate Provider Database* provided basic information on each physician and his or her practice; the *Client Agency Program Enrolment* (CAPE) file provided information on the patient roster for each physician in a PCR practice. OHIP claims data allowed us to identify all services provided by every primary care physician in Ontario. The *Client Agency Program Enrolment* (CAPE) data allowed us to match every patient to a physician enrolled in a PCR practice, and to identify if this beneficiary should be counted towards the targeted population for each incentive payment. This data plus the *Registered Persons Database* provided us the characteristics of the patient population for each practice. *OHIP Claims Database* allowed us to construct the yearly utilization rate of each of the targeted services for every physician. The *Corporate Provider Database* allowed us to identify if a GP was enrolled with any of the PCR models at any point of time during the study period. Together these four databases enabled us to construct for each primary care physician in the province of Ontario, a measure of their practice population each year and a record of all services received by those patients during the period of 1999 to 2008 fiscal years. (See Appendix 2 for all the data sources that we used and the corresponding information that we extracted from each source).

4.2 Study sample

The unit of this analysis is a physician. The analyses focus on community-based GPs that do not specialize in a subset of services. We used to following criteria to select the study sample: (1) include physicians who are GPs throughout the study period; (2) excluded part-time GPs who billed less than $30,000 each year; (3) to limit the study sample to GPs in an established practice, we only included physicians who had at least two consecutive years of practice before study period; (4) include GPs for whom the office-based consultations accounting for the majority of their activities; (5) exclude locums as they are not eligible for bonuses; (6) we excluded GPs affiliated with the PCR models for which we do not have sufficient data for the analyses; (7) we also exclude GPs who converted to FFS for more than one time during the study period for simplicity of the analyses. Table 4 documents how many physicians were excluded by the various criteria when they were applied in the order listed. After applying these criteria, we obtain a core sample of 2,185 GPs.

Since the eligibility scope and implementation dates for the 11 P4P incentives are different for the four PCR models, the composition and the final sample size of the treatment and control groups vary by the P4P incentives. Again for the simplicity of the analysis, we dropped the physicians whose "treatment" status turned on and off for more than one time during the study period[8]. The compositions of control and treatment groups as well as the final sample sizes are presented in Table 5, divided into three subsets of targeted services.

4.3 Variable Specification

4.3.1 Physician responses

Physician responses were measured differently for the preventive care services and the designated services for special payments. Because each of the preventive care bonuses is defined with respect to the proportion of a GP's practice population that has received a specified service as of March 31 each year, the outcome variable is defined as the rate of coverage for the relevant period each year for each preventive care service. For the special payments, the outcome variable is defined as the number of services provided or the number of individuals to whom the designated services had been provided.

Analyzing the impact of the incentive payments requires that we identify each GP's practice population on March 31 of each year. Therefore, we used the following steps to define the practice patient population for each GP. For each year we assigned all patients in the Ontario Health Insurance Program (OHIP) physician claims database to a GP and thereby defined a practice population for each GP on March 31 of each year of the study period. Different methods were used to define practice populations for physicians in FFS and physicians in a PCR. Physicians in traditional FFS practice do not roster patients. We defined the practice population for these physicians using the validated methodology developed in Hutchison et al. (1997). Specifically, a physician's practice population is defined as: all individuals for whom the physician billed OHIP for at least one visit during the previous fiscal year; and all additional patients for whom the physician billed OHIP for at least one

[8] This might be switching back and forth between FFS practice and a PCR model, or switching back and forth between a PCR model which was eligible for the incentives and another PCR model which was not eligible for the incentives yet.

visit in each of the two preceding fiscal years. Patients who met these criteria for more than one physician were assigned to the physician who billed for the largest number of visits; if the number of visits was equal, assignment was based on the physician with most recent visit (details see Appendix 3). Physicians participating in PCR models have both rostered (the sizable majority) and non-rostered patients. For this case we define the practice population as the set of rostered patients (as indicated by the Ministry Client Agency Program Enrollment database) plus non-rostered patients as assigned by the Hutchison et al. algorithm. As a result, all OHIP beneficiaries were assigned to a physician for each year of the study period.

After assigning the patients to each physician based on OHIP claims, we counted the number of patients in each physician's practice who received a targeted service during the relevant period and constructed the dependent variables for the empirical analysis for each targeted service for each GP in each year. It should be noted that for mammogram and senior flu shot, this study requires additional data, because patients can receive these service at specialized clinics whose activity is not captured by the OHIP claims database. For mammogram we were able to merge individual-level data on services used in these clinics[9] and so capture all mammograms in the province. For senior flu shot we were not able to do this, so our data exclude such service provision. This has limited our ability to get an unbiased estimate of P4P incentive effect for this service. We will return to, and discuss this limitation in the method section below.

For the five preventive care bonuses, the dependent variable of each targeted service is defined, as of March 31 each year, as the proportion of a GP's practice population that received the service in question during the relevant period prior to that March 31st. For PCR GPs, this variable is constructed using data from rostered patients only because Ministry's criterion for payment of the bonus is defined in reference to rostered patients only. We conduct a sensitivity analysis (see section 6.1.3 below) using an alternate dependent variable that includes both the rostered and non-rostered patients for GPs in PCR models so to obtain a measure that is more consistent across traditional FFS and PCR physicians. A further complication with this dependent variable definition is that PCR physicians can bill a "tracking code" for patients who receive a flu shot at specialized clinics rather than the GP's office, an option not available to FFS physicians. We conduct sensitivity analyses regarding the use of such codes to define flu shot uptake among PCR practices to test the robustness of the findings to this potential problem.

[9] There is a provincial program —Ontario Breast-cancer Screening Program—from which patients can also receive mammograms but these activities were not included in the OHIP claims. Therefore, we y integrated this part of data provided by Cancer Care Ontario (CCO) into our analysis for mammograms. Unfortunately we couldn't get any data for this type of programs for flu shots so the results for senior flu shots suffer from this data limitation.

For the six special payments the dependent variable of each designated service is defined dichotomously, taking on the value of 1 if the physician's service provision met the criteria for the special payment of interest, and 0 if it did not.

4.3.2 Independent variables

As noted above, the *Corporate Provider Database* allowed us to identify if, and when a GP joined a PCR practice during the study period. Based on this, we constructed a treated/control dummy indicating if a GP was ever eligible for the incentives during the study period, a pre- and post- dummy indicating if an observation was from a period is before or after the implantation of P4P incentives, and a treatment dummy which is an interaction term between the above two dummies, taking on the value of 1 when a GP was eligible for the incentive during the time period in question.

In addition, we included in the analyses a set of independent variables that represent both the supply-side and demand-side characteristics of service utilization. These include characteristics of a physician and the physician's practice, and basic information of the physician's patient population. Physician-specific characteristics are physician age, sex, years in practice, activity level measured by total value of claims submitted each year, and a set of work-load variables including days of work, number of patient visits and number of patient visits per working day. Practice-specific variables include: practice model (FFS, FHN, FHG, CCM and FHO), size of practice population, and a set of practice location characteristics measured by metropolitan influence zone (MIZ) categories and a rurality index of Ontario (RIO). The MIZ categories indicate the degree of influence that metropolitan areas have on the geographic location of a practice; the RIO score indicates the degree of ruralness of a practice location. We also control for a set of patient population characteristics, including the mean age of a physician's patient population, and the proportions of female, infant and elderly patients in the practice. The detailed covariate definitions are listed in Appendix 4.

4.3.3 Descriptive statistics of independent variables

Table 6 presents sample descriptive statistics at the pre-intervention baseline, defined as of March 31, 2003, disaggregated by the control group and the incentive group.[10] The control group GPs differ at baseline from incentive group GPs. First, incentive group GPs are younger and have fewer years of practice experience than control group GPs. This observed difference is not surprising because we expect that GPs whose complying costs are relatively

[10] As noted above, the definition of the control groups differs slightly across some of the incentives, but the patterns are so similar across the cases that we have collapsed them into one table.

smaller are more likely to participate in the PCR models. Younger GPs are more flexible in practice style thus more easily to fit in with the specific rules of the PCR practice. Second, a higher proportion of incentive group GPs are female doctors than control group GPs. This might be due to the fact that female GPs are more interested in, or better at, collaborative team production. Third, for all five bonuses and the special payment on palliative care, incentive group GPs worked more days and more intensively than the control group GPs before the intervention and they had bigger practice size. But this pattern is reversed for the other five special payments on obstetrical deliveries, hospital services, office procedures, prenatal care and home visits. For the five bonuses and the special payments on palliative care, the patient population demographics are similar between incentive and control groups, except for that the incentive group GPs have practices with slightly more female and infant patients. For the other five special payments, the incentive group GPs also have practices with slightly more female and infant patients, but they also have an older patient population. Finally, incentive group GPs are more homogenous (as indicated by smaller standards deviations) than control group GPs.

5. Empirical Methods

5.1 Addressing Possible Sources of Bias

As described above, the policy intervention in Ontario serves as a natural experiment that we can exploit to identify the casual effect of P4P incentives. The treatment of interest is a set of P4P incentives targeted on 11 specific health care services or sets of services. Specifically, this policy intervention conditions the eligibility of the P4P incentives on the PCR model-participation status. A simple difference-in-differences approach can provide us an estimate of the P4P incentive effects by directly comparing the mean change across the PCR model GPs and the FFS GPs. However, voluntary participation generates non-random assignment of GPs to treatment, invalidating the simple difference-in-differences approach (Meyer 1995). In other words, we expect that the "treated" GPs are systematically different from the "non-treated" ones and these differences may contribute to the observed difference in the response of GPs to P4P incentives. Therefore, the identification of causal effect hinges on how well the selected comparison group represents the counterfactual of the treatment group, and on the extent to which we can mitigate selection bias.

As noted in the descriptive statistics above, GPs who join PCR differ from those who stay in FFS at the pre-intervention baseline in a number of ways. For example, PCR GPs are younger and have fewer years of practice experience, and their workload is in general different from those in traditional fee-for-service. These differences in physician characteristics might cause estimation bias generated by both "selection on observables" and "selection on unobservables". We discussed in section 5.2.1 below to discuss the empirical strategies we used to mitigate selection bias.

One might also be concerned about possible confounding from other factors-- it is possible that some of the observed differences in response to P4P incentives between treatment group and control group are actually caused by other unobserved attributes pertaining to the PCR practice rather than the P4P incentives. For example, an important institutional difference between some PCR and FFS practice is that these PCR models are paid by a mixture of FFS and capitation instead of traditional FFS piece rate. One might expect that FFS physicians respond less to P4P bonus related to preventive care services because the opportunity cost may be greater for FFS physicians than for physicians paid by capitation or salary in the sense that doing more preventive care may preclude the provision of other services that generate higher fees per unit time. Another type of confounding may arise if we are concerned about separate initiatives that influence the level of utilization of the services being analyzed. The potential sources of this type of confounding and strategies we used to control for them are described in section 5.2.2 below.

5.2 Identification strategies

5.2.1 Strategies to mitigate selection bias

We employ several identification strategies to mitigate the selection bias that may be generated by both observable and unobservable physician characteristics. First, we control for important aspects of physician characteristics and practice characteristics that might be correlated with the self-selection process and are also important in determining the provision of the targeted services. The data allows us to control for physician characteristics including physician demographics, work experience, and work load measures; and practice characteristics including practice size, geographical location of the practice and patient population characteristics of the practice.

Second, to address selection bias generated by unobservable characteristics, we exploit the longitudinal nature of the data and employ a difference-in-differences approach with individual fixed effects. As noted above, GPs may self-select into PCR models through a process linked to unobserved physician characteristics. This type of selection bias can be reduced to the extent that the unobserved components that determine both the self-selection behaviour and the outcomes are physician-specific and time-invariant, and thereby can be differenced out by a difference-in-differences approach with individual fixed effects.

A potential limitation of the above approach is the lack of control for unobserved temporal individual-specific component that affected the selection into the treatment group and control groups (Blundell and Costa Dias 2000). This could be a problem if some GPs self-selected into PCR models because of temporary shocks that are directly related to the targeted health care services. However, this should not be a big concern in this study for the following reasons. Firstly, participating into a PCR model is unlikely to depend on short-term changes that affect the utilization rates of the targeted services, such as a sudden demand-side change or an onset of other simultaneous policies that are targeted to these specific services. The monetary values of these P4P incentives are a relatively very small proportion of the total income of GPs. So it is unlikely that any temporary changes related to the targeted services caused the conversion behavior. This assumption is reinforced by the fact that only a very small proportion of GPs who converted from FFS practice to PCR models switched back to FFS practice during the study period of ten years. Secondly, any unobserved temporary shocks that are correlated with PCR participation should not play an important role in determining the utilization of the specific services that are targeted by P4P incentives, because most of the treatment group GPs already converted a number of years prior to becoming eligible for the P4P incentives. Hence, the incentives are unlikely to be the reason that motivated the conversion behavior.

5.2.2 Strategies to control for confounding effects

We are concerned about potential confounding from a PCR-practice effect because PCR practices have features (beyond the P4P incentives) not found in traditional fee-for-service practices. We argue that this type of confounding can be controlled in the analyses in the following ways. First, the eligibility timing of the P4P incentives in the PCR models facilitates the reduction of this confounding. The policy intervention provided the P4P incentives to different PCR models in different time periods, but it created essentially three types of physicians groups: non-incentive group, incentive group 1 and incentive group 2 (see Figure 1). The non-incentive group consists of the GPs who remain in FFS over the study period. Since they were never eligible for the incentives, they are used as the legitimate control group in the difference-in-differences design. The Incentive group 1 consists of the GPs who joined a PCR model and simultaneously became eligible for the P4P incentives. This group of physician can be used as part of the treatment group but this is problematic--given the participation in PCR models is a voluntary process, the P4P incentive effect is perfectly confounded by the selection into the PCR model for this group of physicians. The Incentive group 2 consists of the GPs who joined a PCR model before the P4P incentives were introduced and who therefore became eligible for the P4P incentives only after they had participated for some time in a PCR model. This group of GPs pertains to the majority of physicians who were entitled with the incentives in this study. Using this group of physicians as the treatment group can mitigate the problem of confounding: because these physicians chose to participate in PCR before (and with no expectation of future P4P incentives) the introduction of the P4P incentives, the incentive effect is not perfectly confounded by the other PCR-model features. Second, we use alternative treatment groups in the comparison to mitigate confounding from some specific PCR attributes. This approach is possible for this study since we can exploit the variation on several dimensions across different PCR models to conduct falsification tests on the effect of some specific confounders over the P4P incentive effect. For example, to rule out the possibility that the difference in general payment scheme is causing the difference in response, we restrict the treatment GPs as those PCR GPs who were also compensated mainly by fee-for-service scheme and compare their behaviour with the FFS control group GPs. If we still observe the difference in response, we can conclude that it is likely not the general payment scheme causing the observed P4P incentive effects.

Our identification is complicated by potential confounding effects of separate initiatives that could influence the level of utilization of preventive care services during the study period. Potential confounding from such other initiatives is of greatest concern for senior flu shot, breast cancer screening and colorectal screening. The province has invested heavily in its universal flu vaccination program since 2000, both in making the flu shot available through special clinics and in promoting the up-take of the flu shot. Flu shots obtained through a flu-shot clinic rather than in the GP office are not recorded in the OHIP database. Similarly, women can obtain a mammogram through the Ontario Breast Screening Program, which offers specialized clinics for mammograms. Mammograms obtained through these clinics are also not recorded in the OHIP claims database, though, as noted above, we are able to capture such utilization by integrating data from Cancer Care Ontario, the provincial agency that oversees the breast screening program. Finally, beginning in 2004 Ontario launched a pilot program to encourage colorectal cancer screening, and in 2007 launched a population-based colorectal cancer screening program ("ColonCancerCheck") in collaboration with Cancer Care Ontario.

However, none of these initiatives are specific to patients in PCR practices: they offer services to all eligible Ontario residents. Consequently, the inclusion of the fee-for-service control group controls for the general impact of these programs on the receipt of the respective services through GP offices as long as they affected provision equally for physicians in the control and treatment groups. A problem arises only if there is an interaction effect between these programs and treatment/control status. One concern for flu shots and mammograms is that physicians eligible for incentive payments may have differential incentive to encourage their patients to receive the service through the GP office (and captured by the OHIP database) rather than one of the specialized clinics (not captured by OHIP). Because we capture all mammogram utilization (that included in OHIP and that from Cancer Care Ontario) this does not pose a problem for mammogram. But for flu shot we do not capture shots provided in specialized clinics, and in the presence of a differential incentive, this omission would lead to an over-estimate of the effect of the incentive payment.

Finally, the identification of the difference-in-differences with individual-fixed effects approach is based on the assumption of a parallel trend between treatment and control groups. In order to control for the different time trends across treatment group and control group, we use the difference-in-differences adjusting for differential trends approach as suggested by Bell, Blundell and Reenen (1999). This model relaxes the assumption of parallel trends between the control and treatment group GP when these differential trends have different impact on the outcome between P4P system and non-P4P system.

5.3 Empirical specifications

We employ the following empirical approaches to evaluate the impact of the P4P incentives.

5.3.1 Simple difference-in-differences with pooled OLS

The effect of each P4P incentives can be estimated by comparing the treatment and comparison group in the behaviour change before- and after- the exposure to the incentives. Consider the model:

$$Y_{itj} = \beta_j' X_{it} + \gamma_j T_t + \rho_j D_i + \delta_j T_t * D_i + \theta_{tj} + \mu_{itj} \quad (1)$$

where Y_{itj} is the utilization score of service j for physician i in fiscal year t; X_{it} is a set of covariates; T_t is a treatment dummy equal to 1 if this is post-period and 0 otherwise; D_i is a treatment dummy equal to 1 if this physician is in treatment group and 0 other wise; T_t*D_i is the interaction term taking on a value of 1 if GP i was exposed to the P4P incentives at time t. The estimated coefficient of this term, δ_j indicates the difference-in-differences P4P incentive effect. θ_{tj} is a set of year dummies; μ_{itj} is the idiosyncratic term. The above equation is estimated by a pooled linear or nonlinear panel data model.

In order to account for possibly serial correlation of the dependent variable over time, we adjust the standard errors by clustering on individual physician level in the above simple DID estimation and for all the DID models below. This would mitigate the over-rejection problem for DID estimates (see Bertrand, Duflo and Mullainathan 2004) when the inference of the regular t-statistic is based on unadjusted standard errors[11].

5.3.2 Difference-in-differences with individual fixed effects

In order to control for fixed unobserved factors that could influence both selection into a PCR model and provision of the targeted services, we add in a set of individual-specific fixed effects:

$$Y_{itj} = \beta_j' X_{it} + \gamma_j T_t + \rho_j D_i + \delta_j T_t * D_i + \theta_{tj} + \varphi_{ij} + \mu_{itj} \quad (2)$$

where $\varphi_{ij}$is a set of physician dummies; μ_{itj} is the idiosyncratic term. The above equation is estimated by a fixed effect linear or nonlinear panel data model.

5.3.3 Difference-in-differences with differential trend model

To relax the parallel trend assumption we use the difference-in-differences with differential trend model suggested by Bell, Blundell and Reenen (1999). This specification assumes that:

$$\begin{cases} e_{it} = \varphi_i + k_p m_t + \mu_{it} \;\; if \;\; T_t * D_i = 1 \\ e_{it} = \varphi_i + k_n m_t + \mu_{it} \;\; if \;\; T_t * D_i = 0 \end{cases} \quad (3)$$

where e_{it} captures the unobservables and the noise. m_t is an unobserved trend. If the P4P GPs and non-P4P GPs have different trends, the impact of these trends is allowed to differ across the two groups, which is captured by k_p and k_n [12]. This paper follows the regression operationalization of Wagstaff and Moreno-Serra (2009). Incorporating the assumption described in (3), we get the following model:

$$Y_{itj} = \beta_j' X_{it} + \gamma_j T_t + \rho_j D_i + \delta_j T_t * D_i + \varphi_{ij} + k_{nj} m_t + \left(k_{pj} - k_{nj}\right) m_t (T_t * D_i) + \mu_{itj} \quad (4)$$

which can be estimated by a fixed effects model including year dummies and year dummies interacted with the treatment dummy, i.e.

[11] We use the “cluster” option in STATA estimation commands to adjust for standard errors for intragroup correlation among observations over time for each physician. As Bertrand et al. noted (Bertrand, Duflo and Mullainathan 2004), this type of adjustment works well when the number of clusters is large (e.g. N is greater than 50). Our sample size (number of physicians) is sufficiently big for this adjustment to mitigate this problem.

[12] Note that P is the subscript for “P4P” group trend; N is the subscript for “Non-P4P” group trend.

$$Y_{itj} = \beta_j' X_{it} + \gamma_j T_t + \rho_j D_i + \delta_j T_t * D_i + \varphi_{ij} + \sum\nolimits_{\tau=2}^{T} \alpha_{\tau j} Year_\tau + \sum\nolimits_{\tau=2}^{T} \theta_{\tau j} Year_\tau \, (T_t * D_i) + \mu_{itj} \qquad (5)$$

In the above model the impact of P4P incentives varies over time, but the average impact of P4P incentives can be estimated as:

$$\text{Mean P4P impact} = \hat{\delta} + \frac{\sum_{\tau=2}^{T} \theta_{\tau j}}{T-1} \qquad (6)$$

Because the parallel trend assumption implies$k_p = k_n$, this assumption can be tested by testing the nonlinear restriction:

$$\frac{\sum_t m_t (k_{pj} - k_{nj})}{\sum_t m_t \, k_{nj}} = \frac{(k_{pj} - k_{nj}) \sum_t m_t}{k_{nj} \sum_t m_t} = \frac{\sum_{\tau=2}^{T} \theta_{\tau j}}{\sum_{\tau=2}^{T} \alpha_{\tau j}} = 0 \qquad (7)$$

Non-rejection of the hypothesis would suggest that $k_{pj} = k_{nj}$ and provide evidence in favor of the parallel trend assumption and the difference-in-differences with individual fixed effects model.

It should be noted that we could only run a full set of regression analyses described above for the five bonuses, but not for the six special payments. As indicated by table 5, because the eligibility scope and implementation dates for the 11 P4P incentives are different for the four PCR models, the composition and the final sample size of the treatment and control groups vary by the P4P incentives. Most of our P4P GPs became eligible for the five preventive care bonuses in 2006, except colorectal cancer screening, for which most of our P4P GPs became eligible in 2005. For the six special payments, most of the P4P GPs became eligible in 2005, 2006 and 2007, except palliative care special payment, for which most of the P4P GPs became eligible in 2003, 2004 and 2005. Accordingly, for the six special payments, we estimate the difference-in-differences models separately for three subsets of P4P GPs based on the year they became eligible for the payments. Moreover, as the five bonuses were provided to all four PCR models considered in the study while the six special payments were provided to only some of the PCR models (e.g. FHNs and FHOs), there are much fewer GPs constituting the treatment groups in the analysis for the six special payments than for the five cumulative bonuses. As a result, we could estimate the full set of difference-in-differences models and conduct the robustness checks and sensitivity analyses for the five bonuses, but could only estimate the simple pooled difference-in-differences model with the full sample for the six special payments. For the same reason, we could only conduct subgroup analyses for the five bonuses, but not for the six special payments.

6. Empirical Results

6.1 Descriptive trends of physician responses

We can only document the extent to which GPs contacted patients to arrange the receipt of preventive services for the period after the identifying codes were introduced in the fee schedule. Table 7 presents the proportion of eligible physicians who submitted at least one claim for contacting a patient to arrange an appointment to deliver a preventive care service. Two things are notable: (1) the rate of uptake is relatively low — with the exception of a couple of years for the senior flu shot, less than 45% of eligible physicians submitted even a single claim; (2) and there is no noticeable upward trend — in fact, the proportion has been falling in recent years for 4 of the 5 services. Figure 2 presents the mean number of claims per eligible physician for each service. For all services except senior flu shot, the mean number of claims per eligible physician is fewer than 20; even for senior flu shots the mean exceeds 100 for only one year during this period. Overall, there appears to have been little response to these contact incentive payments.

The main outcome measures for this study are the utilization rates of the services that are targeted by the 11 P4P incentives. The unadjusted time paths of compliance level of all the targeted services are shown in Figure 3 to Figure 13. The horizontal axis represents the years from March 31, 1999 to March 31, 2008. For the preventive care services, the vertical axis is the mean proportion of patients who received the targeted services. For the special payments the vertical axis is the proportion of physicians who achieved the targeted performance level. The lines represent the time trends for the control group, and for treatment groups defined in terms of the year when a GP first became eligible for the incentive. For example, incent2003 represent GPs who were first eligible during fiscal year 2002-2003. We can detect a specific pattern of change in trend to the introduction of the P4P incentives for Pap Smear, colorectal cancer screenings, and palliative care. For these services, compared to the control group, provision in the incentive groups started to increase and diverge at the time of exposure to the incentives. This suggests possible effects of the P4P incentive payments for these services. The trend for mammogram displays equivocal evidence. We could not detect any specific pattern in the trend for senior flu shot, toddler immunizations, obstetrical deliveries, hospital services, office procedures, prenatal care and home visits. For most of these 11 services, incentive group GPs started with higher baseline compliance levels so it is possible that selection effect exists in the means of the compliance levels for treatment group versus control group.

6.2 Estimation results for the preventive care bonuses

6.2.1 Estimates for the full sample

Table 8 presents the estimates of the P4P incentive effects for the five preventive care bonuses based on three difference-in-differences models for the full sample. Column (a) lists the baseline compliance level of each targeted service, which is defined as the average utilization rate of this service in 2003. Panel (b), (c) and (d) of Table 8 present the estimates of the P4P incentive effects based on difference-in-differences with pooled OLS model, difference-in-differences with individual fixed effects model, and difference-in-differences with differential trend model, respectively. The marginal effects estimates indicate the percentage change of the service provision due to the introduction of each bonus payment. In order to account for possible correlation of the observations over time for each physician, we calculated the robust standard errors by clustering by individual physician. The results based on difference-in-differences with fixed individual effects model show that the bonus payment had a statistically significant effect on the provision of senior flu shot, Pap Smear, mammogram and colorectal cancer screening, while its effect on the provision of toddler immunization is not statistically significant[13]. The absolute level of increase in compliance is 2.8%, 4.1%, 1.8% and 8.5% for senior flu shot, Pap smear, mammogram and colorectal cancer screening, respectively. It is notable that the marginal effect estimates based on difference-in-differences with pooled OLS model are similar to the above figures, except that the incentive effect estimates are not significantly different from zero for senior flu shot and mammogram in the pooled OLS model. The estimates based on difference-in-differences with differential trend model are consistent with those from the individual fixed effects model while indicating slightly larger effects for all five services.

Panel (e) of Table 8 presents the test statistic and the p-value from the nonlinear restriction test on the parallel-trend assumption as described in equation (7). The results indicate that the null hypothesis of a common trend between the P4P GPs and non-P4P GPs is not rejected at the 5% level for senior flu shot, toddler immunization and mammogram but is rejected for Pap smears and colorectal cancer screening. Therefore, the parallel trend assumption is reasonable for flu shot for senior flu shot, toddler immunization and mammogram. For these services, we can trust the regression results from the difference-in-

[13] As noted by Moulton (1986, 1990), and Donald and Lang (2007), in regression models with mixture of individual and grouped data, the failure to account for the presence of common group errors can generate estimated standard errors that are biased downward dramatically. Our DID estimates may suffer from this problem, under the assumption that physicians in the same practice may have correlated standard errors. Accordingly, adjusting the standard errors for clustering by practice (instead of clustering by individual physician) could correct for the over-rejection problem with our DID estimates. However, we have not done so because we could not properly identify physician practice for FFS GPs and CCM GPs. As a result, our current estimates overstate the statistical significance of the P4P incentive effects, and adjustment for clustering would weaken the evidence of an incentive effect.

differences with fixed individual effects model, while we prefer the results from the differential trend model to the results from the difference-in-differences with fixed individual effects model for pap smears and colorectal cancer screening. Even though the parallel trend assumption does not hold for some services, the magnitude of the estimated incentive effects is similar across these two models.

6.2.2 Robustness checks with alternative samples

The above results may be subject to bias due to the confounding from other PCR model characteristics. As a robustness check we restrict the treatment group to PCR GPs who joined PCR practices before becoming eligible for P4P incentives. This robustness check is conducted only for the five preventive care bonuses. Panel (b) in Table 9 presents the estimated P4P incentive effects from difference-in-differences with individual fixed effects model based on the sample of GPs who joined PCR practices before becoming eligible for P4P incentives. Column (a) lists the baseline compliance level of each targeted service in 2003 for this sample. The regression results show that the estimated P4P incentive effects are robust to this refinement of the treatment group. Therefore, we conclude that the estimated P4P incentive effects are unlikely to be generated by other PCR practice characteristics.

We also used a falsification test to check whether the observed incentive effects are linked to the general payment scheme, which differs between most treatment and control physicians. As a second robustness check we restrict the treatment group to GPs working in PCR practices that are paid primarily by FFS. If we observe the response of this subgroup of GPs is not significantly different from those in traditional FFS practices, we have evidence that compromises the estimated P4P incentive effects. Panel (d) of table 9 presents the estimates of P4P incentive effects based on the sample of GPs in PCR models funded primarily by FFS. The results indicate that refining the treatment group in this way does not change the estimates of the incentive effects from those based on the full sample. This is consistent with the full sample estimation and first robustness check. Overall, the results from these two robustness checks reassure us that the full-sample estimates do not suffer from the possible confounding of other PCR attributes.

6.2.3 Sensitivity analyses on the study design

In order to test the sensitivity of the regression results, we conducted four sensitivity analyses to address limitations on the study design and assumptions.

As noted above, one complication with the current study design is that PCR physicians can bill a "tracking code" for patients who receive a flu shot at specialized clinics rather than the GP's office, an option not available to FFS physicians. We conduct sensitivity analyses regarding the use of such codes to define flu shot uptake among PCR practices to test the sensitivity of the findings to this potential problem. In the first sensitivity analysis, we redefined the dependent variables for the incentive payments by excluding the claims from the PCR doctors with these shadow-billed tracking Q-codes. This would give us a lower bound of the true estimates of the incentive effects. Table 10 presents the estimated marginal effects based on difference-in-difference with individual fixed effects model for the full sample and the two alternative samples we used above. The results from this sensitivity analysis indicate that the base case estimates are robust to this change of definition in that the significance level of the incentive effects stays the same while the magnitude is slightly smaller (as we would expect). The basic conclusions from the main analysis hold.

The second sensitivity analysis aims to test for the consistency of different methods we used to calculate the performance level for the FFS doctors and the PCR doctors. As discussed previously, the dependent variable could not be defined identically for the PCR and the FFS GPs: for PCR GPs, this variable is constructed using data from rostered patients only because Ministry's criterion for payment of the bonus is defined in reference to rostered patients only. Therefore, we added in the non-rosterd patients for the PCR doctors in the calculation of dependent variables and compare the performance level based both rostered and non-rostered patients against that of the FFS doctors. Table 11 presents the estimated marginal effects for this sensitivity analysis based on difference-in-difference with individual fixed effects model for the full sample and the two alternative samples we used above. The results show that the estimates are robust to this change of dependent variable definition. The magnitude of the estimated incentive effect is slightly larger under the second sensitivity analysis. We interpret this as the selection effect introduced by our algorithm of assigning the patients—the non-rostered patients within a PCR doctor's patient population that are assigned by our algorithm are utilizers of the services thus are more likely to get preventive care services from this doctor.

One additional rule for the FHG and CCM doctors to get the bonus payment is that they need to reach a minimum threshold of patient roster size. In the third sensitivity analysis, we estimate the incentive effects over only the subset of FHG and CCM doctors with rosters over the minimum threshold at the time of bonus introduction. The results from this sensitivity analyses represent short-run responses as opposed to long-run responses represented by the base case results, because the patient roster size could be endogenized over time. Table 12 presents the estimated marginal effects for the third sensitivity analysis based on difference-in-differences with individual fixed effects model for the full sample and the two alternative samples we used above. The results from this sensitivity analysis confirm the significance of the incentive effects and show a slightly larger response from the doctors who had already achieved the minimum roster size.

At the beginning of the bonus payment introduction, physicians might be just starting to enroll patients into their practice or ramping up for the incentive payment. So the calculated performance level might be noisy in the sense that the targeted performance is based on the proportion of patients who received the services. In the fourth analyses we dropped the observation of transition year, i.e. the first year that the treatment group GPs became eligible for the bonuses, from the empirical analyses to remove the potential noise in the observed behavior in the first period of transition. Table 13 presents the estimated marginal effects for this sensitivity analysis based on difference-in-difference with individual fixed effects model for the full sample and the two alternative samples we used above. The estimated marginal effects of incentives are not sensitive to this change and are slightly larger in size than the base case results.

6.3 Estimation results for the special payments

Table 14 presents the P4P incentive effects for the six special payments from estimating difference-in-differences with pooled logit model. Column (a) lists the baseline compliance level of each targeted service, which is defined as the proportion of GPs whose pattern of service provision in 2003 exceeded the special payment target level. We present the results separately for the subsets of GPs who became eligible for these incentives in 2005, 2006 and 2007[14] in panel (b), (c) and (d) respectively. The marginal effect estimate indicates the absolute change in the proportion of physicians whose service provision is predicted to exceed the target level as a result of the special payments. There is no statistically significant incentive effect that is consistent over these three subsamples for any of these special payments[15]. Overall, the results suggest little if any response to the special payments: all the estimates are small and not statistically different from zero.

6.4 Estimates of the P4P incentive effects for subgroup analysis

[14] For the palliative care payment, the results are presented separately for the subsets of GPs who became eligible for the incentive since 2003, 2004 and 2005 instead.

[15] Statistically significant results are found only for one subsample for office procedures and prenatal care.

There are reasons to expect that responses may differ by physician age, practice size, and baseline level of compliance. To investigate this we conduct three sets of subgroup analyses and present the results in this section[16]. The subgroup analyses are conducted only for the five bonuses due to the small sample sizes for the six special payments.

The first panel of Table 15 presents the estimates from the difference-in-differences with individual fixed effects model for the subgroup analyses by physician age for the five preventive care bonuses. We see a clear age gradient for Pap smear, mammograms and colorectal cancer screening in which younger physicians respond more to the P4P bonuses than do older physicians. An age-gradient is not discernable for senior flu shot and toddler immunization. For senior flu shot, we observe that only middle-age physicians responded to the incentives. This indicates the possibility that the relatively weak incentive effect from the whole sample analysis for senior flu shot is driven by the response of the middle-age physicians. We only detect a statistically significant incentive effect in the oldest age group for toddler immunization bonus, but this effect is only weakly significant at the 10% level.

The second panel of Table 15 presents estimates by practice size. Overall, the results indicate that physicians with larger practices tend to be more responsive to the P4P incentives. For Pap smear there is essentially no difference in magnitude of the estimated effects across categories of practice size. For mammogram and senior flu shot, there is a statistically significant incentive effect only for the biggest practices but no effect for the small-size or mid-size practices. For colorectal cancer screening, the pattern is clear that the incentive effect is larger for bigger practices.

The third panel of Table 15 present the difference-in-differences estimates for the subgroup analysis by baseline level of compliance. For three of the five preventive services, the response is the greatest for those physicians with the lowest levels of baseline provision (senior flu shot and mammogram) or for physicians with the lowest and middle levels (colorectal cancer screening)[17]. This is consistent with the hypothetic pattern that physicians with lower baseline compliance level tend to respond more, except that for Pap smear, for which physicians in the middle quartiles responded the most.

7. Discussion and conclusions

[16] A hypothesis that response is associated with the size of the group a GP works cannot be tested with our data due to the missing information on group size for the control group GPs.

[17] Note that for mammogram, the incentive effects for middle-quartile and top quartile are about the same; while for colorectal cancer screening, the incentive effects for the lowest quartile and the middle quartile are about the same.

Our estimates of the incentive effects indicate that the cumulative preventive care bonus payments for pap smears, mammograms, senior flu shot and colorectal cancer screening have modestly improved the performances of GPs in the provision of these targeted services. The bonus on toddler immunizations and the special payments on obstetrical deliveries, hospital services, palliative care, office procedures, prenatal care and home visits, had no effect on the provision of these targeted services. The regression results are consistent and similar in magnitude across the series of difference-in-differences models that we use in sequence to partly control for "selection on observables" and "selection on unobservables". The results from the robustness checks with alternative study samples suggest that it is unlikely the baseline estimates are driven by confounding between P4P incentives and other features of PCR practices. The sensitivity analyses also indicate that the main regression results are robust to different definitions of dependent variables and alternative estimation samples, reassuring the validity of our study design.

In general, our empirical results agree with the empirical literature, which indicates little effect of employing P4P incentives to improve the quality of health care. Among the eleven incentives we considered, seven of them did not result in significant improvement of service provision, while the other four only slightly increased the utilization of the targeted services. As noted above, because we could not adjust the standard errors for possible clustering on practice in estimating the difference-in-differences models, our current estimates would overstate the statistical significance of the P4P incentive effects, reinforcing the general conclusion that these incentives were not very effective. Unlike evidence based on P4P programs in the U.S., our findings are derived from observations in a public-funded single-payer system, so the effect of P4P incentives is not confounded by the institution of multiple-payers. Moreover, we found that even for the incentives that generated responses, the magnitude of the response rates varies across targeted services, across physician characteristics and across practice settings. Specifically, physicians responded to the bonuses for preventive care services but not the special payments. Physician responses differ significantly across physician age and initial service provision level.

The different physician responses to the preventive service bonuses and the special payments may be due to a number of factors. First, the costs of complying are different for these two sets of incentives. The preventive care services targeted by the five bonuses do not require special costs in provision and are within the expected competency of a GP. Some preventive care services even can be provided by non-physician staff. However, the services targeted by the six special payments often require a fixed cost. Services like obstetrical deliveries often incur some cost related to insurance premiums and require a commitment to be available for deliveries at all hours of the day. Also, providing services like hospital visits and home visits involves additional time cost and re-organization costs (i.e. re-organizing one's schedule for visits). Therefore, relatively larger financial incentives are required to cover these costs and to generate desired behaviors by GPs. Second, providing preventive care is well documented to be effective and are well established as consistent with high-quality care, but services subject to special payments in Ontario have no strong link to quality of care. Lastly, the preventive care bonuses are complementary to other attributes of the PCR models while this is not true for the special payments. For example, unlike the physicians remaining in FFS practice, the physicians who participated in PCR models were eligible for financial support to adopt electronic medical record systems that can provide automatic reminders when a patient should receive regularly scheduled services. This feature is not as important for the provision of special services as for the provision of preventive care services.

The general take-away message from our empirical results is that physicians do not automatically respond to performance-based financial incentives as we expected. Although the principal-agent theory has suggested the potential to employ P4P incentive to motivate physicians for providing high-quality care, physician responses to such incentives are not easily predicted. The heterogeneity of physician responses found in our study suggests that physician behaviors may be constrained by a complex set of objectives that we do not directly observe. Therefore, more refined positive analyses on physician health care delivery are warranted for future implementations of employing different forms of incentives to elicit desired physician behaviors.

Overall, our results deliver a cautionary message regarding the effectiveness of employing pay-for-performance to increase the quality of health care. The overall small physician responses to the introduction of P4P incentives in Ontario indicate the rather low power of using these incentives to motivate high quality care. One possible reason is that the absolute size of the financial incentives for these services in general is too small to generate desired response from the physicians. After all, the income increase related to these incentives is only a small proportion relative to the total income for most of the GPs, so the increase of marginal utility related to this income increment only works very marginally in physician's service provision decisions. Nonetheless, we would then expect that it will be even more costly to achieve the pre-specified improvement of service provision if we continue to employ the same incentive structure. As indicated in the recent literature of pay-for-performance, the P4P incentives need to be more carefully designed (Christianson et al. 2008; Epstein 2006; Hutchison 2008). As noted above, the cost of complying may vary substantially among different types of procedures. Therefore, tailoring the absolute size of financial incentives for different targeted services according to the relative costs of complying may provide us a more cost-effective solution. Furthermore, our findings also suggest that there is only limited scope of using P4P incentives to increase the provision of targeted services. The employment of P4P incentives is only effective when the targeted performance or tasks are strongly linked to professional standard of high quality care. This is reflected in the fact that physicians tend to be more responsive to P4P incentives targeted on preventive care services, which are unquestionably consistent with medical guidelines of providing high-quality care. Therefore, future implementations of P4P incentives could be confined only to these services. Finally, the P4P incentives should be redesigned so that the target measures are more closely related to real standards of high quality care. For example, financial incentives can be linked to quality indicators that aim to increase the access of health care, or to those representative of evidence-based health care.

Further studies on the performance payment incentives can be extended to several directions. Like much of the current literature on P4P, we could not obtain any patient health outcome measures, so we only rely on utilization rates or the provision levels for the analyses. These measures may not be representative for the health care quality *per se* and patient health outcomes would be better indicators of quality. Therefore, it will be important to document the effect of P4P incentives on patient health outcomes if such data become available in the future. Moreover, it is interesting to test whether there is a “spill-over” effect of only rewarding the provision of a small subset of services. It is likely that physicians only reallocate their time or other resources from the unrewarded services to the rewarded services to obtain more income. Finally, exploring other factors that might be complementary to the P4P incentives will help us to design better P4P programs to elicit the optimal behaviour of physicians.

References

Amundson, G., L. I. Solberg, M. Reed, E. M. Martini, and R. Carlson. (2003) "Paying for quality improvement: Compliance with tobacco cessation guidelines", *Joint Commission Journal on Quality and Safety* 29 (2): 59-65.

Armour B, Pitts M, Maclean R, et al. (2001) "The Effect of Explicit Financial Incentives on Physician Behavior," *Archives of Internal Medicine*;161:1261-6.

Baker, G. (1992) "Incentive contracts and performance measurement". *The Journal of Political Economy*, Vol. 100. no. 3, pp. 598-614

Baker G. (2004) "Pay for performance incentive programs in healthcare: market dynamics and business process". San Francisco, CA: Med-Vantage;

Baker G., Carter B. (2005) "Provider Pay-for-performance incentive programs: 2004 National Study Results". San Francisco, Calif: Med-Vantage;

Beaulieu, ND, Horrigan, DR. (2005) "Putting smart money to work for quality improvement." *Health Services Research*. 2005 Oct; 40(5 Pt 1):1318-34

Bell, B., Blundell, R., Van Reenen, J. (1999) "Getting the unemployed back to work: the role of targeted wage subsidies." *International Tax and Public Finance*, 6 (3), 339-360.

Bertrand M., Duflo E. and Mullainathan S., (2004) "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, vol. 119(1), pages 249-275, February.

Blomqvist, A., (1991). "The doctor as double agent: Information asymmetry, health insurance, and medical care," *Journal of Health Economics*, Elsevier, vol. 10(4), pages 411-432

Blundell, R., Costa Dias, M. (2000) "Evaluation Methods for Non-experimental Data." *Fiscal Studies*, 21 (4), 427-468.

Boland GW, Halpern EF, Gazelle GS. (2010) "Radiologist Report Turnaround Time: Impact of Pay-for-Performance Measures" *AJR Am J Roentgenol.* 2010 Sep;195(3):707-11.

Campbell S, Reeves D, Kontopantelis E, Middleton E, Sibbald B, Roland M. (2007) "Quality of primary care in England with the introduction of pay for performance", *New England Journal of Medicine* 2007 Jul 12; 357(2):181-90.

Campbell S, Reeves D, Kontopantelis E, Sibbald B, and Roland D.M. (2009) "Effects of Pay for Performance on the Quality of Primary Care in England" *New England Journal of Medicine*; 361:368-378

Christianson, J. B., Leatherman S., Sutherland K. (2008). "Lessons From Evaluations of Purchaser Pay-for-Performance Programs: A Review of the Evidence." *Medical Care Research and Review*, 65(6_suppl): 5S-35.

Chung S, Palaniappan LP, Trujillo LM, Rubin HR, Luft HS.(2010) "Effect of physician-specific pay-for-performance incentives in a large group practice." *Am J Manag Care.* 2010 Feb 1;16(2):e35-42.

Coleman, Katie.,Reiter, Kristin L.,Fulwiler, Daniel. (2007) "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, Volume 18, Number 4, November 2007, pp. 966-983

Conrad, D. A., Christianson, J. B. (2004) "Penetrating the "Black Box": Financial Incentives for Enhancing the Quality of Physician Services." *Medical Care Research and Review*, 61 (3), 37S-68S.

Cutler TW, Palmieri J, Khalsa M, Stebbins M.(2007) "Evaluation of the relationship between a chronic disease care management program and california pay-for-performance diabetes care cholesterol measures in one medical group" *J Manag Care Pharm.* 2007 Sep;13(7):578-88.

Donald S. and Lang K., (2007) "Inference with Difference-in-Differences and Other Panel Data," *Review of Economics and Statistics*, vol. 89(2), pages 221-233, 03.

Doran, T., C. Fullwood, et al. (2006). "Pay-for-performance programs in family practices in the United Kingdom." *The New England Journal of Medicine,* 355(4): 375-84.

Eisenberg JM. (1985) "Physician utilization: The state of research about physicians' practice patterns." *Med Care*. 23:461-83.

Eisenberg JM. (1986) "Doctors' Decisions and the Cost of Medical Care". Ann Arbor, MI: Health Administration Press

Ellis R. and McGuire T. (1990) "Optimal payment systems for health services" *Journal of Health Economics*, vol. 9, issue 4, pages 375-396

Epstein, A. M. (2006) "Paying for performance in the United States and abroad," *New England Journal of Medicine*, 355:406-8.

Fairbrother, G., K. L. Hanson, S. Friedman, and G. C. Butts. (1999), "The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates", *American Journal of Public Health* 89 (2): 171-175.

Fairbrother G, Siegel MJ, Friedman S, Kory PD, Butts GC. (2001) "Impact of financial incentives on documented immunization rates in the inner city: results of a randomized controlled trial", *Ambulatory Pediatrics* 2001 Jul-Aug; 1(4):206-12.

Felt-Lisk S, Gimm G and Peterson S. (2007) "Making Pay-For-Performance Work In Medicaid" *Health Affairs*, 26, no. 4 : w516-w527

Frolich, A., Talavera, J. A., Broadhead, P. and Dudley, R. A. (2007) "A behavioral model of clinician responses to incentives to improve quality." *Health Policy*, 80 (1), 179-193.

Gavagan T, Du H, Saver B, Adams G, Graham D, McCray R, and Goodrick K (2010) "Effect of Financial Incentives on Improvement in Medical Quality Indicators for Primary Care" *J Am Board Fam Med*, September-October 2010; 23: 622 - 631.

Gosden, T., F. Forland, et al. (2000). "Capitation, salary, fee-for-service and mixed systems of payment: effects on the behaviour of primary care physicians (Cochrane Review)." *Cochrane Database of Systematic Reviews*(3).

Gosden, T., F. Forland, et al. (2001). "Impact of payment method on behaviour of primary care physicians: a systematic review." *Journal of Health Services Research & Policy* 6(1): 44-55.

Grady KE, Lemkau JP, Lee NR, Caddell C. (1997), "Enhancing mammography referral in primary care", *Preventive Medicine,* 1997 Nov-Dec; 26(6):791-800.

Guido W. Imbens, (2004) "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, MIT Press, vol. 86(1), pages 4-29, 06.

Heckman, J., Ichimura, H., and Todd, P., (1997) "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, Blackwell Publishing, vol. 64(4), pages 605-54, October.

Hickson, G., Altemeier, W., and Perrin J., (1987) "Physician Reimbursement by Salary or Fee-for-Service: Effect on Physician Practice Behavior in a Randomized Prospective Study," *Pediatrics* Vol. 80 No. 3, pp. 344-350

Hillman, AL, Ripley, K, Goldfarb, N, Nuamah, I, Weiner, J, Lusk, E. (1998) "Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care." *American Journal of Public Health*, *88* 1699-1701.

Hillman, A. L., K. Ripley, N. Goldfarb, J. Weiner, I. Nuamah, and E Lusk. (1999), "The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care", *Pediatrics* 104 (4): 931-935.

Hurley, J., R. Labelle and T. Rice (1990), "The relationship between physician fees and the utilization of medical services in Ontario", *Advances in Health Economics and Health Services Research* 11:49-78

Hurley, J., DeCicca,P., Li J. and Buckley G. (2011), "The response of Ontario primary care physicians to pay-for-performance incentives." Final Report (Janurary 26,2011) submitted to the Ontario Ministry of Health and Long-Term Care (MOHLTC)

Hutchison, B. (2008). "Pay for Performance in Primary Care: Proceed with Caution, Pitfalls Ahead (Editorial)." *Healthcare Policy*, 4(1): 5.

Hutchison, B., S. Birch, et al. (1996). "Do physician-payment mechanisms affect hospital utilization? A study of Health Service Organizations in Ontario." *CMAJ*, 154(5): 653-661.

Hutchison, B., J. Hurley, et al. (1997). "Defining the Practice Population in Fee-for-Service Practice." *Health Services Research*, 32 (1): 55-70.

Kouides, RW, Bennettm, NM, Lewis, B, Cappuccio, JD, Barker, WH, LaForce, FM. (1998) "Performance-based physician reimbursement and influenza immunization rates in the elderly.The primary-care physicians of Monroe County." *American Journal of Preventive Medicine*, (14), 89-95.

Krasnik, A., P.P. Groenewegen , P.A. Pedersen, P. von Scholten , G. Mooney, A. Gottschau, H.A. Flierman and M.T. Damsgaard (1990), "Changing remuneration system: effects on activity in general practice", *British Medical Journal*, 300 (6741), 1698-701

Lawrence C. A., James H. B., Steven S. F., Nina L. A., Colleen M. K., Bruce A. C., William S. N., Mark E. L., Jasjit S. A., Marc W. M.,(2008) "A Randomized Trial of a Pay-for-Performance Program Targeting Clinician Referral to a State Tobacco Quitline", *Arch Intern Med.* 2008;168(18):1993-1999.

Lester H, Schmittdiel J, Selby J, Fireman B, Campbell S, Lee J, Whippy A, Madvig P. (2010) "The impact of removing financial incentives from clinical quality indicators: longitudinal analysis of four Kaiser Permanente indicators" *BMJ*; 340:c1898

Levin-Scherz J, DeVita N and Timbie J. (2006) "Impact of Pay-for-Performance Contracts and Network Registry on Diabetes and Asthma HEDIS Measures in an Integrated Delivery Network" *Medical Care Research and Review,* 2006 February vol. 63 no. 1 suppl 14S-28S

Lewis, S. (2009). "Pay for Performance: The Wrong Time, the Wrong Place?" *Healthcare Quarterly*,12(3):8-9.

Ma, Ching-to Albert & McGuire, T. (1997). "Optimal Health Insurance and Provider Payment," *American Economic Review*, vol. 87(4), pages 685-704

Mandel K, Kotagal U. (2007) "Pay for Performance Alone Cannot Drive Quality." *Arch Pediatr Adolesc Med.* ;161(7):650-655.

Maynard, A. (2008). “Payment for Performance (P4P): International experience and a cautionary proposal for Estonia”, World Health Organization Europe - Health Financing Policy Paper, Division of Country Health Systems: 30.

McGuire, T. G., and M.V. Pauly (1991), “Physician response to fee changes with multiple payers”, Journal of Health Economics 385-410

McGuire, Thomas G., (2000) “Physician agency”, Handbook of Health Economics, in: A. J. Culyer & J. P. Newhouse (ed.), Handbook of Health Economics, edition 1, volume 1, chapter 9, pages 461-536 Elsevier.

Meyer, B., (1995) “Natural and quasi-experiments in economics” *Journal of Business and Economic Statistics*, vol. 13, 2 (April) pp. 151-161Millett C, Gray J, Saxena S, Netuveli G and Majeed A, (2007) “Impact of a pay-for-performance incentive on support for smoking cessation and on smoking prevalence among people with diabetes” *CMAJ*, June 5, 2007; 176 (12).page 1705-1710

Moulton, Brent R., (1986)"Random group effects and the precision of regression estimates," *Journal of Econometrics*, vol. 32(3), pages 385-397, August.

Moulton, Brent R, (1990) "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit," *Review of Economics and Statistics*, vol. 72(2), pages 334-38, May.

Mullen K. J., Frank R.G., Rosenthal M. B. (2010) “Can You Get What You Pay For? Pay-For-Performance and the Quality of Healthcare Providers”, *RAND Journal of Economics*, 41: 64–91

Newhouse, J. (1996). "Reimbursing Health Plans and Health Providers: Efficiency in Production versus Selection," *Journal of Economic Literature*, vol. 34(3), pages 1236-1263

Nguyen, N.X., and F.W. Derrick (1997), “Physicians behavioral response to a Medicare price reduction”, Health Services Research 32: 283-298

Pearson S, Schneider E, Kleinman K, Coltin K and Singer J. (2008) "The Impact Of Pay-For-Performance On Health Care Quality In Massachusetts, 2001–2003" *Health Affairs*, 27, no. 4: 1167-1176

Peterson L., Woodard L., Urech T., Daw C., Sookanan S. (2006) “Does pay-for-performance improve the quality of health care”. *Annals of Internal Medicine*, 145(4):265-272

Predengast C. (1999) “The provision of incentives in firms” *Journal of Economic Literature*, vol. 37, issue 1, pages 7-63

Richards, J. (2009). "Is there an elephant in the room?" *British Journal of General Practice,* 59: 376-377.

Robinson, J. C. (2001), “Theory and practice in the design of physician payment incentives”, Milbank Quarterly, 79(2), 149-77

Rockville, M.D. (2004), “Strategies to Support Quality-based Purchasing: A Review of the Evidence”, Agency for Healthcare Research and Quality 2004:04-P024. U.S. Department of Health and Human Services. Public Health Service.

Rosenthal, M. B. and A. Dudley (2007). “Pay-for-Performance: will the latest payment trend improve care?” *Journal of the American Medical Association*, 297 (7): 740-44.

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

Rosenthal MB, Frank RG, Li Z, Epstein AM. (2005) “Early experience with pay-for-performance: from concept to practice.” *Journal of the American Medical Association*, 294 (14), 1788-93.

Rosenthal MB, Landon BE, Normand SL, Frank RG, Epstein AM. (2006) “Pay for performance in commercial HMOs.” *The New England Journal of Medicin*e, 355: 1895-1902

Roski, J., Jeddeloh R., An, L., Lando, H., Hannan, P., Hall, C. and Zhu, SH. (2003) "The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines." *Preventive Medicine,* Volume 36, Issue 3: 291-299

Scott A. (2001) “Eliciting GP’s preferences for pecuniary and non-pecuniary job characteristics.” *Journal of Health Economics*. 20:329-47

Scott, A. and A. Shiell (1997), “Do fee descriptors influence treatment choices in general practice: a multi-level discrete choice model”, Journal of Health Economics 16: 323-342

Shen, Y. (2003) “Selecition incentives in a performance-based contracting system,” *Health Services Research,* 38 (2), 535-552

Steel N, Maisey S, Clark A, Fleetcroft R, Howe A. (2007)"Quality of clinical primary care and targeted incentive payments: an observational study." *British Journal of General Practice*, 2007 June; 57 (539). pp. 449-454.

Stiglitz, Joseph E, (1974). "Incentives and Risk Sharing in Sharecropping," *Review of Economic Studies*, Blackwell Publishing, vol. 41(2), pages 219-55, April

Town R., Wholey D.R., Kralewski J., Dowd B. (2004) “Assessing the influence of incentives on physicians and medical groups.” *Medical Care Research and Review*, 2004 Sep; 61(3 Suppl):80S-118S.

Town R., Kane R., Johnson P., Butler M. “Economic Incenives and physicians’ deliveary of preventive care: a systematic review”. *American Journal of Preventive Medicine*. 2005; 28(2):234-240

Vaghela P, Ashworth M, Schofield P, and Gulliford M. (2009) "Population Intermediate Outcomes of Diabetes under Pay-for-Performance Incentives in England from 2004 to 2008." *Diabetes Care*, March 2009 vol. 32 no. 3 427-429

Wagstaff, A., Moreno-Serra, R. (2009) “Europe and central Asia’s great post-communist social health insurance experiment: Aggregate impacts on health sector outcomes”. *Journal of Health Economics*. 28: 322-340

Wennberg JE, Barnes BA, Zubkoff M. (1982) “Professional uncertainty and the problem of supplier-induced demand.” *Social Science & Medicine*, 16:811-24

Wilson, R. (2006). “Primary Care Renewal in Ontario- Focus on Remuneration”, Presented at the Primary Care Forum, November 26-27, 2006. Airport Marriott, Toronto Ontario. http://toolkit.cfpc.ca/en/remuneration/appendix-2.php

Wooldridge, J.M., (2002) “Econometric analysis of cross section and panel data.” MIT Press, Cambridge.

Yip, W. (1998), “Physician responses to medical fee reductions: changes in the volume and intensity of supply of Coronary, Artery Bypass Graft (CABG) surgeries in the Medicare private sectors”, *Journal of Health Economics* 17: 675-700

Young G, Meterko M, Beckman H, Baker E, et al. (2007) "Effects of Paying Physicians Based on their Relative Performance for Quality", *Journal of General Internal Medicine* 2007; Volume 22, Number 6, 872-876

Figure 1. Groups of physicians with different timing in PCR participation and P4P incentive exposure

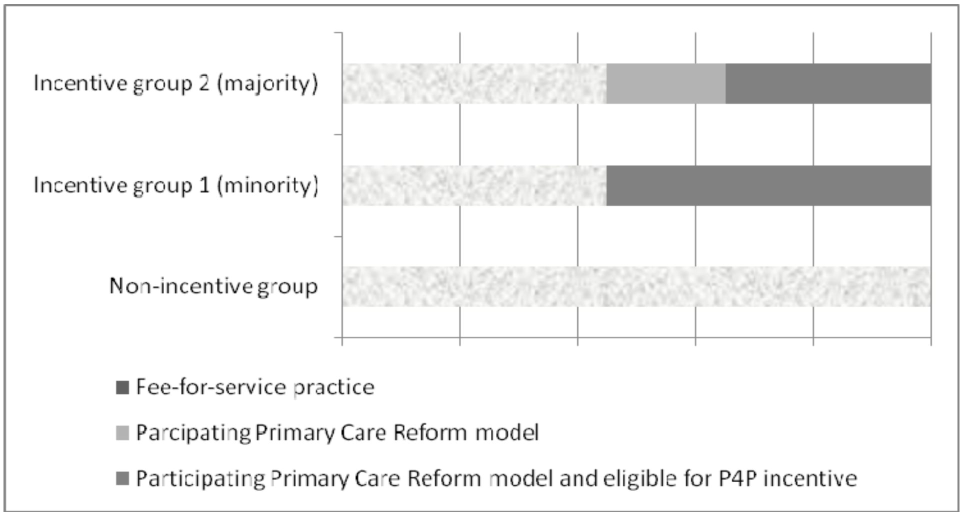

Figure 2: Mean Number of Claims for Contact Incentive Payments per Eligible Physician

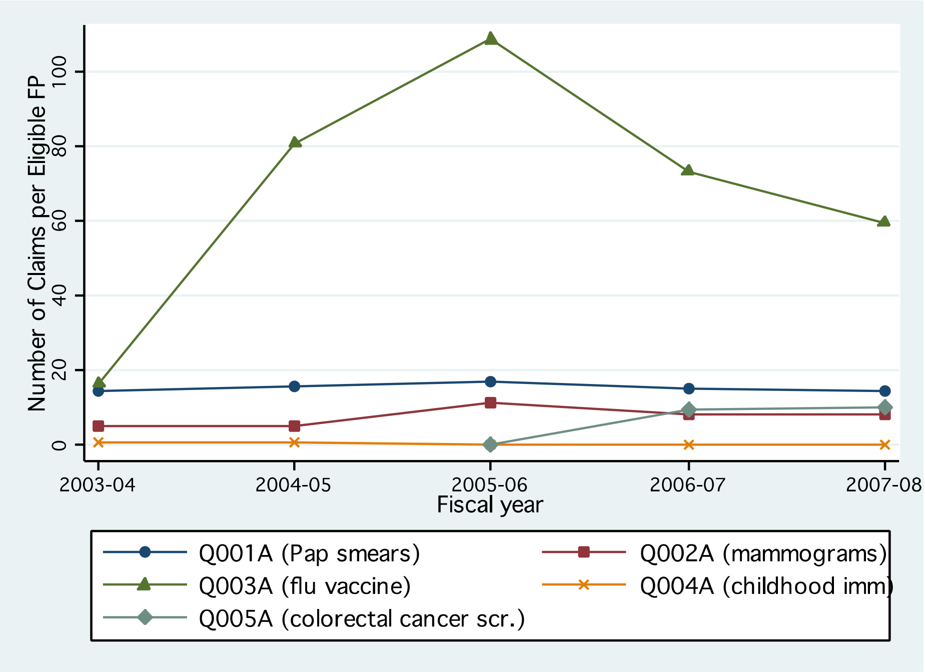

Figure 3: Share of Target Practice Population Receiving Targeted Service-- Senior Flu Shot

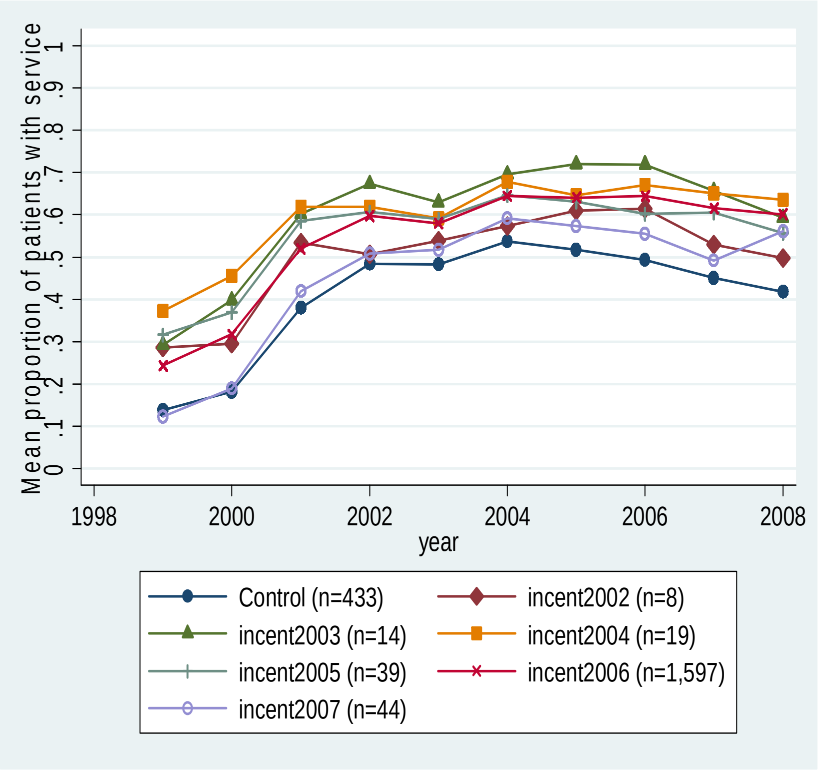

Figure 4: Share of Target Practice Population Receiving Targeted Service-- Toddler Immunzation

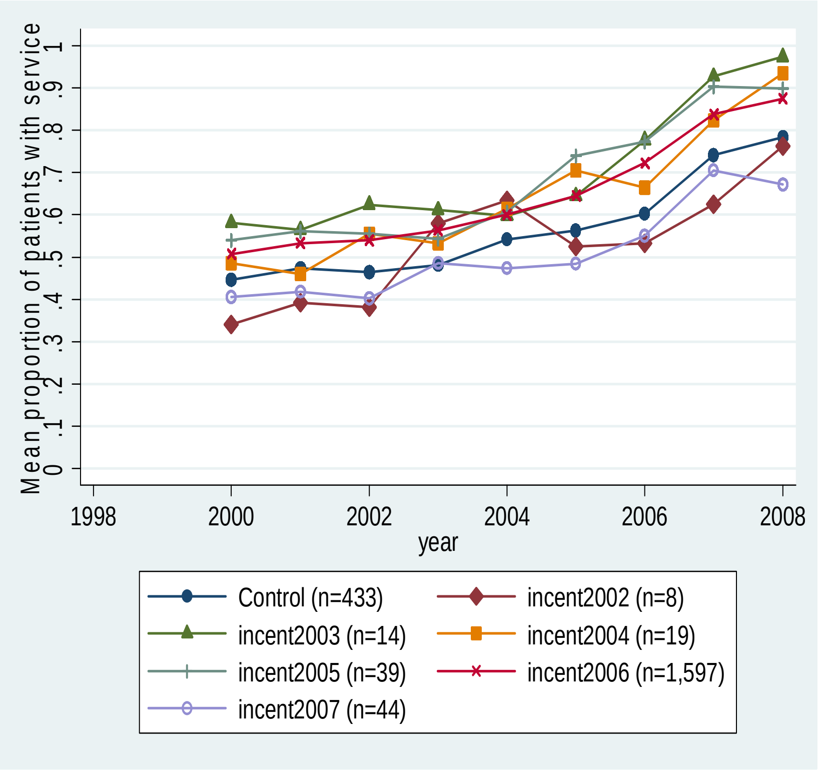

Figure 5: Share of Target Practice Population Receiving Targeted Service -- Pap Smear

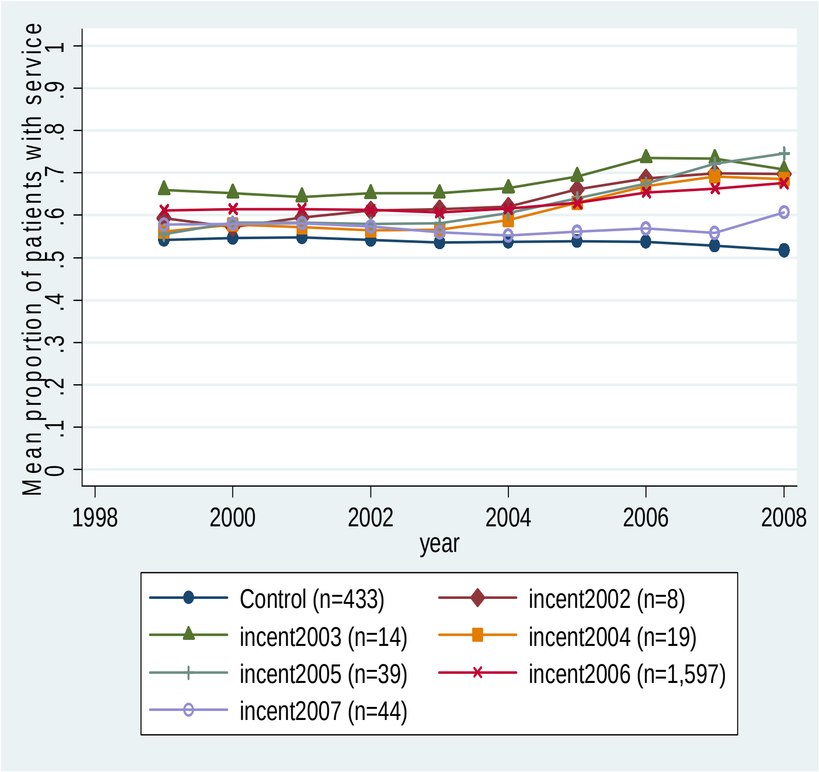

Figure 6: Share of Target Practice Population Receiving Targeted Service -- Mammogram

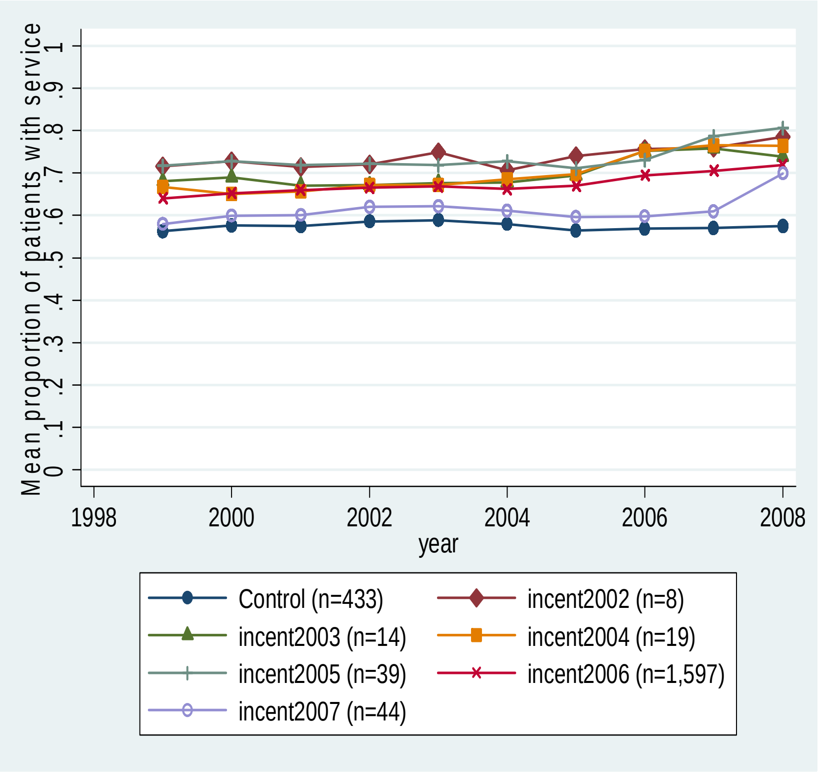

Figure 7: Share of Target Practice Population Receiving Targeted Service — Colorectal cancer screening

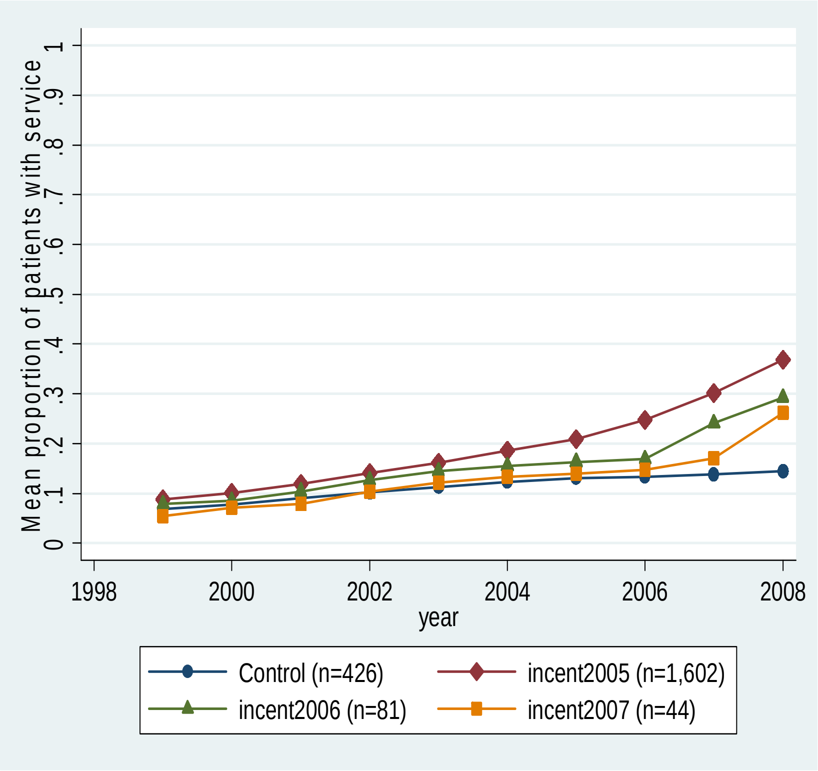

Figure 8: Proportion of Physicians Achieving the Targeted Performance Level of Service — Obstetrical deliveries

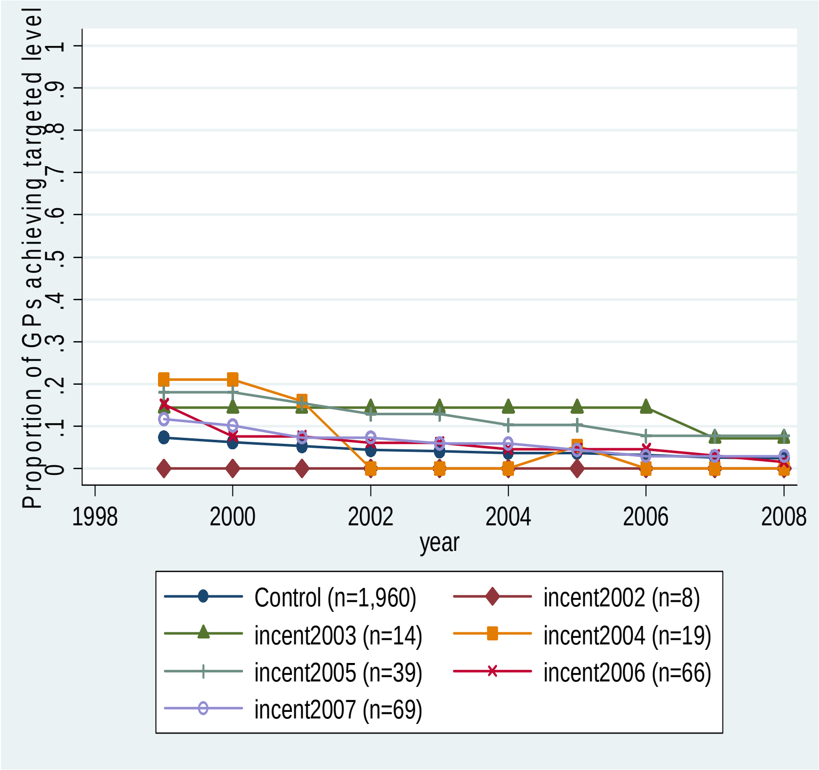

Figure 9: Proportion of Physicians Achieving the Targeted Performance Level of Service — Hospital services

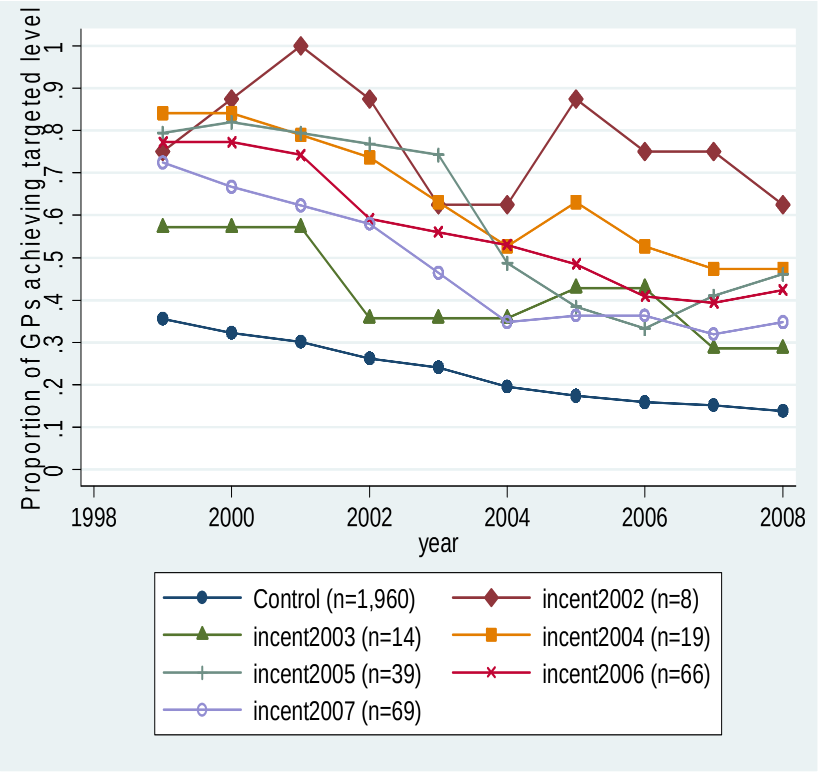

Figure 10: Proportion of Physicians Achieving the Targeted Performance Level of Service — Palliative care

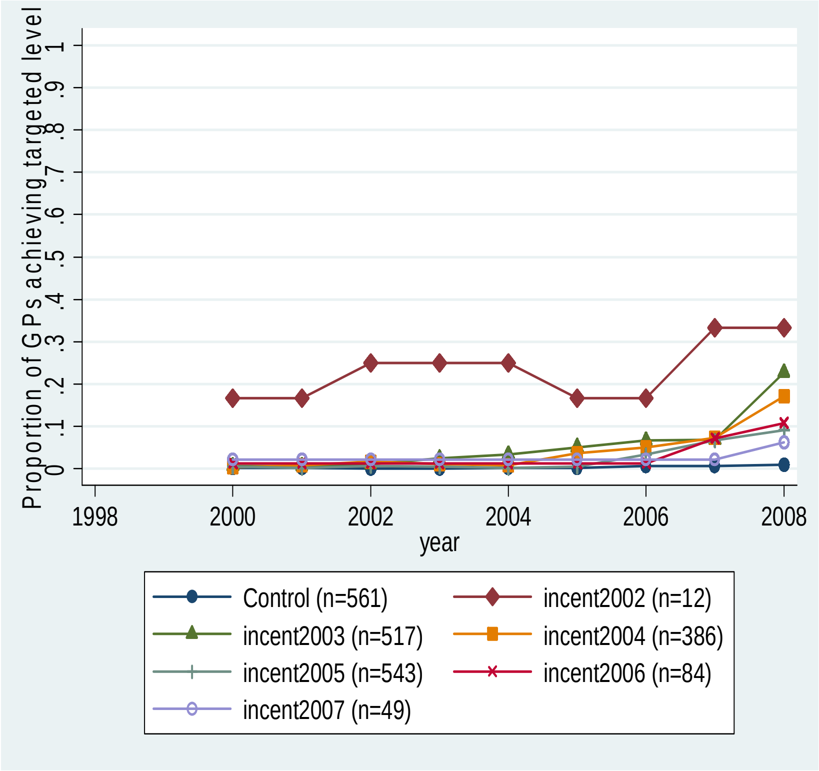

Figure 11: Proportion of Physicians Achieving the Targeted Performance Level of Service — Office procedures

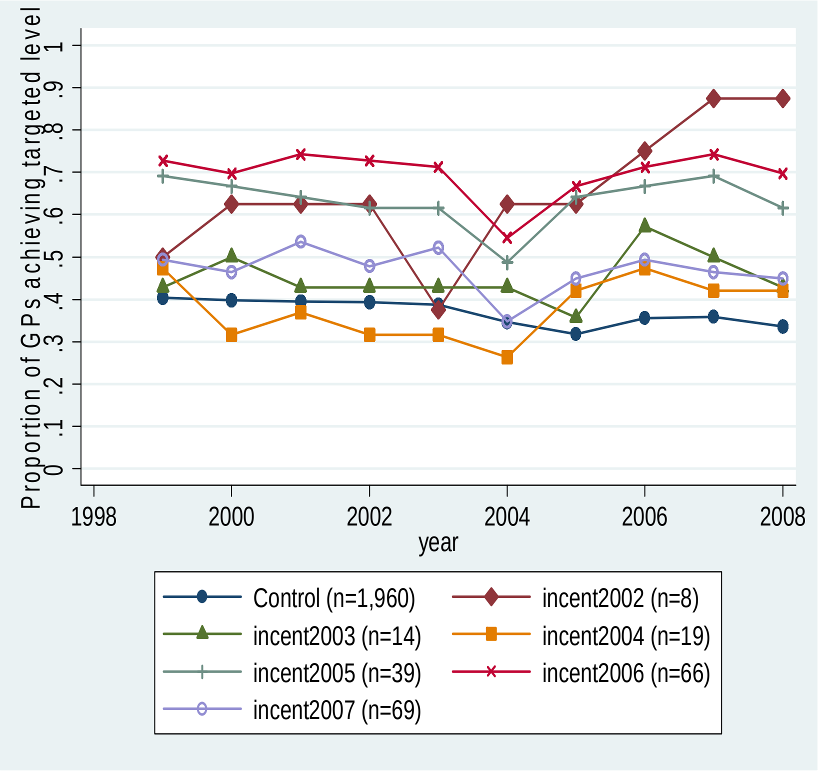

Figure 12: Proportion of Physicians Achieving the Targeted Performance Level of Service — Prenatal care

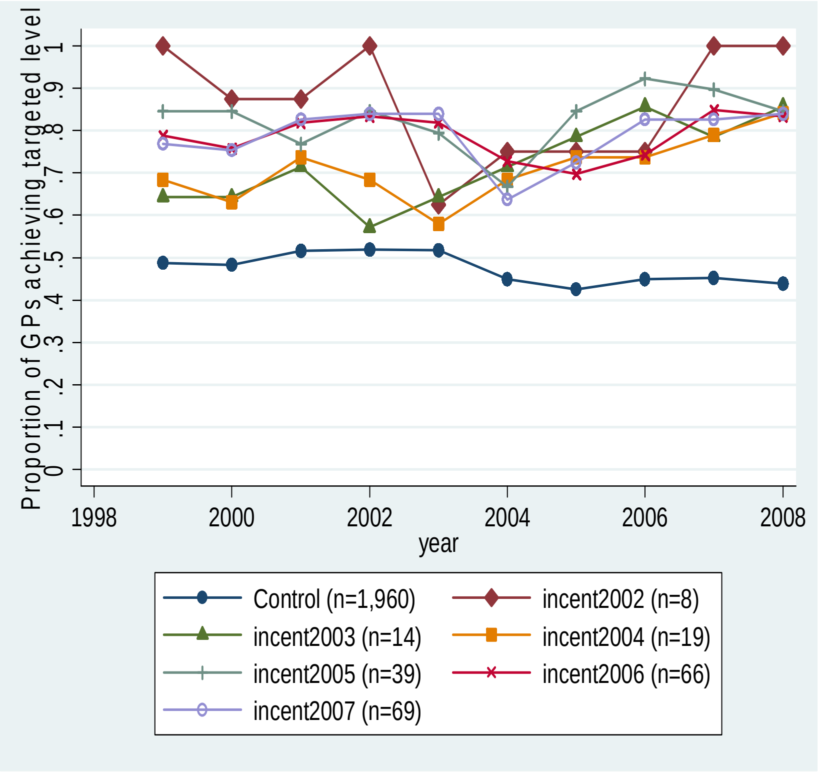

Figure 13: Proportion of Physicians Achieving the Targeted Performance Level of Service — Home visits

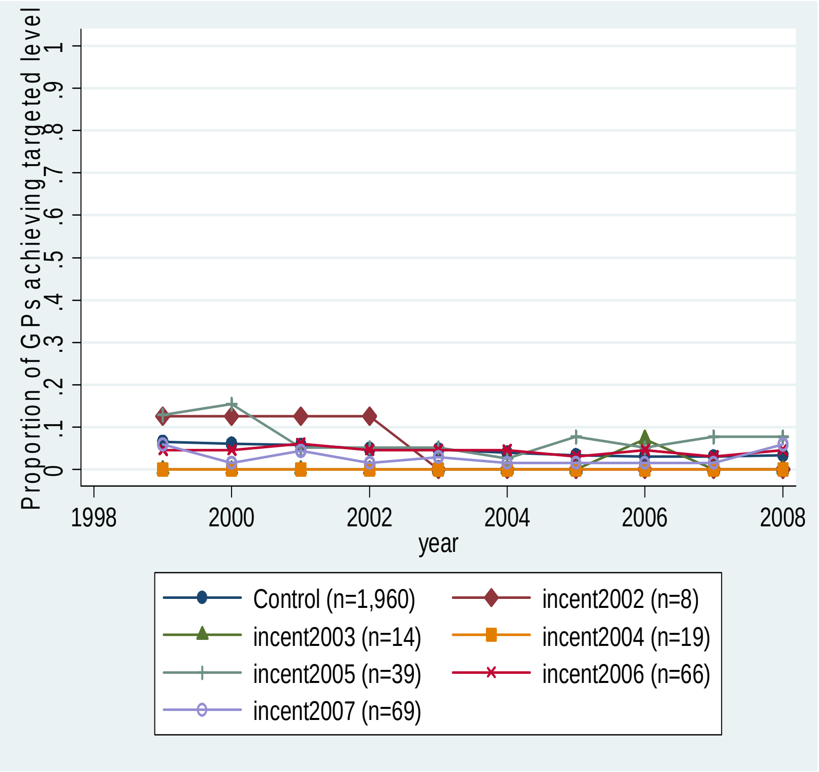

Table 1. Description of Eleven Financial Incentives under Analysis

#	Financial incentive	Eligibility condition	Bonus Payment
		Preventive Care Service Enhancement Payments	
Contact Payment			
-- payment of $6.86 for each documented contact for eligible patients to obtain the preventive service			
Cumulative Care Preventive Service Bonus			
1	Seniors' influenza immunizations	Bonus payment based on the proportion of the physician's eligible (aged 65 or more) and rostered patients on March 31 who received the flu shot in the previous flu season.	• $220 (60% of patients) • $440 (65% of patients) • $770 (70% of patients) • $1,100 (75% of patients) • $2,200 (80% of patients)
2	Pap smear	Bonus payment based on the proportion of the physician's eligible(females aged 35 to 69) and rostered patients on March 31 who received a pap smear for cervical cancer screening during the last 30 months.	• $220 (60% of patients) • $440 (65% of patients) • $660 (70% of patients) • $1,320 (75% of patients) • $2,200 (80% of patients)
3	Mammogram	Bonus payment based on the proportion of the physician's eligible (females aged 50 to 69) and rostered patients on March 31 who received a mammogram for breast cancer screening during the last 30 months.	• $220 (55% of patients) • $440 (60% of patients) • $770 (65% of patients) • $1,320 (70% of patients) • $2,200 (75% of patients)
4	Toddler immunizations	Bonus payment based on the proportion of the physician's eligible (children aged 30 to 42 months) and rostered patients on March 31 who received 5 immunizations by the age of 30 months.	• $440 (85% of patients) • $1,100 (90% of patients) • $2,200 (95% of patients)
5	Colorectal cancer screening	Bonus payment based on the proportion of the physician's eligible (aged 50 to 74) and rostered patients on March 31 who was administered a colorectal screening test by Fecal Occult Blood Testing during the last 30 months.	• $220 (15% of patients) • $440 (20% of patients) • $1,100 (40% of patients) • $2,200 (50% of patients)
		Annual Special Payments	
6	Obstetrical deliveries	Payment if 5 or more obstetrical services were delivered to 5 or more patients in a fiscal year.	$3,200 (increased to $5,000 in October 2007)
7	Hospital services	Payment if hospital services provided to all patients are totaled at least $2,000 in a fiscal year	$5,000 (increased to $7,500 in April 2005 for those with a Rurality Index of Ontario score greater than 45)
8	Palliative care	Payment if palliative care services are delivered to four or more patients in a fiscal year.	$2,000
9	Office procedures	Payment if office procedures provided to enrolled patients are totaled at least $1,200 in a fiscal year	$2,000
10	Prenatal care	Payment if prenatal care services are provided to five or more enrolled patients in a fiscal year	$2,000
11	Home visits	Payment if 100 or more home visits are provided to enrolled patients in a fiscal year.	$2,000

Table 2: Eligibility for Preventive Care Bonuses and Special Payments

	2002	2003	2004	2005	2006	2007	2008
Preventive Care Bonuses							
Senior Flu Immunization. Toddler Immunization, Pap Smear and Mammogram							
FHN	April						
FHG						April	
CCM						April	
FHO						April	
Colorectal Cancer Screening							
FHN					April		
FHG					April		
CCM					April		
FHO						April	
Special Payments							
Obstetrical Services, Hospital Services, Office Procedures, Prenatal Care and Home Visits							
FHN	April						
FHG				- never eligible -			
CCM				- never eligible -			
FHO					November		
Palliative Care							
FHN	April						
FHG		July					
CCM				- never eligible -			
FHO					November		

Note: Date PCR models introduced: FHN: April 2002; FHG: July 2003; CCM: October 2005; FHO: November 2006.

Table 3: Key Characteristics of Primary Care Models Included in This Study

Model (Year Introduced)	**Size/Rostering**	**Funding**
Traditional Fee-for-service (whole period)	No size regulation No rostering	• Fee-for-service
Family Health Network (2002)	*Physician:* • At least 3 GPs *Rostering:* • Minimum total roster of 2400 for group of 3 • Financial penalty for average rosters > 2400 per GP	Blended Capitation: • Age-sex adjusted capitation for rostered patients for 57 core services (about 80% of gross income) • 10% of FFS rate for core services to rostered patients • Access Bonus: 20.65% of the base capitation payment less value of outside use by rostered patients • Monthly comprehensive care capitation payments for formally rostered patients • 100% of FFS rate for core services to non-rostered patients up to $45,000 per physician • 100% of FFS rate for excluded services to either rostered or non-rostered patients
Family Health Group (2003)	*Physician:* • At least 3 GPs *Rostering:* • Voluntary	Blended fee-for-service: • 100% FFS as usual • 10% premium on FFS rate for specified comprehensive care services provided to rostered patients (Ministry-assigned and formally rostered) • Comprehensive care premium • Monthly comprehensive care capitation payments for formally rostered patients • Some premiums/ bonuses paid only for formally rostered patients
Comprehensive Care Model (2005)	*Physician:* • Solo practice *Rostering* • Required • No size regulation	Blended fee-for-service: • 100% FFS as usual • Monthly comprehensive care capitation for rostered patients
Family Health Organization (2006)	*Physician:* • At least 3 GPs *Rostering* • Required • No size regulation	Blended capitation: • Capitation for core services to enrolled patients • Access bonus: maximum of 18.59% of the base rate payment less outside use by rostered patients • 100% FFS for excluded services to all patients and for non-enrolled patients • Monthly comprehensive care capitation payments for formally rostered patients • 100% of FFS rate for core services to non-rostered patients up to $45,000 per physician • 100% of FFS rate for excluded services to either rostered or non-rostered patients

Note: Some of these elements were present for different periods for the different PCR model

Table 4. Criteria for Selecting the Sample of GPs for Analysis

Criterion	Rationale	Resulting Sample Size
All Ontario physicians present in claims data at any point between April 1998 and March 2008		37,422
Exclude physicians not present in all 10 years of the study period	Exclude physicians who interrupted their practice or who left province during the study period	- 21,415 = 16,007
Exclude physicians whose specialty is not general/family practice during entire study period[a]	Only GPs are eligible for incentives; we exclude those billing as GP while attaining specialization	- 8,533 = 7,474
Exclude physicians who billed less than $30,000 annually	Exclude part-time physicians	- 1,304 = 6,170
Exclude physicians without two consecutive years of practice before start of study (i.e., April 1996 to March 1998)	Exclude new GPs who are newly establishing their practice at the start of the study period	- 3,835 = 2,335
Exclude GPs for which billings for A001, A003, and A007 constitute less than 70% of activity, and GPs for which billings for A001, A003 and A007 constitute less than 50% of all activity and a single "non A-code" category constitutes over 15% of activity[c]	Exclude GP specialists whose main activity is other than providing traditional family medicine visits and consultations[b]	- 95 = 2,240
Exclude locums[d]	Locums are not eligible for the incentives[e]	- 19 = 2,221
Exclude GPs affiliated with the following primary care groups: RNPGA, HSO, PCN, SEAMO, GHC or ICHA[f]	Such GPs did not submit claims data or submitted only shadow billing claims; available data are insufficient for the analysis	- 32 = 2,189
Exclude GPs who converted between FFS and PCR practices for more than one time during the study period	Such GPs do not represent typical observations in service provision behavior	- 4 = 2,185

[a] A physician's specialty was defined as the specialty under which the largest share of services were billed (based on the fee approved).

[b] An additional criterion whereby a GP with more than 25% of billings for K-codes was classified as a GP psychotherapist was rendered superfluous by this 70% rule.

[c] The following ophthalmology codes are treated as a "non A" category: A009A, A110A, A111A, A112A, A114A, A115A, A237A, A238A, A239A, A240A, E077A.

[d] Locums were identified using information on the Group Type in the Corporate Provider Database. We were able to identify locums only imperfectly.

[e] GPs in walk-in-clinics do not regularly provide the services eligible for the financial incentives under study and should therefore be excluded from the analysis. We considered excluding GPs with a high proportion of billings for code A888 (Emergency Department Equivalent – Partial Assessment); however, once all of the above criteria were applied, this criterion was redundant.

[f] RNPGA: Rural and Northern Physician Group Agreement; HSO: Health Service Organization; PCN: Primary Care Network; SEAMO: Southeastern Ontario Medical Organization; GHC: Group Health Care; IHCA: Inner City Health Associates.

Table 5: Definitions and Sample Sizes of Control and Treatment Groups for each Performance Incentive

	Preventive Care Incentives	
Senior Flu Shot, Toddler Immunization, Pap Smear, Mammogram		
Control Group	• FFS	433 physicians
Treatment Group	• FHN starting from April 2002 • FHG starting from April 2007 • CCM starting from April 2007 • FHO starting from April 2007	1,722 physicians
Colorectal Cancer Screening		
Control Group	• FFS	427 physicians
Treatment Group	• FHN starting from April 2006 • FHG starting from April 2006 • CCM starting from April 2006 • FHO starting from April 2007	1,730 physicians
	Special Payments	
Obstetrical Care, Hospital Services, Office Procedures, Prenatal Care, and Home Visits		
Control Group	• FFS • FHG • CCM	1,962 physicians
Treatment Group	• FHN starting from April 2002 • FHO starting from November 2006	218 physicians
Palliative Care		
Control Group	• FFS • CCM	560 physicians
Treatment Group	• FHN starting from April 2002 • FHG starting from July 2003 • FHO starting from November 2006	1,596 physicians

Table 6. Descriptive Statistics in pre-intervention period: Control and Incentive Groups

	Control Group			Incentive Group			Equal Means	Equal Variance
	Mean	Median	St. Dev	Mean	Median	St. Dev	p-value	p-value
Preventive Care Incentives for Senior Flu Shot, Toddler Immunization, Pap Smear, Mammogram, Colorectal Cancer Screening and Special Payment for Palliative Care								
Physician Characteristics								
Age	54.0	54.0	10.8	49.0	49.0	8.4	0.000	0.000
Female	0.213	-	-	0.270	-	-	-	-
Years Licensed	22.1	20.0	11.6	18.1	16.0	8.8	0.000	0.000
Practice								
Size	1,408	1,345	668	1,605	1,572	573	0.000	0.000
Patient Age	39.7	39.3	8.5	38.6	38.0	6.2	0.000	0.000
Proportion Female	0.528	0.493	0.121	0.544	0.506	0.111	0.000	0.000
Proportion Infants	0.016	0.014	0.014	0.023	0.021	0.014	0.000	0.009
Proportion Elderly	0.144	0.119	0.111	0.135	0.119	0.079	0.000	0.000
Workload								
Annual Workdays	250.6	249.8	44.2	263.7	261.8	38.8	0.000	0.000
Annual Visits	7,663.0	7,337.8	3,531.0	8,466.2	8,307.5	3066.3	0.000	0.000
Visits/Workday	30.3	29.1	12.8	31.9	31.0	10.2	0.000	0.000
Special Payments for Obstetrical Care, Hospital Services, Office Procedures, Prenatal Care, and Home Visits								
Physician Characteristics								
Age	50.2	50.0	9.3	47.8	47.0	8.1	0.000	0.000
Female	0.255	-	-	0.284	-	-	-	-
Years Licensed	19.1	17.5	9.6	17.8	15.0	8.8	0.000	0.000
Practice								
Size	1,571	1,526	608	1,497	1,491	478	0.000	0.000
Patient Age	38.8	38.2	6.9	39.1	38.8	5.3	0.091	0.000
Proportion Female	0.538	0.500	0.114	0.565	0.528	0.102	0.000	0.000
Proportion Infants	0.021	0.019	0.014	0.027	0.025	0.013	0.000	0.000
Proportion Elderly	0.136	0.117	0.088	0.151	0.147	0.069	0.000	0.000
Workload								
Annual Workdays	260.8	260.3	40.4	264.2	261.8	38.9	0.000	0.018
Annual Visits	8,392	8,209	3,230	7,424	7,383	2,415	0.000	0.000
Visits/Workday	31.9	31.0	10.9	27.8	27.8	7.4	0.000	0.000

Notes: the null hypothesis of the t-tests on the equality of means is that the variable has the same mean for the treatment and control groups; the null hypothesis of the F-tests for the homogeneity of variances is that the variable has the same standard deviation for the treatment and control groups.

Table 7: Proportion of Eligible Family Physicians who Submitted at Least One Claim for a "Contact Incentive Payment"

	Pap Smear	**Mammogram**	**Senior Flu Vaccine**	**Toddler Immunization**	**Colorectal Cancer Screening**
2003-04	0.43	0.30	0.22	0.17	-
2004-05	0.28	0.19	0.62	0.13	-
2005-06	0.44	0.36	0.62	0.06	0.02
2006-07	0.37	0.27	0.47	0.06	0.18
2007-08	0.30	0.25	0.40	0.04	0.19

Note: Only physicians in FHNs were eligible from 2003-04 to 2005-06 for the contact incentive payments; FHN and FHO physicians were eligible for 2006-07 and 2007-08; FHGs and CCMs were never eligible during the study period; they became eligible April 1, 2008. FFS physicians were never eligible and remain ineligible.

Table 8: Main Results: Preventive Care Bonuses, Estimated Marginal Effects, Difference-in-Difference Models-- Full Sample

		DID with pooled OLS model			DID with physician-specific fixed effects model			DID with differential trend model			Specification test	
	(a)	(b)			(c)			(d)			(e)	
	Baseline Compliance in 2003	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Wald Test Statistics	P-value
Senior Flu Shot	0.554	0.013 (0.010)	19,866 (2,029)	0.371	0.028*** (0.007)	19,866 (2,029)	0.470	0.036*** (0.009)	19,866 (2,029)	0.469	2.71	0.100
Toddler Immunization	0.543	-0.007 (0.013)	16,826 (1,999)	0.278	0.011 (0.011)	16,826 (1,999)	0.356	0.004 (0.014)	16,826 (1,999)	0.356	0.66	0.417
Pap Smear	0.589	0.031*** (0.006)	19,926 (2,029)	0.433	0.041*** (0.004)	19,926 (2,029)	0.115	0.050*** (0.006)	19,926 (2,029)	0.115	12.17	0.001
Mammogram	0.646	0.004 (0.007)	19,888 (2,029)	0.351	0.018*** (0.005)	19,888 (2,029)	0.158	0.022*** (0.006)	19,888 (2,029)	0.158	1.44	0.230
Colorectal Cancer Screening	0.150	0.095*** (0.009)	19,918 (2,027)	0.217	0.085*** (0.005)	19,918 (2,027)	0.373	0.113*** (0.006)	19,918 (2,027)	0.379	66.30	0.000

*** Indicates statistical significance at the 1% level; ** indicates statistical significance at the 5% level; * indicates statistical significance at the 10% level.

Table 9: Robustness checks: Preventive Care Bonuses, Estimated Marginal Effects, Difference-in-Difference Estimator with Physician-specific Fixed Effects—Alternative estimation samples

	GPs who joined PCR before introduction of bonuses as treatment group				GPs in PCR models funded primarily by fee-for-service as treatment group			
	(a)	(b)			(c)	(d)		
	Baseline Compliance in 2003	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Baseline Compliance in 2003	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2
Senior Flu Shot	0.561	0.028*** (0.007)	19,073 (1,948)	0.468	0.554	0.024*** (0.007)	18,550 (1,893)	0.471
Toddler Immunization	0.548	0.010 (0.011)	16,162 (1,919)	0.356	0.543	0.010 (0.011)	15,669 (1,863)	0.352
Pap Smear	0.591	0.041*** (0.004)	19,130 (1,948)	0.117	0.589	0.040*** (0.004)	18,607 (1,893)	0.111
Mammogram	0.653	0.017*** (0.005)	19,093 (1,948)	0.152	0.646	0.018*** (0.005)	18,569 (1,893)	0.163
Colorectal Cancer Screening	0.144	0.079*** (0.006)	13,158 (1,341)	0.364	0.150	0.085*** (0.006)	17,778 (1,808)	0.355

*** Indicates statistical significance at the 1% level; ** indicates statistical significance at the 5% level; * indicates statistical significance at the 10% level.

Table 10: Sensitivity analysis 1: Preventive Care Bonuses, Estimated Marginal Effects, Difference-in-Difference Estimator with Physician-specific Fixed Effects—excluding Q-codes/exclusion codes

		Full Sample			GPs who joined PCR before introduction of bonuses			GPs in PCR models funded primarily by FFS		
	(a)	(b)			(c)			(d)		
	Baseline Compliance in 2003	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2
Senior Flu Shot	0.554	0.013* (0.007)	19,866 (2,029)	0.469	0.013* (0.007)	19,073 (1,948)	0.468	0.011 (0.007)	18,550 (1,893)	0.471
Toddler Immunization	0.543	0.008 (0.011)	16,826 (1,999)	0.352	0.007 (0.011)	16,162 (1,919)	0.352	0.007 (0.011)	15,669 (1,863)	0.349
Pap Smear	0.589	0.024*** (0.004)	19,926 (2,029)	0.084	0.024*** (0.004)	19,130 (1,948)	0.085	0.024*** (0.004)	18,607 (1,893)	0.084
Mammogram	0.653	0.017*** (0.005)	19,888 (2,029)	0.162	0.017*** (0.005)	19,093 (1,948)	0.156	0.017*** (0.005)	18,569 (1,893)	0.167
Colorectal Cancer Screening	0.150	0.068*** (0.005)	19,918 (2,027)	0.341	0.061*** (0.005)	13,158 (1,341)	0.334	0.067*** (0.006)	17,778 (1,808)	0.325

*** Indicates statistical significance at the 1% level; ** indicates statistical significance at the 5% level; * indicates statistical significance at the 10% level.

Table 11: Sensitivity analysis 2: Preventive Care Bonuses, Estimated Marginal Effects, Difference-in-Difference Estimator with Physician-specific Fixed Effects—adding in non-rostered patients

		Full Sample			GPs who joined PCR before introduction of bonuses			GPs in PCR models funded primarily by FFS		
	(a)	(b)			(c)			(d)		
	Baseline Compliance in 2003	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2
Senior Flu Shot	0.554	0.043*** (0.006)	20,207 (2,029)	0.466	0.044*** (0.007)	19,398 (1,948)	0.465	0.041*** (0.007)	18,847 (1,893)	0.468
Toddler Immunization	0.543	0.007 (0.010)	17,416 (1,999)	0.372	0.006 (0.010)	16,732 (1,919)	0.372	0.007 (0.010)	16,200 (1,863)	0.367
Pap Smear	0.589	0.042*** (0.004)	20,245 (2,029)	0.099	0.043*** (0.004)	19,435 (1,948)	0.102	0.041*** (0.004)	18,885 (1,893)	0.094
Mammogram	0.646	0.027*** (0.004)	20,228 (2,029)	0.150	0.028*** (0.004)	19,419 (1,948)	0.147	0.027*** (0.004)	18,868 (1,893)	0.153
Colorectal Cancer Screening	0.150	0.089*** (0.005)	20,217 (2,027)	0.365	0.085*** (0.005)	13,364 (1,341)	0.365	0.089*** (0.005)	18,029 (1,808)	0.344

*** Indicates statistical significance at the 1% level; ** indicates statistical significance at the 5% level; * indicates statistical significance at the 10% level.

Table 12: Sensitivity analysis 3: Preventive Care Bonuses, Estimated Marginal Effects, Difference-in-Difference Estimator with Physician-specific Fixed Effects—only FHGs and CCMs achived minimum roster size

		Full Sample			GPs who joined PCR before introduction of bonuses			GPs in PCR models funded primarily by FFS		
	(a)	(b)			(c)			(d)		
	Baseline Compliance in 2003	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2
Senior Flu Shot	0.560	0.031*** (0.007)	18,156 (1,843)	0.469	0.030*** (0.007)	17,706 (1,798)	0.469	0.028*** (0.007)	16,956 (1,720)	0.471
Toddler Immunization	0.548	0.010 (0.011)	15,480 (1,814)	0.370	0.011 (0.011)	15,094 (1,770)	0.370	0.009 (0.011)	14,423 (1,691)	0.366
Pap Smear	0.591	0.043*** (0.004)	18,212 (1,843)	0.120	0.043*** (0.004)	17,762 (1,798)	0.120	0.043*** (0.004)	17,010 (1,720)	0.116
Mammogram	0.649	0.029*** (0.005)	18,179 (1,843)	0.183	0.028*** (0.005)	17,729 (1,798)	0.181	0.029*** (0.005)	16,977 (1,720)	0.188
Colorectal Cancer Screening	0.154	0.091*** (0.006)	17,877 (1,806)	0.380	0.083*** (0.006)	12,443 (1,262)	0.367	0.090*** (0.006)	15,918 (1,607)	0.363

*** Indicates statistical significance at the 1% level; ** indicates statistical significance at the 5% level; * indicates statistical significance at the 10% level.

Table 13: Sensitivity analysis 4: Preventive Care Bonuses, Estimated Marginal Effects, Difference-in-Difference Estimator with Physician-specific Fixed Effects—dropping transition year/first year of incentive exposure

		Full Sample			GPs who joined PCR before introduction of bonuses			GPs in PCR models funded primarily by FFS		
	(a)	(b)			(c)			(d)		
	Baseline Compliance in 2003	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2	Marginal Effect (St. Error)	Sample size: # obs (# GPs)	R^2
Senior Flu Shot	0.554	0.037*** (0.009)	18,329 (2,029)	0.475	0.036*** (0.009)	17,607 (1,948)	0.473	0.033*** (0.009)	17,136 (1,893)	0.476
Toddler Immunization	0.543	0.002 (0.015)	15,365 (1,999)	0.318	0.001 (0.015)	14,760 (1,919)	0.318	0.001 (0.015)	14,328 (1,863)	0.316
Pap Smear	0.589	0.049*** (0.006)	18,389 (2,029)	0.106	0.050*** (0.006)	17,665 (1,948)	0.107	0.050*** (0.006)	17,193 (1,893)	0.102
Mammogram	0.646	0.035*** (0.006)	18,352 (2,029)	0.173	0.035*** (0.006)	17,628 (1,948)	0.169	0.034*** (0.006)	17,156 (1,893)	0.177
Colorectal Cancer Screening	0.150	0.111*** (0.006)	18,381 (2,027)	0.390	0.105*** (0.007)	12,259 (1,341)	0.377	0.112*** (0.007)	16,457 (1,808)	0.372

*** Indicates statistical significance at the 1% level; ** indicates statistical significance at the 5% level; * indicates statistical significance at the 10% level.

Table 14: Main Results: Special Payments, Estimated Marginal Effects, Difference-in-Difference Estimator (no physician-specific fixed effects)

		GPs Eligible for Special Payments in 2005			GPs Eligible for Special Payments in 2006			GPs Eligible for Special Payments in 2007		
	(a) Baseline Compliance in 2003	Marginal Effect (St. Error)	(b) Sample size: # obs (# GPs)	Pseudo R^2	Marginal Effect (St. Error)	(c) Sample size: # obs (# GPs)	Pseudo R^2	Marginal Effect (St. Error)	(d) Sample size: # obs (# GPs)	Pseudo R^2
Obstetrical Services	0.043	-0.0004 (0.005)	19,934 (1,998)	0.302	-0.004 (0.004)	20,187 (2,025)	0.308	0.013 (0.024)	20,196 (2,028)	0.302
Hospital Services	0.272	-0.013 (0.035)	19,777 (1,985)	0.481	-0.005 (0.074)	20,052 (2,012)	0.482	-0.019 (0.037)	20,138 (2,021)	0.482
Office Procedures	0.405	0.006 (0.064)	19,897 (1,995)	0.167	0.075 (0.127)	20,175 (2,022)	0.171	-0.141*** (0.053)	20,209 (2,026)	0.165
Prenatal Care	0.544	0.314*** (0.107)	19,857 (1,991)	0.295	0.106 (0.070)	20,109 (2,016)	0.295	0.184 (0.127)	20,151 (2,020)	0.294
Home Visits	0.045	0.007 (0.007)	18,814 (1,893)	0.225	0.003 (0.012)	19,251 (1,934)	0.230	0.084 (0.078)	19,557 (1,961)	0.226
		GPs Eligible for Special Payments in 2003			GPs Eligible for Special Payments in 2004			GPs Eligible for Special Payments in 2005		
Palliative Care	0.011	0.009 (0.012)	9,681 (1,078)	0.305	0.004 (0.005)	8,495 (946)	0.347	0.032 (0.031)	9,928 (1,104)	0.301

*** Indicates statistical significance at the 1% level.

Table 15: Estimated Marginal Effects: Preventive Care Bonuses by Physician Age, Practice Size, and Baseline Level of Compliance, Difference-in-Difference Estimator with Physician-specific Fixed Effects

	By Age					Practice Size					Baseline Compliance				
	GP Age	Baseline in 2003	Marg. eff. (St Err)	# obs. (# GPs)	R^2	Size	Baseline in 2003	Marg. eff. (St Err)	# obs.(#	R^2	Quartile	Baseline in 2003	Marg. eff. (St Err)	# obs. (# GPs)	R^2
Senior Flu Shot	< 40	0.549	0.024 (0.022)	2,320 (237)	0.48	< 1K	0.547	0.025 (0.016)	3,270 (336)	0.44	Q1 (Lowest)	0.304	0.036*** (0.014)	4,887 (503)	0.42
	40-55	0.577	0.049*** (0.010)	10,508 (1,073)	0.47	1K-1.5K	0.572	0.017 (0.012)	5,935 (606)	0.46	Q2 and Q3	0.593	0.027*** (0.010)	9,908 (1,009)	0.51
	> 55	0.537	-0.001 (0.010)	7,038 (719)	0.47	> 1.5K	0.557	0.031*** (0.010)	10,661 (1,087)	0.49	Q4 (Highest)	0.747	0.020 (0.015)	5,071 (517)	0.49
Toddler Immunization	< 40	0.527	0.027 (0.026)	2,044 (234)	0.56	< 1K	0.496	0.014 (0.034)	2,402 (313)	0.28	Q1 (Lowest)	0.217	0.026 (0.025)	3,810 (455)	0.40
	40-55	0.577	-0.010 (0.015)	9,095 (1,065)	0.39	1K-1.5K	0.560	0.004 (0.021)	5,042 (601)	0.32	Q2 and Q3	0.558	0.012 (0.014)	8,795 (1,023)	0.44
	> 55	0.503	0.037* (0.020)	5,687 (700)	0.26	> 1.5K	0.552	0.021 (0.014)	9,382 (1,085)	0.43	Q4 (Highest)	0.838	0.015 (0.024)	4,221 (521)	0.30
Pap Smear	< 40	0.620	0.059*** (0.013)	2,334 (237)	0.20	< 1K	0.630	0.040*** (0.011)	3,289 (335)	0.10	Q1 (Lowest)	0.374	0.024** (0.010)	4,929 (503)	0.18
	40-55	0.612	0.053*** (0.006)	10,503 (1,069)	0.14	1K-1.5K	0.617	0.038*** (0.007)	5,926 (604)	0.12	Q2 and Q3	0.594	0.050*** (0.006)	9,904 (1,009)	0.13
	> 55	0.549	0.019*** (0.007)	7,029 (717)	0.09	> 1.5K	0.564	0.037*** (0.006)	10,651 (1,084)	0.14	Q4 (Highest)	0.800	0.038*** (0.010)	5,033 (511)	0.14
Mammogram	< 40	0.653	0.045*** (0.016)	2,322 (237)	0.26	< 1K	0.682	0.017 (0.011)	3,284 (336)	0.11	Q1 (Lowest)	0.438	0.034*** (0.009)	5,031 (515)	0.25
	40-55	0.671	0.016** (0.007)	10,532 (1,073)	0.16	1K-1.5K	0.673	-0.003 (0.008)	5,937 (606)	0.20	Q2 and Q3	0.672	0.018*** (0.006)	9,873 (1,007)	0.16
	> 55	0.625	0.014*** (0.007)	7,034 (719)	0.16	> 1.5K	0.632	0.028*** (0.006)	10,667 (1,087)	0.19	Q4 (Highest)	0.829	0.014 (0.010)	4,984 (507)	0.17
Colorectal Cancer Screening	< 40	0.187	0.147*** (0.019)	2,197 (224)	0.48	< 1K	0.159	0.068*** (0.015)	3,243 (330)	0.37	Q1 (Lowest)	0.015	0.091*** (0.007)	4,918 (502)	0.41
	40-55	0.164	0.081*** (0.008)	10,744 (1,092)	0.39	1K-1.5K	0.145	0.079*** (0.010)	6,059 (617)	0.38	Q2 and Q3	0.078	0.102*** (0.007)	9,938 (1,012)	0.45
	> 55	0.116	0.067*** (0.008)	6,977 (711)	0.30	> 1.5K	0.150	0.090*** (0.007)	10,616 (1,080)	0.39	Q4 (Highest)	0.428	0.071*** (0.015)	5,062 (513)	0.43

*** Indicates statistical significance at the 1% level; ** indicates statistical significance at the 5% level; and * indicates statistical significance at the 10% level.

Appendix 1-- Empirical Studies on Physician Response to P4P incentives

Study (Authors)	Study design	Incentives involved (Form of incentives, targeted services)	Incentive level	Results	Context	Intervention duration	Sample size / scale of the experiment
Grady et al. 1997	RCT, random assignment 3 arms: 20 education and reward; 18 education; 23 control (61 practices in total)	Reward with education Mammography referrals	Physician	No effect	U.S. 61 primary care practices in greater Dayton, Ohio and Springfield, Massachusetts	6 months	95 physicians in total
Kouides et al. 1998	RCT, non-random assignment 2 arms: 27 practices in treatment, 27 practices in control	Bonus Influenza immunization rates	Provider group	Positive effect in immunization rate in a Medicare population	U.S. For Medicare population	4 months	62 physician in treatment, 82 in control
Hillman et al. 1998	RCT, random assignment 2 arms: 26 PC sites intervention; 26 PC sites control	Bonus(based on the part of capitation payment)+ feedback regarding compliance with guidelines; Cancer screening guidelines: mammography, breast cancer, pap smear, colorectal screening	Provider group	No effect in compliance scores	U.S. Medicaid HMO (contract with numerous other health plans)	18 months period	52 primary care practices, relatively small sample
Hillman et al. 1999	RCT, random assignment 3 arms: control, feedback of performance only, feedback+ bonus payment	Bonus(based on the part of capitation payment)+ feedback Pediatric immunization	Provider group	No effect in compliance scores	U.S. Medicaid HMO (contract with numerous other health plans)	18 months period	53 pediatric practices, relatively small sample

Fairbrother et al. 1999	RCT, random assignment 4 arms: 15 doctors in control, 15 feedback, 15 feedback+ bonus, 15 enhanced FFS+ bonus	Bonus with performance feedback Childhood immunization rate	Physician level	Partial effect: only feedback+bonus improved childhood immunization rate, but primarily achieved through better documentation	U.S. A low-income urban population	12 months	60 physicians in total
Fairbrother et al. 2001	RCT, random assignment 3 arms: 24 bonus, 12 FFS, 21 control	Bonus Pediatric immunizations	Physician level	Partial effect: significant increase in coverage levels, but the increase is primarily due to better documentation not to better immunizing practices	U.S. A low-income urban population	16 months	57 physicians in total
Roski et al. 2003	RCT, random assignment 3 arms: 15 clinics control; 15 bonus; 10 bonus + computerized patient registry	Bonus Smoking cessation	Provider group	Partial effect: improved in adherence to guidelines (documentation of smoking status and providing advice to quit), but no effect in quitting rate	U.S.	12 months	40 clinics in total
Amundson et al. 2003	Observational Before-after analysis No control group	Bonus + performance feedback Tobacco cessation	Provider group	Positive effect improve physician compliance with the tobacco treatment guideline	U.S. HealthParterners system in Minneapolis	3 years	20 medical groups

Beaulieu and Horrigan 2005	Observational Before-after analysis with control group Treatment group: 21 primary care doctors contracted with Independent Health in upstate New York; Control group: provider groups in the Pacific Northwest	Performance bonus + offered with diabetic registry and group discussion process Process and outcome measures for diabetic care	Physician level	Partial effect: patients treated by the doctors contracted with the program had improvement on 7 out of 9 measures	U.S. New York.	8 months	21 physicians as treatment group
Rosenthal et al. 2005	Observational Before-after analysis with control Treatment group: 163 provider groups contracted with PacifiCare Health systems in California; Control group: 42 provider groups contracted with PacifiCare in the Pacific Northwest	Bonus in Quality Incentive Program Process measures: cervical cancer screening, mammography, Haemoglobin A1c test	Provider group	Partial effect: improved only in cervical cancer screening, not improved in mammography, haemoglobin A1c test	U.S. A large network HMO, PacifiCare Health System introduced Quality Incentive Program to contracted medical groups in California in March 2002	10 months	163medical groups eligible for the bonus
Doran et al. 2006	Observational Cross-sectional regression, Not Before-after analysis	Performance contracting: performance w.r.t. 146 quality indicator	Physician level	Positive effect: high levels of reported achievement	U.K. Large scale, national level pay-for-performance contract in 2004	1 year	8,105 phyisicians

Levin-Scherz et al. 2006	Observational Retrospective cohort study using before-after trend comparison Treatment group: health plans participating in PCHI P4P contracts; Comparison group: national and Massachusetts State measures	P4P contracts: bonus based on network performance compared to previously agreed targets Adult diabetes and pediatric asthma HEDIS scores	Network level	Positive effect: improvement compared to state and national level	U.S. Beginning in 2001, a provider network Partners Community HealthCare, Inc (PCHI) and the health plans began P4P contracts with bonus payments.	2001-2003	18-75 health plans in PCHI as treatment group
Campbell et al. 2007	Observational Before-after analysis, No comparison group	Performance contracting: Clinical indicators on coronary heart disease, asthma, type 2 diabetes	Physician practices	Partial effect: improved in asthma, type 2 diabetes measures. Not improved in coronary heart disease measures	U.K. Large scale, national level pay-for-performance scheme for family practice in 2004	1998, 2003, 2005	42 family practices, national representative
Millett et al. 2007	Observational Before-after analysis, No comparison group	Performance contracting: smoking cessation among patients with diabetes Proportion of patients with documented smoking cessation advice, prevalence of smoking among patients with diabetes	Physician practices	Positive effect: increased the provision of support for smokers with diabetes in primary care settings	U.K. Large scale, national level pay-for-performance scheme for family practice in 2004	2003, 2005	36 primary care practices

Steel et al. 2007	Observational Before-after analysis, No comparison group	Performance contracting: quality of care for two common chronic conditions: asthma and hypertension Six quality indicators referred to asthma and hypertension subject to incentive payments	Physician practices	Positive effect: significant increase for the six indicators referred to asthma and hypertension linked to incentive payments	U.K. Large scale, national level pay-for-performance scheme for family practice in 2004	2003, 2005	18 primary care practices
Mandel and Kotagal 2007	Observational Before-after mean comparison, No control group	Pay for performance coupled with additional improvement interventions related to the collaborative. Flu shot percentage, controller medication percentage for children with persistent asthma, written self-management plan percentage.	Primary care practices	Positive effect: The initiative resulted in substantive and sustainable improvement in all measures	U.S. The Physician-Hospital Organization (PHO) affiliated with Cincinnati Children's Hospital Medical Center launched an asthma improvement collaborative in October 2003	2003-2006	44 pediatric practices

Felt-Lisk et al. 2007	Observational Before-after mean comparison with comparison group Treatment group: five Medicaid-focused health plans in California participating the LIRR Collaborative Comparison groups: national and state benchmarks, and two plans that were part of the collaborative but no incentives	Bonus payments for improving the measure A HEDIS measure for “well-baby visits” that requires six visits by age fifteen months	Health plans	Partial effect: only small effect in some of the plans	U.S A collaborative P4P effort among seven Medicaid-focused health plans in California during 2003–2005, known as the Local Initiative Rewarding Results (LIRR) Collaborative Demonstration.	2002 versus 2003–05	7 health plans
Young et al. 2007	Observational Retrospective cohort study using before-after trend comparison Treatment group: physicians participating in the program; Comparison group: national and New York State RIPA scores	Incentive program placing physicians at financial risk to receive rewards based on their performance relative to other physicians in the program. 4 diabetes performance measures	Individual physician	Partial effect: modest effect provider adherence to quality standards for a single measure of diabetes care	U.S. A pay-for-performance program of the Rochester (New York) Individual Practice Association (RIPA) between 2000 and 2001.	1999-2004	334 primary care physicians
Coleman et al. 2007	Observational Before-after regression analysis No comparison group	Performance-based compensation HbA1c testing for diabetes care in a low-income patient population	Physician level	Partial effects: dramatic improvements in rate of patients receiving recommended number of HbA1c tests; no effect on rate of physicians providing first HbA1c test, nor improvement in patient outcomes	U.S. A performance-based provider compensation program implemented in January 2004 at Access Community Health Network (ACCESS), a large system of federally qualified health centers (FQHCs) in Chicago.	2002-2004	46 physicians

Cutler et al. 2007	Observational Before-after mean comparison with comparison group; Treatment group: patients with diabetes who were followed by the P4P program (CDCM); Comparison group: patients followed by routine care group	P4P incentive payments Percent of eligible patients received an LDL-C test and attained LDL-C control	Medical groups	Positive effect: Higher rates of performance in the P4P program	U.S. Chronic disease care management (CDCM) program received by the Mercy Medical Group (MMG), which is a 160-provider, multispecialty medical group and has participated in the California P4P initiative	2003-2004	165 patients in the incentive program, 1,694 patients as control
Lawrence et al. 2008	RCT, random assignment 2 arms: 24 clinics with P4P payments, 25 usual care clinics	Bonus payment intervention based on tobacco quitline referrals; Rates of referrals	Clinics	Positive effect: increased referrals rates compared to usual care clinics	U.S. A P4P program targeting clinician referral to statewide quitline services in Minnesota.	September 2005 to June 2006	24 clinics as treatment, 25 clinics as control
Pearson et al. 2008	Observational Before-after mean comparison with comparison group Treatment group: physician groups that receive incentives Groups Comparison group: groups that were matched with incentivized groups on their baseline performance but that did not subsequently receive any incentive	Multiple P4P programs introduced into physician group contracts during 2001–2003 by the five major commercial health plans operating in Massachusetts; 13 Health Care Employer Data And Information Set (HEDIS) Measures	Physician groups	No significant effect: no distinguishable different trends among the treatment and the matched comparison group.	U.S. Multiple P4P programs introduced into physician group contracts during 2001–2003 by the five major commercial health plans operating in Massachusetts.	2001-2003	154 physician groups in total

Campbell et al. 2009	Observational Before-after interrupted time-series analysis, No control group	Performance contracting: Clinical indicators on coronary heart disease, asthma, type 2 diabetes	Physician practices	Partial effect: by 2005, improvement quality for asthma and diabetes but not for heart disease. By 2007, the rate of improvement had slowed for all three conditions.	U.K. Large scale, national level pay-for-performance scheme for family practice in 2004	1998, 2003, 2005, 2007	42 family practices, national representative
Vaghela et al. 2009	Observational Before-after analysis, No comparison group	Performance contracting 3 measures related to diabetes outcomes	Physician practices	Positive effect: significant increase	U.K. Large scale, national level pay-for-performance scheme for family practice in 2004	2004-2005, 2007-2008	Around 8,423 practices
Lee et al. 2010	Observational Before-after mean comparison with comparison group; Treatment group: patients with diabetes who were enrolled in the P4P program; Comparison group: randomly sampled patients with diabetes who had never joined the P4P program	Financial Incentives for increasing comprehensive follow-up visits for diabetes care Number of essential exams/tests; numbers of diabetes-related physician visits and hospital admissions	Physician level	Positive effect: P4P program for diabetes was associated with a significant increase in regular follow-up visits and evidence-based services, and significantly lower hospitalization costs.	Taiwan A pay-for-performance (P4P) program for diabetes care operated by the Bureau of National Health Insurance (NHI) in Taiwan.	2005-2006	12,499 patients as intervention group; 26,172 patients as comparison group
Gavagan et al. 2010	Observational Before-after mean comparison with comparison group; Treatment group: clinics received incentives; Comparison group: clinics with no incentives	Financial incentive for achieving group targets in preventive care; Cervical cancer screening, mammography, and pediatric immunization	Physician level	No significant effect: no significant effect on performance of preventive care	U.S. In 2002, 11 public community health centers in Houston/Harris County were provided performance incentives on 3 quality indicators in preventive care	2002	6 clinics as treatment group; 5 clinics as comparison group

Chung et al. 2010	Observational Before-after mean and trend comparison; No comparison group	Bonus payment to physicians: based on individual physicians' performance on 15 ambulatory quality measures, with a composite score	Physician level	No significant effect: no evident effect of physician-specific incentives	U.S. In 2007, all primary care physicians at Palo Alto Medical Clinic (PAMC), California participated in the physician incentive program.	2007	179 physicians
Boland et al. 2010	Observational Before-after mean comparison; No comparison group	Bonus payments were to be made if the radiologists met goals. Three radiologist report turnaround times (RTAT) Components	Provider individual level	Positive effect: significant decrease in turnaround time after the program	U.S. Massachusetts General (MGPO) Physicians Organization at the Massachusetts General Hospital (MGH) initiated a hospital wide department specific radiologist PFP initiative.	July 2006–March 2009	81 radiologists, 11 subspecialty divisions
Lester et al. 2010	Observational Before-after trend comparison; No comparison group	Financial incentive related to quality indicator; Screening for cervical cancer, control of hypertension, diabetes control, and screening for diabetic retinopathy	Medical facilities	Positive effect: upward trend when incentives were in place and downward trend when incentives were removed.	U.S. (and U.K.) Four of original financial incentives removed for 35 outpatient facilities owned and operated by Kaiser Permanente Northern California.	1999-2007	35 outpatient medical facilities

Mullen et al. 2010	Observational Before-after regression analysis with control (DID) Treatment group: provider groups contracted with PacifiCare Health systems in California; Control group: provider groups in the Pacific Northwest	Bonus in Quality Incentive Program, Another annual bonus program by the Integrated Healthcare Association (IHA), bonus based on cervical cancer screening, mammography, Haemoglobin A1c test, asthma medication	Provider group	Partial effect: improved only in cervical cancer screening, not improved in mammography, haemoglobin A1c test, asthma medication	U.S. A large network HMO, PacifiCare Health System introduced Quality Incentive Program to contracted medical groups in California in March 2002; One year later, PacifiCare with five other big health plans introduced another larger P4P program by the Integrated Healthcare Association (IHA).	2002-2004	Treatment groups size (77-186) medical groups; Control group size: (7-32) medical groups

Appendix 2-- All data sources and main relevant variables

Data Source	Relevant information

Ontario Health Insurance Program (OHIP) physician claims data	Claim records to calculate the services provided by physicians;
	Basic provider information;
	Basic patient information;
	Fee paid/billed;
	Patient encrypted health number as linking variable;
	Provider encrypted number as linking variable

Corporate Provider Database (CPDB) data	Physician demographic variables;
	Physician practice variables;
	Physician PCR group participation, effective dates;
	Provider encrypted number as linking variable

Client Agency Program Enrollment (CAPE) data	Patient member status;
	Patient roster dates;
	Patient encrypted health number as linking variable;
	Provider encrypted number as linking variable

Registered Persons Database (RPDB) data	Demographics of registered persons; Postal code of residence; Patient encrypted health number as linking variable

Appendix 3-- Assigning Patients to Primary Care Physicians' Practices

The Hutchison methodology (Hutchison, Hurley, Birch, Lomas, & Stratford-Devai, 1997) was implemented for all FPs in the province, not just those in the analysis sample. This ensured that an individual was assigned to an FP as called for by the algorithm, regardless of whether the FP was included in the analysis (if we focused only on the analysis sample, we would have falsely assigned some patients to sample physicians when the individual's real family physician was not included in the sample). A FP's practice population is defined as:

- All persons for whom the physician billed OHIP for at least one visit (see below for how a visit was defined) during the previous fiscal year; and
- All additional patients for whom the physician billed OHIP for at least one visit in each of the two preceding fiscal years.
- Patients who met these criteria for more than one physician were assigned to the physician who billed for the largest number of visits in the most recent year.
- When an equal number of visits were made to more than one physician in the most recent year, assignment is made to the made to the physician who billed for the most recent visit.

A service is defined as a FP visit if:

The attending physician is a FP and the fee code is one of the following 74 visit codes from the Ontario Schedule of Benefits

Fee Schedule Code	**Description**
A001A	Minor assessment
A003A	General assessment
A004A	General re-assessment
A005A	Consultation
A006A	Repeat consultation
A007A	Intermediate assessment or well baby care
A008A	Mini assessment
A110A	Periodic oculo-visual assessment, aged 19 years and below
A112A	Periodic oculo-visual assessment, aged 65 years and above
A115A	A major eye examination
A888A	Emergency department equivalent – Partial assessment
A901A	House call assessment – First patient seen
A902A	House call assessment – Pronouncement of death in the home
A903A	Pre-dental/operative general assessment (maximum of 2 per 12-month period)
A905A	Limited consultation

Fee Schedule Code	Description
A933A	On-call admission assessment
A945A	Special palliative care consultation
E070A	Geriatric Geriatric Age Premium: Gen. Practice - Geriatic Gen. Assess. Premium - 75 or Older
E071A	
E075A	Geriatric general assessment premium – patient aged 75 or older (maximum 1 per 12 month period)
E077A	Identification of patient for a Major Eye Examination
G212A	Diagnostic and Therapeutic Procedure – Hyposensitisation, including assessment and supervision – When sole reason for visit
G271A	Diagnostic and Therapeutic Procedure – Cardiovascular – Anticoagulant supervision – long-term, telephone advice
G365A	Diagnostic and Therapeutic Procedure – Papanicolaou Smear – periodic
G372A	Diagnostic and Therapeutic Procedure – Intramuscular, subcutaneous or intradermal – With visit (each injection)
G373A	Diagnostic and Therapeutic Procedure – Intramuscular, subcutaneous or intradermal – Sole reason (first injection)
G538A	Active immunization – Injection of unspecified agent – with visit (each injection)
G539A	Active immunization – Injection of unspecified agent – sole reason (first injection)
G590A	Active Immunization – Injection of influenza agent – With visit
G591A	Active Immunization – Injection of influenza agent – Sole reason
K004A	Family psychotherapy – 2 or more family members in attendance at the same time
K005A	Primary mental health care - Individual care – Per half hour
K006A	Hypnotherapy – Individual care – Per half hour
K007A	Psychotherapy – Individual care – Per half hour
K010A	Psychotherapy – Additional units per member (maximum 6 units per patient per day)
K011A	Hypnotherapy – Group – for induction and training for hypnosis (maximum 8 people), per member, per half hour
K012A	Psychotherapy, Group – Per member of a group of 4, first 12 units per day
K013A	Counselling – Individual care – Per half hour
K017A	Annual health or annual physical examination – Child after second birthday
K019A	Psychotherapy, Group – Per member of a group of 2, first 12 units per day
K020A	Psychotherapy, Group – Per member of a group of 3, first 12 units per day
K022A	HIV Primary Care – Individual care per half hour
K023A	Palliative care support – Individual care per half hour
K024A	Psychotherapy, Group – Per member of a group of 5, first 12 units per day
K025A	Psychotherapy, Group – Per member of a group of 6 to 12, first 12 units per day
K026A	Family Practice & Practice in General - Certification of medical eligibility for Ontario Hepatitis C Assistance Program
K027A	Family Practice & Practice in General – Certification of medical eligibility for Ontario Hepatitis C Assistance Program (OHCAP)
K028A	Family Practice & Practice in General – Sexually transmitted disease management
K030A	Family Practice & Practice in General – Diabetic management assessment
K031A	Health Protection and Promotion Act – Physician Report – Completion of Physician Report in accordance with Section 22.1 of the Health Protection and Promotion Act.
K033A	Counselling – Individual care – Additional units per patient per provider per 12-month period
K040A	Group counselling – 2 or more persons
K041A	Group counseling – 2 or more persons – Additional units
K070A	Home care application – Application

Fee Schedule Code	**Description**
K071A	Home care supervision – Acute Home Care Supervision (maximum 1 every 2 weeks for the first 12 weeks following admission to home care program).
K072A	Home care supervision – Chronic Home Care Supervision (maximum 1 per month commencing in the 13th week following admission to the home care program).
K623A	Certification of mental illness – Form 1 – Application for psychiatric assessment in accordance with the *Mental Health Act* – includes necessary history, examination, notification of the patient, family and relevant authorities and completion of form.
P004A	Obstetrics, prenatal care – Minor prenatal assessment
W001A	Non-emergency long-term care in-patient services – Subsequent visits – Chronic care or convalescent hospital – additional subsequent visits (maximum 4 per patient per month)
W002A	Non-emergency long-term care in-patient services – Subsequent visits – Chronic care or convalescent hospital – first 4 subsequent visits per patient per month
W003A	Non-emergency long-term care in-patient services – Subsequent visits – Nursing home or home for the aged – first 2 subsequent visits per patient per month
W004A	Emergency or Out-Patient Department (OPD) & Non-Emergency Long-Term Care In-Patient Services – General re-assessment of patient in nursing home
W008A	Non-emergency long-term care in-patient services – Subsequent visits – Nursing home or home for the aged – additional subsequent visits (maximum 2 per patient per month)
W102A	Emergency or Out-Patient Department (OPD) & Non-Emergency Long-Term Care In-Patient Services – Admission assessment – Type 1
W104A	Emergency or Out-Patient Department (OPD) & Non-Emergency Long-Term Care In-Patient Services – Admission assessment – Type 2
W105A	Emergency or Out-Patient Department (OPD) & Non-Emergency Long-Term Care In-Patient Services – Consultation
W106A	Emergency or Out-Patient Department (OPD) & Non-Emergency Long-Term Care In-Patient Services – Repeat Consultation
W109A	Emergency or Out-Patient Department (OPD) & Non-Emergency Long-Term Care In-Patient Services – Admission assessment – Annual physical examination
W121A	Non-emergency long-term care in-patient services – Subsequent visits – Nursing home or home for the aged – Additional visits due to intercurrent illness
W107A	Emergency or Out-Patient Department (OPD) & Non-Emergency Long-Term Care In-Patient Services, Admission assessment – Type 3
W777A	Emergency or Out-Patient Department (OPD) & Non-Emergency Long-Term Care In-Patient Services – Intermediate assessment - Pronouncement of death
W872A	Non-emergency long-term care in-patient services – Subsequent visits – Nursing home or home for the aged – Palliative care
W882A	Non-emergency long-term care in-patient services – Subsequent visits – Chronic care or convalescent hospital – Palliative care
W903A	Emergency or Out-Patient Department (OPD) & Non-Emergency Long-Term Care In-Patient Services – Pre-dental/pre-operative general assessment (maximum of 2 per 12-month period)

Appendix 4-- Independent Variable Specification

Variable Name	Variable Description
treatment	Treatment dummy indicating whether the GP was eligible for the financial incentive on that date.
post	Pre- & post- dummy indicating whether it is post-intervention period
treated	Treated & untreated dummy indicating whether the this physician was entitled to collect bonus/special payment during out study period
ffsd fhgd fhnd ccmd fhod	PCR dummies indicating the physician was affiliated with the model on that date ffsd: 1 if FFS, 0 otherwise fhgd: 1 if FHG, 0 otherwise fhnd: 1 if FHN, 0 otherwise ccmd: 1 if CCM, 0 otherwise fhod: 1 if FHO, 0 otherwise
doc_age	Age of physician in years on March 31 2002
doc_ageg1 doc_ageg2 doc_ageg3 doc_ageg4	Physician age-- four categories doc_age1: physician age < 45; doc_age2: 45 ≤ physician age < 50; doc_age3: 50 ≤ physician age < 60; doc_age4: 60 ≤ physician age
doc_male	Physician sex: 1 if male; 0 if female
cmaca	Practice location$_{S:}$ metropolitan area influence measured by Metropolitan Influence Zone (five categories: 0 if CMA, 4 levels otherwise: strong, medium, weak, no influence)
riotype	Practice location$_{S:}$ urban/rural level measured by RIO score (Rurality Index of Ontario) riotype =0 if RIO score > 45; riotype =1 if RIO score ≤ 45
yrslicyr	Years since licensing year as of March 31 2008
lnbsbill	Log of billings: log of total value (fee approved) of claims submitted in 1998-99
workdays	Days of working: total number of days billed in this fiscal year
sum_visit	Number of visits: total # of patient visits in this year
sum_visitg1 sum_visitg2 sum_visitg3 sum_visitg4	Total # of patient visits in this year: four categories sum_visitg1: < 5,000; sum_visitg2: 5,000 ≤ x < 7,500; sum_visitg3: 7,500 ≤ x < 10,000; sum_visitg4: ≥ 10,000
visitpd	Number of patient visits per working day: the number of visits divided by number of days worked in this year
sum_pt	Practice size: number of assigned patients seen per year
sum_ptg1 sum_ptg2 sum_ptg3 sum_ptg4	Practice size: number of assigned patients seen per year, four categories sum_ptg1: < 1,000; sum_ptg2: 1,000 ≤ x < 1,500; sum_ptg3: 1,500 ≤ x < 2,000; sum_ptg4: ≥ 2,000
meanage	Average practice age: average age of eligible patient population on each snapshot date
fmpercent	Proportion of females in the practice: proportion of female patients as of eligible patient population on each snapshot date
clpercent	Proportion of children patients: Proportion of the eligible patients that are under 2 years of age
elpercent	Proportion of elderly patients: Proportion of the eligible patients that are over 65 years of age
year	Year fixed effects:10 dummy variables for each snapshot year from March 31 1999 to March 2008
cdname	Geographical fixed effects: 49 dummy variables defined by Census Division codes