DISCUSSION PAPER SERIES

IZA DP No. 6474

Link between Pay for Performance Incentives and Physician Payment Mechanisms: Evidence from the Diabetes Management Incentive in Ontario

Jasmin Kantarevic
Boris Kralj

April 2012

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Link between Pay for Performance Incentives and Physician Payment Mechanisms: Evidence from the Diabetes Management Incentive in Ontario

Jasmin Kantarevic
Ontario Medical Association
and IZA

Boris Kralj
Ontario Medical Association

Discussion Paper No. 6474
April 2012

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 6474
April 2012

ABSTRACT

Link between Pay for Performance Incentives and Physician Payment Mechanisms: Evidence from the Diabetes Management Incentive in Ontario*

Pay for performance (P4P) incentives for physicians are generally designed as additional payments that can be paired with any existing payment mechanism such as salary, fee-for-service, and capitation. However, the link between the physician response to performance incentives and the existing payment mechanisms is still not well understood. In this paper, we study this link using the recent primary care reform in Ontario as a natural experiment and the Diabetes Management Incentive (DMI) as a case study. Using a comprehensive administrative data and a difference-indifferences matching strategy, we find that physicians in a blended capitation model are more responsive to the DMI than physicians in an enhanced fee-for-service model. We show that for a given payment mechanism this result implies that the optimal size of P4P incentives varies negatively with the degree of supply-side cost sharing. These results have important implications for the design of P4P programs and the cost of their implementation.

JEL Classification: I10, I12, I18

Keywords: pay for performance, physician remuneration, diabetes management

Corresponding author:

Jasmin Kantarevic
Ontario Medical Association
525 University Avenue
Toronto, ON M5G 2K7
Canada
E-mail: jasmin.kantarevic@oma.org

* The views expressed in this paper are strictly those of the authors. No official endorsement by the Ontario Medical Association is intended or should be inferred.

1. Introduction

Pay for performance (P4P) programs have become increasingly popular in recent health care reforms. Two well-known examples include the Quality and Outcomes Framework in the U.K. and the California Pay for Performance Program in the U.S., but there are similar programs in many other countries[1]. The P4P programs provide incentives to health care providers for achieving selected performance targets, such as improving preventive and chronic care, patient experience, and the use of information technology. The broad goal of these programs is to enhance healthcare quality, which is expected to improve long-term patients' health and reduce healthcare costs[2]. Such promising outcomes put the P4P programs at the front and centre of many recent health care reforms.

Changing physician practice is a critical step for implementing successful P4P programs. However, recent empirical evidence on the impact of P4P programs on physician practice is quite mixed[3]. This puzzling result can be explained in at least two ways. First, there are significant differences across studies in the type of evaluation methodology used to identify the P4P impact, such as the sample size, the nature of comparison group, and the set of included confounding factors. Second, there is wide variation in the structure of P4P programs, such as the size of financial incentives, the use of absolute versus relative targets, and the use of individual-based versus group-based payments. Consequently, it is not clear whether the lack of consensus in the literature on the impact of P4P programs is due to methodological shortcomings or because some P4P programs are just poorly designed. In this study, we focus on the second question of the optimal design of P4P programs. Compared to the literature on whether availability of a specific P4P program affects physician behaviour, the empirical evidence on this

[1] For an overview of these programs, see for example Smith and York (2004) for the U.K., the Integrated Healthcare Association (2006) for California, and references in Frolich et al. (2007) for other countries.
[2] See for example Dusheiko et al. (2011) for the impact of the Quality and Outcomes Framework on reducing hospital costs and mortality.
[3] For recent surveys, see for example Armour et al. (2011), Christianson et al. (2008), Li et al. (2011), Petersen et al. (2006), Town et al. (2005), and Rosenthal and Frank (2006).

question is still quite limited, which reduces our ability to design and implement successful P4P programs[4].

We contribute to this literature by examining how the optimal size of P4P incentives depends on the supply-side cost sharing in the physician compensation mechanisms. This cost sharing refers to the degree to which physicians are reimbursed for incremental services, after receiving any fixed payment. The two extreme examples of cost sharing are the fee-for-service model, with no cost sharing, where physicians receive the full value of incremental services but no fixed payment, and the pure capitation model, with full cost sharing, where physicians receive a fixed payment per patient but no reimbursement for incremental services. This question is of policy interest in many countries in which physicians practice in models with various degrees of cost sharing, such as Canada and the U.S., in which policy makers have to determine the size of P4P incentives. The question is also relevant in countries with a single predominant type of physician compensation mechanism, such as the U.K., where the introduction of new P4P programs may be contemplated along with changes in the degree of supply-side cost sharing.

In Section 2, we show that the relation between the optimal size of P4P incentives and the supply-side cost sharing depends critically on the link between the physician response to the P4P programs and the type of physician compensation mechanism. We study this link empirically using the recent primary care reform in Ontario as a natural experiment. In this reform, new compensation models with varying degrees of supply-side cost sharing were sequentially introduced. We use the differential timing of the introduction of these models and the physician transition between the models as a main source of identification. Specifically, we study the physician response to the Diabetes Management Incentive (DMI), an annual bonus that physicians receive for planned, ongoing management of diabetic patients, between physicians practicing in an enhanced fee-for-service model (the Family Health Groups) and a blended capitation model (the Family Health Organizations). These two models are currently the most prevalent payment models in Ontario, comprising about two thirds of all primary care physicians. We provide more institutional background on these two models and on the DMI in Section 3.

[4] For a recent review, see Frolich et al. (2007).

Participation of physicians in the new payment models is voluntary, which generates concerns about the selection bias if, as expected, factors that affect physician participation in a model also affect their response to the DMI. To address this problem, we use a difference-in-difference matching strategy which allows us to control for unobserved, time-invariant physician heterogeneity. This empirical strategy is discussed in detail in Section 4. In addition, the matching approach is particularly appealing in our study because of availability of rich administrative data, described in Section 5, which includes medical profiles of almost all physicians in Ontario that can be used to predict the physician choice of the compensation model. Our focus on Ontario is also attractive because it is a single payer system with universal health insurance coverage. Therefore, all physicians within a compensation model face the same financial incentives and demand for medical services is unlikely to be affected by changes in the incentives offered to physicians.

Our results, presented in Section 6, indicate that physicians in the blended capitation model are about 12 percent more likely to participate in the DMI than physicians in the enhanced fee-for-service model. We also find that diabetic patients enrolled to the capitation physicians are about 8 percent more likely to receive the DMI services than diabetic patients enrolled to the fee-for-service physicians. These results suggest that the physician response to the P4P incentives varies positively with the degree of supply-side cost sharing. Furthermore, these results imply that for a given compensation mechanism the optimal size of the P4P incentives varies negatively with the degree of cost sharing. Additional comments and our conclusions are presented in Section 7.

Our analysis contributes to the existing literature in three main ways. First, as mentioned, understanding the link between P4P programs and physician payment mechanisms has important implications for both the design of effective P4P programs and the cost of their implementation. Second, diabetes is one of the most common and costly of all chronic diseases[5]. In addition, it is relatively well

[5] According to the International Diabetes Federation (2010), the estimated diabetes prevalence for 2010 rose to 285 million, representing 6.4% of the world's adult population, with a prediction that by 2030 the number of people with diabetes will have risen to 438 million. In Ontario, diabetes costs are estimated at C$4.9 billion, or about 10% of the total healthcare budget (2012 Ontario Budget Speech). Dali et al. (2010) estimate that the U.S. national economic

understood medically, and there is broad-based agreement on how to manage the disease. Despite this professional knowledge, however, there is widespread concern that diabetes is poorly managed and that it can be significantly improved through incentive programs. Lastly, understanding the impact of different payment models on quality of patient care has been an important policy question for long time[6]. Most of the earlier literature focused on the case where quality cannot be observed or verified. Relatively less is known about the impact of payment models when verifiable and contractible indicators of quality are available, such as the DMI in Ontario and many P4P programs in other jurisdictions.

2. Optimal Size of P4P Incentives and Physician Compensation Mechanisms

P4P incentives for physicians are generally designed as additional payments that are paired with the existing physician payment mechanism such as fee-for-service and capitation. In this section, we develop a simple model to reflect this policy problem with the aim of determining the optimal size of a P4P incentive, given the existing payment mechanism. Our model builds on the recent contributions by Eggleston (2005) and Kaarboe and Siciliani (2011)[7].

We assume that a policy maker wishes to maximize the patient benefit (B) net of physician payment (I)[8]:

(1) $W = B - I$

The patient benefit depends on the quantity (q) and quality (e) of medical services according to $B(q,e)$, with B_q, $B_e > 0$ and B_{qq}, $B_{ee} \leq 0$[9]. The sign of B_{qe} depends on whether e and q are complements ($B_{qe} > 0$) or substitutes ($B_{qe} < 0$) in the patient benefit function.

---

burden of pre-diabetes and diabetes reached US$218 billion in 2007, with the average annual cost of US$9,677 for type 2 and US$14,856 for type 1.

[6] For recent surveys of this literature, see for example McGuire (2000) and Leger (2008).

[7] Our model can also be interpreted as a special case of the classic multitasking problem where both tasks are perfectly observable and the principal cares about the agent's welfare (e.g. Holmstrom and Milgrom, 1991).

[8] This is the same welfare function studied by Kaarboe and Siciliani (2011). See also Boadway et al. (2004) and Chalkley and Malcomson (1998) for alternative specifications.

The physician payment per patient can be represented in a general way as:

(2) $I = R + rq + pe$

where R represents the fixed payment per patient, r represents the reimbursement rate for incremental services, and p represents the quality bonus such as the DMI. The degree of supply-side cost sharing is captured by parameter r. With the full cost sharing, as in the pure capitation model, $R > 0$ and $r = 0$. With no cost sharing, as in the pure fee-for-service model, $R = 0$ and $r = 1$. In a mixed capitation model that is common in many countries, $R > 0$ and $r \in (0, 1)$.

The policy maker's problem in our environment is to choose p, given the existing payment mechanism (R, r). This problem is also subject to two additional types of constraints: the physician participation constraint and the incentive compatibility constraint.

The participation constraint requires that the physician utility from participating in the P4P program is at least as large as her outside option from not participating. Without loss of generality, we normalize this outside option to 0. The physician utility can be expressed as:

(3) $U = \alpha B + I - C(q,e)$

where $\alpha \geq 0$ represents the extent of physician's altruism and $C(.)$ represents the physician disutility function, with C_q, $C_e > 0$ and C_{qq}, $C_{ee} \geq 0$. The sign of C_{qe} depends on whether e and q are complements ($C_{qe} < 0$) or substitutes ($C_{qe} > 0$) in the physician disutility function. The participation constraint is then $U \geq 0$, which in equilibrium binds with equality.

The incentive compatibility constraint requires that the policy maker incorporates the physician best response to any given contract (R, r, p) into the decision making process. The physician best response can be described by the first-order conditions to the problem of choosing (q, e) to maximize U given the compensation contract. For the interior solution, these conditions are:

(4) $\alpha B_q(q^*,e^*) + r - C_q(q^*,e^*) = 0$

(5) $\alpha B_e(q^*,e^*) + p - C_e(q^*,e^*) = 0$

[9] We refer to q as the number of medical services, while it is more properly interpreted as the value of medical services, where the price per service is normalized to one. Therefore, other prices in the model (R and r) should be interpreted as relative to the price of medical services.

The solution to these two conditions is the physician best response functions $q(r, p)$ and $e(r, p)$[10]. It is straightforward to show, using the Cramer's rule, that $\partial q/\partial r = (C_{ee} - \alpha B_{ee})/D > 0$ and $\partial e/\partial p = (C_{qq} - \alpha B_{qq})/D > 0$, where $D=U_{qq}U_{ee}-U_{eq}^2 > 0$ by the second-order necessary condition. Therefore, as expected, the physician provision of quantity and quality depends positively on their own prices. In addition, it is easy to show that $\partial e/\partial r = \partial q/\partial p = (\alpha B_{eq} - C_{eq})/D$. The sign of this parameter is in general ambiguous and depends on whether q and e are complements or substitutes in the patient benefit and physician disutility functions. To gain some intuition, consider the standard case of effort substitution ($C_{eq} > 0$) where q and e compete for physician time. In this case, $\partial e/\partial r < 0$ as the physician re-allocates his time from quality to quantity as the marginal return to quantity increases. This opportunity cost explanation is the only mechanism through which r affects e when the physician does not care about the patient's benefit ($\alpha=0$). When the physician is altruistic, the negative impact of r on e due to the opportunity cost is amplified by the physician concern for the patient if q and r are also substitutes in the production of health ($B_{eq} < 0$). In the opposite case, the physician's concern for the patient mitigates the negative impact of r on e and the net impact depends on the relative magnitudes of α, B_{eq}, and C_{eq}.

Using the physician participation and incentive compatibility constraints, the policy maker's objective function can be expressed as:

(6) $W = (1+\alpha)B(q(r,p),e(r,p)) - C(q(r,p),e(r,p))$

The first-order condition for the quality bonus p is then equal to:

(7) $[(1+\alpha)B_q - C_q]\partial q/\partial p + [(1+a)B_e - C_e]\partial e/\partial p = 0$

Using the first-order conditions (4) and (5) for the physician's problem and the fact that $\partial e/\partial r = \partial q/\partial p$, equation (7) can be expressed as:

(8) $p = B_e + (B_q - r)(\partial e/\partial r)/(\partial e/\partial p)$

This equation relates the optimal size of P4P incentive p to the degree of supply-side cost sharing r. Given that $\partial e/\partial p > 0$, this relation depends critically on the sign of $\partial e/\partial r$, which is a priori ambiguous,

[10] Note that q and e do not depend on the fixed payment R, which plays a role only in the participation constraint.

as we discussed earlier. In our empirical analysis, we aim to determine the sign of $\partial e/\partial r$ using the variation in physician response (e) to the DMI between physicians practicing in an enhanced fee-for-service model ($r = 1$) and a blended capitation model ($0 < r < 1$). We describe these two payment models and the DMI in more detail in the next section.

3. Institutional Background

Until the early 2000s, almost all primary care physicians in Ontario practiced in a traditional fee-for-service model. In response to long-standing criticisms of this model, the government sequentially introduced a variety of new payment models[11]. The common elements in these models include patient enrolment, extended hours, and eligibility for a set of performance-based incentives, such as preventive care bonuses, special payments for providing targeted services, incentives to enrol patients with no regular family doctor, and chronic disease management incentives. The main difference between the new models is in their base compensation, with two main options of fee-for-service and capitation.

Currently, about 80 percent of primary care physicians participate in the various new payment models. In this study, we focus on the two most prevalent new models, known as the Family Health Groups (FHG) and the Family Health Organizations (FHO). As of March 2011, there were over 6,500 physicians practicing in these two models, comprising about 60 percent of all primary care physicians in Ontario. The FHG is an enhanced fee-for-service model that was introduced in 2003. In this model, physicians receive a full fee-for-service value for services provided to their enrolled patients ($r = 1$), in addition to a premium for selected comprehensive care services. The FHO is a blended capitation model that was introduced in 2007. In this model, physicians receive an age-sex adjusted capitation rate for each enrolled patient (R) and a discounted fee-for-service value for selected services ($r = 0.15$)[12].

[11] For an overview of these new models, see for example Glazier et al. (2009), Kantarevic et al. (2011), and Li et al. (2011).

[12] A more detailed description of the payment mechanism in these two models is presented in Kantarevic and Kralj (2012) and in Appendix C.

The FHG and FHO models are identical in almost all other aspects, including the eligibility for the Diabetes Management Incentive (DMI). The DMI was introduced on April 1, 2006 in response to a number of concerns related to the management of diabetic patients. Specifically, prior to 2006, primary care physicians were compensated through a variety of fee codes for services provided to diabetic patients, such as the intermediate assessment and the diabetic management assessment. These codes paid physicians for services provided during the patient visit, but not for services provided over an extended period of time. As a result, this fee-for service payment method did not explicitly encourage a planned approach to on-going management of diabetic patients. In addition, none of the existing codes required that the physician complies with all of the best clinical practice guidelines, such as those recommended by the Canadian Diabetes Association (CDA)[13].

In contrast, the DMI is paid for services provided to diabetic patients over the previous twelve months. Specifically, the DMI is paid for a planned, ongoing management of diabetic patients according to elements required by the Canadian Diabetes Association (CDA) Clinical Practice Guidelines. These elements include: "(a) tracking lipids, cholesterol, HbA1C, blood pressure, weight and body mass index (BMI), and medication dosage, (b) discussion and offer of preventive measures including vascular protection, influenza and pneumococcal vaccination, (c) health promotion counselling and patient self-management support, (d) tracking of albumin to creatinine ratio, (e) discussion and offer of referral for dilated eye examination, and (f) foot examination and neurologic examination."[14] To meet these guidelines, the physician must see the patient at least twice during the last twelve months. Physicians who meet the requirements may claim a code Q040 and receive an annual bonus of $60 per patient[15]. This bonus is payable in addition to the existing codes for services provided during the patient visit.

When it was introduced, the DMI was restricted to services provided by physicians in the patient enrolment models to their enrolled patients. As of April 1, 2009, this restriction was removed and

[13] The diabetic management assessment requires that physician complies with a subset of the guidelines specified by the CDA. This subset includes elements described in part (a) for the DMI, discussed later in this Section.
[14] Schedule of Benefits, Physician Services under the Health Insurance Act (September 1, 2011), Ontario Ministry of Health and Long Term Care, page A39.
[15] As a reference, this is equivalent to the fee for about two regular office visits (i.e. intermediate assessments).

eligibility was extended to all family physicians and both enrolled and non-enrolled patients. At the same time, the value of the DMI increased from C$60 to C$75 per patient.

As mentioned previously, our main empirical goal is to determine how the reimbursement rate affects the physician quality effort ($\partial e/\partial r$). To do so, we use the variation in r between the FHG and FHO models to identify its impact on the physician response to the DMI (e). Again, this comparison is particularly appealing because other payment elements, including the DMI, are nearly identical between the two models[16]. However, a simple comparison between the two models may not be appropriate because physicians freely choose which model to join. This voluntary participation raises concerns about the selection bias if, as expected, factors that affect physician participation in a model also affect their response to the DMI. In the next section, we present our empirical approach to deal with this potential problem.

4. Difference in Difference Matching

4.1 Parameter of Interest

We wish to evaluate the difference in the physician response to the DMI between physicians participating in the FHG and FHO model. This evaluation problem can be studied within a potential outcomes framework[17] in which we can precisely define the parameter of interest and clarify the assumptions needed to identify it.

Consider a simple setup in this framework with two periods and treatment in the second period only. Specifically, let t = 0 denote the period prior to the introduction of the FHO model and let t = 1 denote the period after its introduction. In addition, let d_{it} denote the treatment indicator for whether

[16] The minor differences include the Group Management and Leadership funding and the eligibility for the Continuing Medical Education grants, which apply only to the FHO model. However, these elements for non-clinical work represent a minor source of income for physicians participating in the FHO model.
[17] This model is also known as the Rubin causal model. See for example Rosenbaum and Rubin (1983, 1985).

physician *i* participates in the FHO model at time *t*. In this setup, $d_{i0}=0$ for all physicians, $d_{i1}=0$ for the FHG physicians, and $d_{i1}=1$ for the FHO physicians. Lastly, let y^1_{it} and y^0_{it} denote the potential outcomes (i.e. the physician response to the DMI) conditional on participating in the FHO and FHG model, respectively. For each physician, we can observe only y^1 or y^0 at any time. This observed outcome can be expressed as $y_{it} = d_{it}y^1_{it}+(1-d_{it})y^0_{it}$.

Given this setup, we can precisely define any parameter we wish to study. In the literature, two commonly studied parameters are the mean impact of treatment (ATE) and the mean impact of treatment on the treated (ATT)[18]. In this paper, we focus on the ATT because its identification requires much weaker assumptions that the identification of the ATE, as we discuss below. In addition, given the voluntary participation in the new models, the ATE may be less policy relevant. In our setup, the ATT can be defined as $E[y^1_{i1}-y^0_{i1}|d_{i1}=1]$, which represents the mean difference between actual and potential outcomes for the group of treatment physicians. One limitation of this definition is that it uses data from the post-treatment period only. To exploit data from both pre-treatment and post-treatment periods, we use an equivalent definition of the ATT that can be expressed as:

(9) $ATT \equiv E[y^1_{i1}-y^0_{i0}|d_{i1}=1] - E[y^0_{i1}-y^0_{i0}|d_{i1}=1] = E[\Delta y_{it}|d_{i1}=1] - E[\Delta y^0_{it}|d_{i1}=1]$.

4.2 Identification Assumptions

Without further assumptions, the ATT cannot be identified because we only observe $E[\Delta y_{it}|d_{i1}=1]$ but not the counterfactual outcome $E[\Delta y^0_{it}|d_{i1}=1]$. In this study, we construct this missing counterfactual using the sample of comparison FHG physicians and estimate the ATT using the difference-in-difference (DD) matching estimators[19].

The identification of the ATT in the DD matching framework relies on two main assumptions. The first assumption, known as the Conditional Independence Assumption (CIA), requires that

[18] See for example Blundell and Costa Dias (2009) and Imbens and Wooldridge (2009).

[19] See for example Heckman, Ichimura, and Todd (1997, 1998), Smith and Todd (2005), and Ham et al. (2011). For implementation in STATA, see Leuven and Sianesi (2003) and Becker and Ichino (2002).

(10) $E[\Delta y^0_{it}|X_i, d_{i1}=1] = E[\Delta y^0_{it}|X_i, d_{i1}=0]$.

where X_i is an appropriate set of observable covariates unaffected by treatment. This assumption states that, conditional on X, the mean change in potential outcomes for the treatment physicians had they not joined the FHO model would be the same as the mean change in actual outcomes for the comparison FHG physicians. The CIA is a rather strong condition, but its plausibility in our study comes from the fact that it only needs to hold after unobserved time invariant individual characteristics that affect both treatment and outcomes have been differenced out. Furthermore, because we focus on the ATT and not the ATE, the CIA needs to hold only for Δy^0 and not for Δy^1. In other words, the DD matching estimators that we implement allow for selection on fixed unobservable characteristics and on potential treatment outcomes.

In practice, matching on all variables in X becomes impractical as the number of covariates increases. Rosenbaum and Rubin (1983) show that if Δy^0 is mean independent of treatment status given X, then it is also mean independent of treatment status given $p(X_i)=Pr(d_{i1}=1|X_i)$, where $p(X_i)$ is known as the propensity score. As a consequence, matching can be done using the propensity score alone instead of using all variables in X, and the CIA in (10) can be replaced by

(11) $E[\Delta y^0_{it}| p(X_i), d_{i1}=1] = E[\Delta y^0_{it}| p(X_i), d_{i1}=0]$.

The second assumption required for identifying the ATT in the DD matching models is that

(12) $Pr(d_{i1}=1|X_i) < 1$.

This assumption, known as the common support or overlap assumption, requires a positive probability of observing comparison physicians at each level of X. Note that we do not require that $Pr(d_{i1}=1|X_i) > 0$ because we focus on the ATT and not the ATE.

4.3 Alternative DD Matching Estimators

The alternative DD matching estimators that we consider in this study can be represented by the following general form:

(13) $\widehat{ATT} = n^{-1} \sum_i \left\{\Delta y_{it} - \sum_j w(i,j) \Delta y_{jt}\right\}$

where *i* and *j* denote, respectively, the treatment and comparison physicians in the region of common support, *n* is the number of treatment physicians in the region of common support, and $w(i,j)$ are the matching weights with $\Sigma_j w(i,j)=1$. Therefore, the DD matching estimators construct the missing counterfactual outcome Δy^0 for each treatment physician *i* by taking a weighted average of actual outcomes for comparison physicians who are matched to physician *i*.

Alternative matching estimators differ in how they construct the matching weights. We consider three commonly used matching estimators: nearest neighbour, conventional kernel, and local linear kernel.

In the nearest neighbour estimator, each treatment physician is matched on the propensity score to the nearest comparison physician. The weighting scheme for this estimator assigns the weight of one to the closest comparison physician and the weight of zero to all other comparison physicians. In a sampling with replacement version of this estimator, which we implement, a single comparison physician can be matched to more than one treatment physician. This is in general preferred to the sampling with no replacement if the distribution of propensity scores is very different between the treatment and comparison groups[20].

The nearest neighbour estimator is in general inefficient because it matches each treatment physician to a single comparison physician. This may be partially addressed by expanding the matched comparison group to $n > 1$ physicians, in which case each matched comparison physician receives an equal weight of $1/n$. However, this weighting scheme is problematic because close and distant matches receive the same weight in constructing the missing counterfactual. The conventional kernel estimator addresses this problem by matching all comparison physicians to each treatment physician and assigning a higher weight to comparison physicians closer to the matched treatment physician. Specifically, the weight that each comparison physician receives is equal to $w(i,j) = G(z_j)/\sum G(z_j)$, where $G(.)$ is the kernel function, $z_j = (p_i - p_j)/h$ is the standardized distance in the propensity score between treatment

[20] See for example Dehejia and Wahba (2002).

physician *i* and comparison physician *j* , and *h* is the bandwidth. To implement the kernel estimator, the kernel function and the bandwidth must be specified. As our baseline case, we used the bi-weight kernel, which is equal to $15/16(z^2-1)^2$ for $|z|<1$ and 0 otherwise. As a specification check, we also explore several alternative kernels. For the bandwidth selection, we use the Silverman's (1986) optimal plug-in selector, which produces the bandwidth of about 0.1 in our application, but we also experiment with alternative bandwidth values[21].

The conventional kernel estimator constructs the missing counterfactual for each treatment physician non-parametrically as the weighted average of Δy^0 among the comparison physicians, which can be interpreted as a kernel-weighted regression of Δy^0 on a constant. The local linear kernel extends this regression model to include a linear term in p_i-p_j, which is helpful whenever comparison group observations are distributed asymmetrically around the treatment observations[22]. Given the more desirable properties of this estimator compared to the conventional kernel and nearest neighbour, we use the local linear kernel as our baseline estimator.

4.4 Standard Error Estimation

Due to the complexity of the propensity score matching, most empirical studies rely on bootstrapping to compute the standard errors for the effect of treatment. This approach is expected to work well for the kernel and local linear kernel matching estimators, but it is in general not valid for the nearest neighbour due to its extreme non-smoothness (Abadie and Imbens, 2008). In implementing the bootstrap, we choose the optimal number of repetitions using the three-step methodology developed by Andrews and Buchinsky (2000, 2001)[23]. In our application, this optimal number of repetitions is about 200.

[21] This bandwidth selector is described in detail in Appendix A.
[22] For example, Fan (1992, 1993) shows that the local linear estimator has a faster rate of convergence near boundary points and greater robustness to different data design densities than the conventional kernel estimator.
[23] This methodology is described in detail in Appendix B.

5. Data

The data comes from several administrative sources maintained by the Ontario Ministry of Health and Long-Term Care. Specifically, the Corporate Provider Database provides information on physician affiliation with a patient enrolment model, the Client Agency Program Enrolment database provides the list of all enrolled patients, and the Ontario Health Insurance Plan database provides detailed, claim-level data on physician services provided to each patient. These sources can be linked together using encrypted physician and patient numbers to construct a comprehensive database that includes almost all family physicians and enrolled patients in Ontario and their entire profile of medical services.

The study period for our analysis includes fiscal years 2006 and 2010, one year before and three years after the FHO model was introduced in Ontario. For these two years, we focus on a cohort of physicians affiliated with the FHG model as of April 1, 2006. This cohort includes 4,455 physicians, or about 40 percent of all primary care physicians in Ontario. Of this cohort, 441 physicians ceased to practice in Ontario between 2006 and 2010 for various reasons such as retirement and migration. Further, 197 physicians switched to a patient enrolment model other than the FHO. These physicians were excluded from our analysis because our main focus is on the comparison between FHG and FHO physicians. Lastly, we excluded 162 physicians who had no enrolled patients in either 2006 or 2010[24]. The final sample used for the analysis therefore includes 3,655 physicians. Of this sample, about 42 percent of physicians switched to the FHO model by 2010. For our purposes, these 1,521 physicians are defined as treatment physicians, while the other 2,134 physicians who remained in the FHG model are defined as comparison physicians.

The outcome of interest is measured in two complementary ways to capture the extensive and intensive margins of physician response to participating in the FHO model. On the extensive margin, we use a binary indicator for whether the physician participated at all in the DMI (i.e. whether the physician

[24] Unfortunately, it is difficult to determine the impact of these exclusions on our results. For physicians not present in 2010, we cannot calculate changes in outcomes because we have only one observation per physician; for physicians who switched to other models, it is difficult to disentangle the impact of these models from the supply-side cost sharing that we are interested in; and for the physicians with no enrolled patients, we cannot calculate one of our outcomes (the percent of enrolled diabetic patients with the DMI), as we explain later in this Section.

provided any Q040 services). One important advantage of using this measure is that it is expected to be measured with virtually no error. In addition, the results concerning this outcome may be particularly informative if factors that affect the decision to participate differ from factors that affect the decision on how many Q040 services to provide conditional on participation. On the intensive margin, we use the percent of enrolled diabetic patients who received Q040 services[25]. This measure is appealing because it reflects the targeted patient population and because it accounts for differences in the share of enrolled diabetic patients and the workload between physicians.

The set of covariates includes matching variables that we expect to belong to the propensity score model. The choice of the appropriate matching variables is critical for consistently estimating the treatment effect[26], which makes matching particularly attractive in our study because we have access to rich data on physician practices in the pre-treatment period. Specifically, the included matching variables are related to: (1) physician characteristics (physician age, sex, and experience with the patient enrolment models (as measured by the number of days in the FHG model as of April 1, 2006)); (2) practice characteristics (the geographic location of practice, the number of enrolled patients, the annual number of patient visits, and the number of other physicians in practice); (3) patient characteristics (the patient complexity (as measured by the risk-adjustment factors based on patients' age and gender) and the share of enrolled diabetic patients), (4) the expected income gain (estimated using the actual service and patient profiles in fiscal 2006 and the administrative payment rules in the FHG and FHO models[27]), and (5) past outcomes (an indicator for physician participation in the DMI in fiscal 2006, the percent of enrolled diabetic patients who received the DMI in fiscal 2006). To ensure that the included covariates are not determined by treatment, all of the variables are measured prior to the introduction of the FHO model.

[25] To identify the diabetic patients, we use a methodology similar to the Institute for Clinical and Evaluative Studies (2003). Specifically, the patients are identified as diabetic patients if they had any services over the last year with the Diabetes Mellitus ICD -10 diagnosis code or using fee codes that are provided exclusively to the diagnosed diabetic patients (the full list is available upon request). Using this methodology, we identified 724,237 diabetic patients in 2006 and 850,067 diabetic patients in fiscal 2010, which is within the range of published estimates.
[26] See for example Heckman, Ichimura, and Todd (1997, 1998) and Smith and Todd (2005).
[27] See appendix C for details.

Descriptive statistics for the sample included in the analysis are presented in Table 1. The first two columns contain variable names and definitions. The next three columns present the means for the whole sample, the treatment sample, and the comparison sample, respectively. The last column presents the difference in means between treatment and comparison physicians. Standard errors for the sample means are presented in parentheses.

The upper panel of Table 1 shows the outcomes of interest. On the intensive margin, the percent of enrolled diabetic patients who received Q040 services was 22 percent in fiscal 2006 and 34 percent in fiscal 2010. This outcome was significantly larger for the treatment physicians in both years, with the difference growing over time from about 8 percent in 2006 to about 13 percent in 2010. The simple difference-in-difference estimate of the FHO impact is about five percent and it is statistically significant. On the extensive margin, about 49 percent of sample physicians provided Q040 services in 2006 and 68 percent in fiscal 2010. Again, this outcome is significantly larger for the treatment physicians in both years, with the difference growing from about 13 percent in 2006 to about 21 percent in 2010. Furthermore, the simple difference-in-difference estimate of about 8 percent is statistically significant. These unadjusted comparisons of outcomes suggest that the treatment physicians responded to the DMI more than the comparison physicians on both extensive and intensive margins.

The lower panel of Table 1 shows the distribution of covariates across the two groups of physicians as of fiscal 2006. These statistics indicate that the treatment physicians are on average two years younger and about two percent less likely to live in the Toronto region. In addition, the treatment physicians enrol more patients, practice in smaller groups, provide fewer annual visits, and have been affiliated with the FHG model for a longer time. Perhaps most significantly, the expected income gain from joining the FHO model is about C$57,000 for the treatment physicians and about -$15,000 for the comparison physicians. All of these differences are statistically significant and suggest that physicians who joined the FHO model were a selected, non-random group of FHG physicians. This selection on observed covariates may also be indicative of selection on unobserved characteristics. These preliminary

results confirm the need to address the potential selection bias when estimating the impact of participating in the FHO model.

6. Results

6.1 Propensity Scores

Table 2 presents the propensity score logit estimates for participation in the FHO model. With the exception of gender, group size, and the intensive measure of past outcomes, all coefficients are statistically significant[28]. However, some coefficients do not have signs expected from the descriptive statistics reported in Table 1. This is not surprising because some covariates are highly correlated, such as the number of enrolled patients and the number of annual visits. In addition, this is not a serious concern because the propensity score model does not necessarily represent a structural behavioural relationship, as its main role in matching is to provide a good model for predicting treatment.

The estimated model has a good fit. The likelihood ratio test clearly rejects the hypothesis that included variables are jointly insignificant[29]. In addition, McFadden's R^2 is about 0.24[30]. Furthermore, the model correctly predicts treatment for about 72 percent of sample physicians. This prediction metric is constructed by comparing the actual treatment status of each physician to their estimated probability of treatment. A prediction is considered to be correct if the estimated propensity score is above 0.42 for the treatment physician and below 0.42 for the comparison physician, where 0.42 represents the percent of sample physicians in the treatment group.

[28] The quadratic forms for age, visits, and roster size are all statistically significant.
[29] The LR chi-square statistic with 30 degrees of freedom is about 1,194, with the associated p-value < 0.000.
[30] This R^2 is calculated as 1-L(B)/L(0), where L(B) denotes the fitted log-likelihood value of the model and L(0) denotes the value of log-likelihood in a constant-only model. The lower and upper bounds of this pseudo R^2 are 0 and 1, but this pseudo R^2 is not a measure of proportion of variance of the dependent variable explained by the model.

We chose the functional form of the variables included in the model to ensure that they are distributed similarly across the treatment and matched comparison physicians using balancing tests originally proposed by Rosenbaum and Rubin (1985). Specifically, for a given functional form, we tested whether our empirical model balanced the sample via paired t-tests and joint F-tests. The paired t-tests examine whether the mean of each covariate for the treatment group is equal to that of the matched comparison sample. The joint F-tests examine whether, at each quintile of the propensity score distribution, the mean of all covariates are jointly different between treatment and comparison physicians.

Table 3 shows these balancing tests, using the full sample of treatment physicians and matched sample of comparison physicians obtained using the nearest neighbour matching. The upper panel shows the paired t-tests. These tests indicate that matching balances the two groups of physicians on each pre-treatment covariate quite well, as none of the reported differences are significant at the standard test levels. The lower panel shows the joint F-tests. For the middle three quintiles, the F-tests cannot reject the hypothesis that these covariates are jointly insignificant. However, the F-tests are significant at the 1st and 5th quintiles, unless further restrictions are imposed on the propensity score distribution. Specifically, the F-tests are insignificant at the standard test levels only when the sample excludes observations with propensity score below 0.05 and over 0.95. Rather than imposing this restriction on our analysis, we present all of our results using the unrestricted sample and conduct the analysis with the restricted sample as a specification check[31].

Lastly, the estimated propensity scores can be used to evaluate the validity of the overlap assumption in our sample. Figure 1 presents the distribution of the propensity scores for the treatment and comparison physicians. This figure shows that the empirical support of the two distributions is very similar, although, as expected, the treatment physicians have a higher average probability of joining the FHO model than the comparison physicians. However, the overlap assumption fails for a small number of physicians at the extremes of the propensity score distribution. Specifically, 36 comparison physicians have a propensity score lower than the minimum score for treatment physicians (0.015) and 52 treatment

[31] Our main empirical results are not sensitive to this restriction. Results are available upon request.

physicians have a propensity score higher than the maximum propensity score of comparison physicians (0.968). In our analysis, we impose the common support condition by excluding these 88 physicians, or about one percent of the sample from each tail of the propensity score distribution[32]. In addition, as a specification check, we also exclude an additional *q* percent of treatment physicians for which the propensity score density of the comparison physicians is the lowest.

6.2 Main Results

Table 4 presents our main results. The first row shows the baseline model, which is the local linear regression model using the bi-weight kernel, the bandwidth of 0.1, and the trimming level of five percent. These results indicate that patients enrolled to the FHO physicians are about 8 percent more likely to receive the DMI services than patients enrolled to the FHG physicians. Similarly, physicians practicing in the FHO model are about 12 percent more likely to participate in the DMI than physicians practicing in the FHG model. Both of these effects are statistically significant. In addition, both effects are quite large compared to the pre-treatment means of 22 and 49 percent, respectively.

The remaining panels in Table 4 show the sensitivity of our results to using alternative matching estimators, bandwidth values, kernel functions, and trimming levels. With respect to the alternative estimators, we considered the nearest neighbour matching, with one and ten neighbours, and conventional kernel estimator. In addition, we considered the bandwidth values that are half as large (0.05) and twice as large (0.2) as our baseline value of 0.1. With respect to the kernel functions, the alternatives we considered were the Epanechnikov, Normal, Tri-cube, and Uniform functions. Lastly, we estimated the baseline model with no trimming and with the alternative trimming level of 0.1. Our baseline results are quite robust with respect to these alternative specifications. In each specification, the FHO impact remains positive and statistically significant for both outcomes, and its magnitude is quite similar to our baseline estimates.

[32] Note that this is more than what is required by the overlap assumption in (12), which only requires the exclusion of 52 physicians for which $p(d_{it}=1|X_i)=1$.

6.3 Specification Checks

The CIA can never be directly verified because the counterfactual outcomes in the non-treatment state cannot be observed for any treatment physician. However, we conduct three specification checks to shed some light on the validity of this assumption.

The pre-treatment test, originally proposed by Heckman and Hotz (1989), relies on data on outcomes in the pre-treatment period and knowledge of future treatment status of sample physicians. The test is based on the idea that a consistent estimator applied to the pre-treatment data should make the outcomes of future treatment and comparison physicians similar. The results from this test are presented in the second panel of Table 5. For convenience, the first panel reproduces our baseline results from Table 4. These results indicate that our baseline estimator, the local linear regression, aligns the treatment and comparison physicians quite well in the pre-treatment period. Specifically, the estimated coefficients on both outcomes are small and statistically insignificant, as would be expected if the CIA holds.

The second test is based on the idea that the treatment effect of joining the FHO model, if any, should be similar across successive cohorts of future treatment physicians. The results from this test are presented in the third panel of Table 5. In our sample, there were three main cohorts of physicians joining the FHO model, in 2008 (370 physicians), in 2009 (745 physicians), and in 2010 (406 physicians). To facilitate the comparison of estimates across cohorts, we used the same group of comparison physicians in each model. The results show a positive and significant FHO impact on the percent of diabetic patients receiving the DMI services for all three cohorts of treatment physicians. In addition, the magnitude of this impact is quite similar across the three cohorts and also to our baseline estimate. On the other hand, the estimated impact on the probability of physician participation in the DMI is positive for all three cohorts, with the similar magnitude across cohorts and to our baseline estimates, although the impact is estimated imprecisely for the 2008 cohort. Again, these cohort-specific results are largely consistent with what would be expected if the CIA holds.

Lastly, we examine the sensitivity of our results to the choice of matching variables included in the propensity score model. As mentioned earlier, this choice is critically important for consistently

estimating the ATT[33]. At the same time, this choice is quite difficult because it must simultaneously satisfy the requirements of both the CIA and the common support assumption. In particular, the set of matching variables must be rich enough to ensure that the potential outcomes in the non-treated state (y^0) are similar between the treatment and comparison physicians, but including any additional variables will make the common support assumption more likely to fail. To examine this issue, we estimated our baseline model using the successively richer sets of matching variables. Specifically, we start with the model that includes only variables related to physician characteristics, and then successively add those related to practice characteristics, patient characteristics, the expected income gain, and the past outcomes. The results of this analysis, presented in Table 6, indicate that the estimates are uniformly smaller whenever we use less than the full set of matching variables. At the same time, the estimates are positive and statistically significant in all models, suggesting that our results are not overly sensitive to these permutations of matching variables. Perhaps most significantly, our baseline results presented in Table 4 depend most critically on the inclusion of two past outcomes. In fact, including only the past outcomes in the propensity score model produces estimates nearly identical to our baseline results. This finding is particularly comforting because the past outcomes could contain all the relevant information on the unobservable physician characteristics as they are partly determined by such factors.

6.4 Alternative Empirical Strategies

In addition to the DD matching estimators, we also present results from several alternative estimators that rely on different identification assumptions. First, we estimated the standard DD regression models. These models are based on the same CIA as the DD matching models, but unlike the matching models, they do not explicitly impose the common support condition and they rely on a correct parametric specification of potential outcomes[34]. Second, we implemented the cross-section matching estimator

[33] For example, Heckman and Lozano (2004) show that bias may result if the conditioning set of variables is not the right and complete one. Specifically, if the relevant information is not all controlled for, adding additional relevant information but not all that is required may increase rather than reduce bias.
[34] See for example Blundell and Costa Dias (2009).

using outcomes in the post-treatment period only. This estimator relies on a stronger version of the CIA which requires that matching balances all unobserved characteristics, both fixed and time variant, between the treatment and comparison physicians. Lastly, we estimated the cross-section regression models using outcomes in the post-treatment period only. These models do not account for fixed effects, do not explicitly impose the common support condition, and rely on the correct parametric specification of potential outcomes.

Table 7 shows the estimated FHO impact using these alternative estimators. Importantly, the estimated impact remains positive and significant using all estimators. The difference between the alternative estimators is mainly in the magnitude of the estimated impact. For example, the estimates in the DD regression models are smaller than our baseline estimates, while the estimates in the cross-section regression model are in general larger than our baseline estimates. On the other hand, the estimates from the cross-section matching model are quite similar to our baseline estimates. Overall, these results indicate that the sign of estimated FHO impact is not overly sensitive to the mentioned variations in identification assumptions.

6.5 Sub-group Analysis

Our main results reported in Table 4 represent the average impact of joining the FHO model. In this section, we examine how this impact varies for specific groups of treatment physicians defined by age, gender, location of practice[35], and experience with the new primary care models, using the same baseline DD matching model as in Table 4. The results, presented in Table 8, suggest that there is some heterogeneity in the physician response to the DMI. For example, the FHO impact on both outcomes is somewhat larger for female physicians and physicians over 50 years of age. Despite this heterogeneity,

[35] We differentiate between practices in rural and urban areas. The rurality is measured using the Rurality Index of Ontario (see Kralj, 2000). This index is used by the Ontario Ministry of Health and Long-Term Care in many programs (e.g. Continuing Medical Education) that provide additional incentives to physicians living in rural or remote areas. The index ranges between 0 and 100, where the lower values of this index indicate more urban areas. About half of physicians reside in an area with a RIO score of zero, which we use as a threshold value to differentiate between rural and urban areas.

however, the estimated impact is positive and statistically significant for each physician group and for each outcome. These results suggest that the FHO impact was not restricted to a particular group of physicians.

6.6 Policy Implications

Our empirical results suggest that physicians in a blended capitation model are more responsive to the DMI than physicians in an enhanced fee-for-service model. In terms of our theoretical model, this result implies that the quality effort varies positively with the degree of supply-side cost sharing ($\partial e/\partial r < 0$). Further, given that $p = B_e + (B_q - r)(\partial e/\partial r)/(\partial e/\partial p)$ from equation (8), this result implies that for a given compensation mechanism, the optimal size of P4P incentive varies negatively with the degree of supply-side cost sharing ($\partial p/\partial r > 0$).

The main policy implication of this result is that the design of P4P programs must take into account the underlying physician payment mechanism. Our analysis shows that the P4P incentive should be higher when the degree of cost sharing is lower. This argument must be interpreted with caution, however, because it strictly applies to comparing payment mechanisms with small differences in the cost sharing parameter r. Therefore, it does not necessarily follow from this argument that the P4P incentives in the capitation model should be smaller than in the fee-for-service model. Rather, the argument is more policy relevant when contemplating changes in the size of P4P incentives following small changes in the physician compensation mechanism. For example, the Ontario Ministry of Health and Long-Term Care increased the cost sharing parameter in the FHO model, known as the shadow billing rate, from 0.1 to 0.15 on October 1, 2010. Our analysis implies that this change should be accompanied with a corresponding increase in the DMI and other P4P performance incentives.

A corollary of our results is that the quality bonus p should be set above the marginal impact of quality on patient's health ($p > B_e$) whenever there is an over-provision of medical services ($B_q < r$). Similarly, $p < B_e$ whenever there is an under-provision of medical services ($B_q > r$). The intuition for this

result is simple. The introduction of a P4P program creates incentives to reallocate physician effort from 'quantity' to 'quality'. To the extent that there is an existing distortion in quantity, the quality bonus addresses the twin goals of improving quality and reducing distortions in quantity.

The possible distortions in the quantity of medical services may arise mainly because the policy maker takes the physician compensation mechanism as given when introducing a new P4P program. Clearly, such distortion can be eliminated and welfare improved if the introduction of P4P programs is determined jointly with changes to the physician payment mechanisms. Such a wholesome approach to the healthcare reform may be welfare improving because, as our analysis suggests, there exist important links between the physician response to the P4P programs and the type of physician compensation mechanism.

8. Conclusion

In this study, we compare physician response to the Diabetes Management Incentive (DMI), a new pay-for-performance incentive in Ontario, between physicians practicing in an enhanced fee-for-service model and a blended capitation model. Using a comprehensive administrative data and a difference-in-difference matching strategy, we find that physicians in a blended capitation model are more responsive to the DMI than physicians in an enhanced fee-for-service model. We show that for a given payment mechanism this result implies that the optimal size of P4P incentives is negatively related to the degree of supply-side cost sharing. These results suggest that the optimal design of P4P programs is importantly linked to the physician payment mechanisms. More generally, our analysis suggests that a joint approach to both the physician payment reform and the design of P4P programs may be welfare improving.

Future research can build on our analysis in at least two ways. First, our analysis is based on two types of physician payment mechanisms (enhanced fee-for-service and blended capitation) and a single P4P incentive (the DMI). Future analysis can examine the physician response between these two models to other P4P incentives, such as preventive care bonuses and incentives to enrol new patients. Similarly,

the difference in physician response to the P4P incentives can be studied using other types of physician payment mechanisms, such as the traditional fee-for-service model and salary. Second, we studied the physician uptake of the DMI, which involves a planned, on-going management of diabetic patients using the best practice clinical guidelines. Ultimately, the policy importance of this incentive is its impact on patients' health and healthcare costs. This remains a promising area for future research, following advances already made in the literature[36].

[36] See for example, Dusheiko et al. (2011).

References

Abadie, A. and Imbens, G.W. 2008. On the failure of the bootstrap for matching estimators. *Econometrica* 76: 1537-1557.

Andrews, D.W.K. and Buchinsky, M. 2000. A three-step method for choosing the number of bootstrap repetitions. *Econometrica* 67, 23–51.

Andrews, D.W.K. and Buchinsky, M. 2001. Evaluation of a three-step method for choosing the number of bootstrap repetitions. *Journal of Econometrics* 103, 345–386.

Armour B, Pitts M, Maclean R, et al. 2001. The effect of explicit financial incentives on physician behavior. *Archives of Internal Medicine* 161:1261-1266.

Becker, S. and Ichino, A. 2002. Estimation of average treatment effects based on propensity scores. *The Stata Journal* 2(4): 358-377.

Richard Blundell, R. and Costa Dias, M. 2009. Alternative Approaches to Evaluation in Empirical Microeconomics. *Journal of Human Resources* 44(3): 565-640.

Boadway R, Marchand M, and Sato M. 2004. An optimal contract approach to hospital financing. *Journal of Health Economics* 23(1): 85–110.

Chalkley, M. and Malcomson, J.M., 1998. Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics* 17(1), 1–19.

Christianson, J. B., Leatherman S., and Sutherland K. 2008. Lessons from evaluations of purchaser pay-for-performance programs: A Review of the Evidence. *Medical Care Research and Review*, 65(6): 5S-35.

Dali, T.M., Zhang, Y. Chen Y.J. Quick W.W., Yang, W.G. and Fogli, J. 2010. The economic burden of diabetes. *Health Affairs* 29: 1-7.

Dehejia, R.H. and Wahba, S. 2002. Propensity-score matching methods for nonexperimental causal studies. *The Review of Economics and Statistics* 84(1): 151-161.

Dusheiko, M., Gravelle, H., Martin S., Rice, N., and Smith PC. 2011. Does disease management in primary care reduce hospital costs? Evidence from English primary care. *Journal of Health Economics* 2:30(5):919-32.

Eggleston, K. 2005. Multitasking and mixed systems for provider payment. *Journal of Health Economics* 24, 211–223.

Fan J. 1992. Design adaptive nonparametric regression. *Journal of the American Statistical Association* 87: 998-1004.

Fan J. 1993. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21: 196-216.

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.

Glazier, R.H., Klein-Geltnik, J., Kopp, A., and Sibley, L.M. 2009. Capitation and enhanced fee-for-service models for primary care reform: a population-based evaluation. *Canadian Medical Association Journal* 180(11): E72-E81.

Ham, J.C., Li, X., and Reagan, P.B. 2011. Matching and semiparametric IV estimation, a distance-based measure of migration, and the wages of young men. *Journal of Econometrics* 161(2): 208-227.

Heckman, J. and Hotz, J., 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *Journal of the American Statistical Association* 84 (408), 862–880.

Heckman, J., Ichimura, H., and Todd, P. 1997. Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program. *The Review of Economic Studies* 64(4): 605-654.

Heckman, J., Ichimura, H., and Todd, P. 1998. Matching as an econometric evaluation estimator. *The Review of Economic Studies* 65 (2), 261–294.

Heckman, J. And Lozano, S. 2004. Using matching, instrumental variables and control functions to estimate economic choice models. *The Review of Economics and Statistics* 86(1): 30-57.

Holmstrom, B. and Milgrom, P., 1991. Multitask principal-agent analyses: incentive contracts, job design. *Journal of Law, Economics and Organization* 7 (special issue), 24–52.

Institute for Clinical Evaluative Studies. 2003. Diabetes in Ontario – an ICES practice atlas. Available at: www.ices.on.ca/file/DM_Intro.pdf.

Imbens, G.W. and Wooldridge, J.M. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1): 5–86.

Integrated Healthcare Association. 2006. Advancing quality through collaboration: The California pay for performance program. www.iha.org/pdfs_documents/p4p_california/P4PWhitePaper1_February2009.pdf.

International Diabetes Federation. 2010. Annual report. Available at: www.idf.org/sites/default/files/Annual-Report-2010-FINAL-EN_0.pdf

Kaarboe, O. and Siciliani, L. 2011. Multi-tasking, quality and pay for performance. *Health Economics* 20: 225-238.

Kantarevic, J., Kralj, B. and Weinkauf, D. 2011. Enhanced Fee-for-Service Model and Physician Productivity: Evidence from Family Health Groups in Ontario**.** *Journal of Health Economics*, 30(1): 99-111.

Kantarevic, J. and Kralj, B. Quality and Quantity in Primary Care Mixed Payment Models: Evidence from Family Health Organizations in Ontario. *Canadian Journal of Economics* (forthcoming).

Kralj, B. 2000. Measuring 'rurality' for purposes of health-care planning: an empirical measure for Ontario. *Ontario Medical Review*.

Léger, P.T. 2008. Physician Payment Mechanisms. In: Financing Health Care: New Ideas for a Changing Society, Lu, M. and Jonsson, E. (Eds.), Wiley, pp. 149-176.

Leuven, E. and Sianesi, B. 2003. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. http://ideas.repec.org/c/boc/bocode/s432001.html. Version 3.1.5.

Li, J., Hurley, J., DeCicca, P., and Buckley, G. 2011. Physician response to pay-for-performance – evidence from a natural experiment. NBER Working Paper 16909.

McGuire, T.G., 2000. Physician agency. In: Culyer, A.J., Newhouse, J.P. (Eds.), *Handbook of Health Economics*, vol. 1A. North-Holland, Amsterdam, pp. 461–536.

Ontario Ministry of Health and Long Term Care, Schedule of Benefits, Physician Services under the Health Insurance Act (September 1, 2011).

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

Rosenbaum, P.R. and Rubin, D.B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41-55.

Rosenbaum, P.R. and Rubin, D.B. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39(1): 33-38.

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.

Silverman, B., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.

Smith, P.C. and York, N. 2004. Quality incentives: the case of U.K. general practitioners. *Health Affairs* 23(3): 112-118.

Smith, J. and Todd, P. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* 125: 305-353.

Town R., Kane R., Johnson P., and Butler M. 2005. Economic incentives and physicians' delivery of preventive care: a systematic review. *American Journal of Preventive Medicine* 28(2):234-240

Table 1.
Variable definitions and descriptive statistics

Variable name	Variable description	Whole Sample	Treatment Sample	Comparison Sample	Difference
N	Sample size (number of physicians)	3,655	1,521	2,134	
Treat	=1 if in FHO model in 2010, =0 if in FHG in 2010	0.4161	1	0	
Intensive_2006	Percent of enrolled DM patients with Q040 claim, 2006	0.22	0.27	0.19	0.08
		(0.30)	(0.31)	(0.28)	(0.01)
Intensive_2010	Percent of enrolled DM patients with Q040 claim, 2010	0.34	0.42	0.29	0.13
		(0.31)	(0.31)	(0.29)	(0.01)
Extensive_2006	Percent of physicians with any Q040 claims, 2006	0.49	0.57	0.44	0.13
		(0.50)	(0.50)	(0.50)	(0.02)
Extensive_2010	Percent of physicians with any Q040 claims, 2010	0.68	0.81	0.60	0.21
		(0.46)	(0.39)	(0.49)	(0.02)
Age	Physician age (in years), in 2006	49.73	48.48	50.63	-2.16
		(9.57)	(9.17)	(9.75)	(0.32)
Male	=1 if male physician	0.64	0.63	0.64	-0.01
		(0.48)	(0.48)	(0.48)	(0.02)
Toronto LHIN	=1 if physician resides in Toronto region, in 2006	0.12	0.11	0.13	-0.02
		(0.33)	(0.31)	(0.34)	(0.01)
Roster	Number of enrolled patients, April 1, 2006	860.7	949.3	797.5	151.8
		(530.0)	(519.1)	(528.8)	(17.6)
Share DM	Percent of enrolled patients with DM, April 1, 2006	0.08	0.07	0.08	-0.01
		(0.05)	(0.03)	(0.05)	(.002)
Age-sex modifier	Risk-adjustment factor based on age and sex, April 1, 2006	1.13	1.14	1.13	0.003
		(0.14)	(0.12)	(0.16)	(.005)
Income gain	Potential gain from switching to FHO (C$), in 2006	14,857	57,185	-15,312	72,496
		(100,552)	(81,573)	(101,935)	(3,154)
Group size	Number of physicians in FHG group, April 1, 2006	45.9	34.2	54.2	-20.1
		(60.2)	(46.2)	(67.3)	(2.0)
FHG days	Number of years since joining FHG model, April 1, 2006	1.5	1.7	1.4	0.3
		(0.8)	(0.8)	(0.8)	(0.03)
Visits	Number of annual visits, fiscal 2006/7	7,538	7,104	7,846	-742
		(3,588)	(3,197)	(3,813)	(120)

Table 2.
Propensity score logit estimates for participation in FHO

Variable	Coefficient	Standard Error
Age	0.0013	0.0386
Age^2	-0.0005	0.0004
Age×Male	0.0151	0.0106
Male	-0.1722	0.5173
Roster	-0.2777	0.2667
$Roster^2$	-0.0003***	0.0001
Visits	0.1241**	0.0548
$Visits^2$	-0.1660	0.2650
Group size	-0.0010	0.0008
Income gain	0.0126***	0.0010
Income $gain^2$	0.0008**	0.0004
Age-sex modifier	8.7542***	2.7239
Age-sex $modifier^2$	-3.0855***	1.0858
FHG days	0.0008***	0.0002
Share DM	-6.5936***	1.1581
Intensive_2006	0.2085	0.2028
Extensive_2006	0.2624**	0.1173
Constant	-5.5839***	1.8782

Notes:
(1) *** significance at 1% level, ** significance at 5% level, and * significance at 10% level.
(2) To improve readability, the coefficients on Roster, $Roster^2$, Visits, and Income gain have been multiplied by 10^3 and the coefficients on $Visits^2$ and Income $gain^2$ by 10^8.
(3) The model also includes 14 indicators for Local Health Integration Networks (LHINs).
(4) The sample size is 3,588 physicians.
(5) The likelihood ratio chi-square statistic is 1,194 with 30 degrees of freedom. The McFadden's pseudo R^2 is 0.24.

Table 3.
Balancing Tests

Paired *t* tests

	Difference: Unmatched	Difference: Matched	p-value of paired *t* statistics
Age	-2.16	-0.07	0.85
Male	-0.01	0.02	0.17
Toronto	-0.02	0.01	0.25
Roster	152	-41	0.08
Visits	-742	-215	0.07
Share DM	-0.01	-0.001	0.46
Age-sex modifier	0.00	-0.001	0.74
Income gain	72,496	-958	0.74
Group size	-20.1	0.6	0.72
FHG days	109	3.0	0.78
Intensive_2006	0.08	-0.01	0.61
Extensive_2006	0.13	-0.01	0.43

F test statistics

	Sample size	F-statistic	p-value
1st quintile	536	1.71	0.06
2nd quintile	596	0.46	0.94
3rd quintile	594	0.33	0.98
4th quintile	595	1.30	0.21
5th quintile	522	1.55	0.10

Notes:
(1) All tests are based on nearest neighbour matching. The unmatched difference is the difference between the full sample of treatment and comparison physicians for each covariate, while the matched differences are for the full sample of treatment physicians and only the matched sample of comparison physicians.

Table 4.
Difference-in-Difference Matching Estimates of FHO Impact

	Enrolled Diabetic Patients with DMI	Physicians with DMI
Baseline Model	0.0843***	0.1153***
	(0.0146)	(0.0223)
Alternative Estimators		
Nearest Neighbour (1 neighbour)	0.0972***	0.1271***
	(0.0182)	(0.0284)
Nearest Neighbour (10 neighbours)	0.0847***	0.1108***
	(0.0152)	(0.0236)
Kernel	0.0803***	0.1147***
	(0.0135)	(0.0216)
Alternative Bandwidth Values		
0.05	0.0846***	0.1142***
	(0.0142)	(0.0255)
0.20	0.0836***	0.1086***
	(0.0148)	(0.0232)
Alternative Kernel Functions		
Normal	0.0815***	0.1076***
	(0.0145)	(0.0231)
Uniform	0.0839***	0.1127***
	(0.0149)	(0.0227)
Epanechnikov	0.0841***	0.1145***
	(0.0147)	(0.0224)
Tricube	0.0830***	0.1095***
	(0.0148)	(0.0234)
Alternative Trimming Levels		
No trimming	0.0818***	0.1067***
	(0.0135)	(0.0219)
10 percent	0.0760***	0.1157***
	(0.0130)	(0.0214)

Notes:
(1) The baseline model is the local linear regression model, using the bi-weight kernel, the bandwidth of 0.1, and imposing a common support by dropping treatment observation whose propensity score is higher than the maximum or less than the minimum propensity score of the comparison physicians and by dropping 5 percent of the treatment observations at which the propensity score density of the comparison observations is the lowest.
(2) The sample size is 3,588 physicians.
(3) Bootstrap standard errors in parentheses, using 200 bootstrap repetitions.
(4) *** significance at 1% level, ** significance at 5% level, * significance at 10% level.

Table 5.
Pre-treatment Impact and Impact by Year of Switch

	Sample	Enrolled Diabetic Patients with DMI	Physicians with DMI
Baseline Model	3,588	0.0843*** (0.0146)	0.1153*** (0.0223)
Pre-treatment Impact	3,595	-0.0005 (0.0233)	-0.0141 (0.0299)
Impact by year of switch			
2008 cohort	2,444	0.1018***	0.0899
		(0.0330)	(0.0843)
2009 cohort	2,816	0.0691***	0.1095***
		(0.0164)	(0.0247)
2010 cohort	2,482	0.0762**	0.1176**
		(0.0173)	(0.0280)

Notes:
(1) For the baseline model, see note (1) in Table 4.
(2) The pre-treatment impact specification uses the outcomes in and the baseline matching model.
(3) The cohort for each year represents treatment physicians who joined FHO in that year. The set of comparison physicians is the same for each of the cohort models. Each row represents a separate model for each cohort.
(4) Bootstrap standard errors in parentheses, using 200 bootstrap repetitions.
(5) *** significance at 1% level, ** significance at 5% level, * significance at 10% level.

Table 6.
Choice of Matching Variables

Set of Matching Variables	Enrolled Diabetic Patients with DMI	Physicians with DMI
(1) Physician Characteristics	0.0585***	0.0788***
	(0.0114)	(0.0170)
(2) Practice Characteristics + (1)	0.0567***	0.0650***
	(0.0134)	(0.0224)
(3) Patient Characteristics + (2)	0.0488***	0.0429*
	(0.0138)	(0.0238)
(4) Expected Income Gain + (3)	0.0630***	0.0544*
	(0.0245)	(0.0332)
(5) Past Outcomes + (4)	0.0843***	0.1153***
	(0.0146)	(0.0223)

Notes:
(1) For the baseline model, see note (1) in Table 4.
(2) Physician characteristics include age, sex, and days in FHG model as of April 1, 2006; practice characteristics include geographic location, number of enrolled patients, number of annual visits, and group size; patient characteristics include the average age-sex modifier and the percent of enrolled patients that are diabetic; the expected gain is the calculated income gain from switching from the FHG to FHO model; and past outcomes include the percent of enrolled diabetic patients that receive DMI in 2006 and the percent of physicians participating in the DMI in 2006. Further details are available in Section 5.
(3) Bootstrap standard errors in parentheses, using 200 bootstrap repetitions.
(4) *** significance at 1% level, ** significance at 5% level, * significance at 10% level.

Table 7.
Estimates of FHO Impact from Alternative Empirical Strategies

Estimation Model	Enrolled Diabetic Patients with DMI	Physicians with DMI
Baseline Model	0.0843***	0.1153***
	(0.0146)	(0.0223)
Regression Difference-in-Difference Model		
Pooled OLS	0.0575***	0.0800***
	(0.0098)	(0.0167)
Fixed Effects	0.0578***	0.0805***
	(0.0099)	(0.0169)
Cross-Section Estimators		
LLR Matching	0.0859***	0.0939***
	(0.0200)	(0.0272)
OLS - controlling for age, sex, location only	0.1081***	0.1791***
	(0.0109)	(0.0157)
OLS - controlling for all covariates in 2006	0.1046***	0.1685***
	(0.0113)	(0.0165)

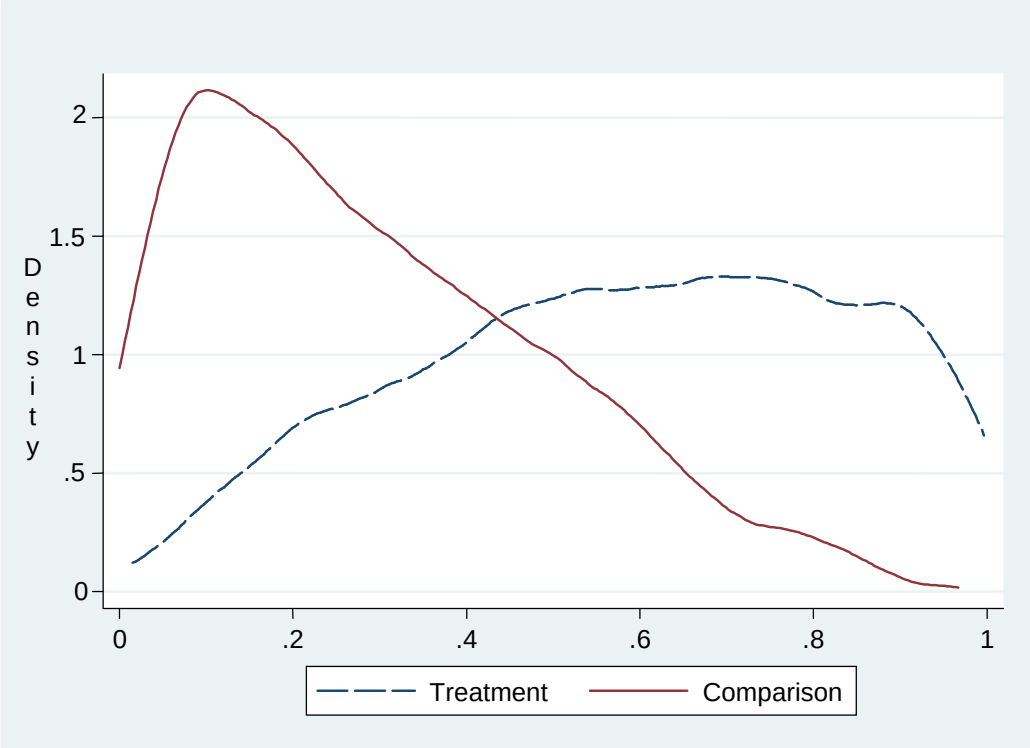
Notes:
(1) For the baseline model, see note (1) in Table 4.
(2) The pooled OLS and fixed effects DD models also include controls for age, age squared, male, age-male interaction, and 14 indicators for Local Health Integration Networks (LHINs). Robust standard errors, clustered at the physician level.
(3) The LLR cross-section matching estimator uses the same propensity score model as the baseline model in Table 4. Bootstrap standard errors, based on 200 bootstrap repetitions.
(4) The available covariates in 2006 in the OLS cross-section model include all variables used in the propensity score model described in Table 2. Robust standard errors in parentheses.
(5) *** significance at 1% level, ** significance at 5% level, * significance at 10% level.

Table 8.
Estimates of FHO Impact by Sub-groups

	Sample	Enrolled Diabetic Patients with DMI	Physicians with DMI
Baseline Model	3,588	0.0843***	0.1153***
		(0.0146)	(0.0223)
Female	1,298	0.0951***	0.1153***
		(0.0247)	(0.0380)
Male	2,284	0.0683***	0.1076**
		(0.0237)	(0.0461)
Age < 50	1,755	0.0688***	0.1162*
		(0.0235)	(0.0702)
Age ≥ 50	1,827	0.1076***	0.1248***
		(0.0242)	(0.0340)
Urban	1,749	0.0557***	0.1210***
		(0.0186)	(0.0328)
Rural	1,839	0.0977***	0.0914**
		(0.0209)	(0.0423)
In FHG < 18 months	1,808	0.0895**	0.1102**
		(0.0311)	(0.0489)
In FHG ≥ 18 months	1,774	0.0807**	0.1285**
		(0.0178)	(0.0363)

Notes:
(1) For the baseline model, see note (1) in Table 4.
(2) The rural/urban distinction is determined using the Rural Index of Ontario (RIO) score of 0 (>0 for rural, =0 for urban). See footnote (36) in Section 6.5 for more details on the RIO.
(3) The length in the FHG model is measured as the number of months between April 1, 2006 and the date that the physician joined a FHG model.
(4) Bootstrap standard errors in parentheses, using 200 bootstrap repetitions.
(5) *** significance at 1% level, ** significance at 5% level, * significance at 10% level.

Figure 1.
Distribution of estimated propensity scores



Appendix A: Selecting the Bandwidth Value

The Silverman's rule of thumb method consists of finding the bandwidth value that minimizes the mean squared integrated error for the kernel density estimator and then replacing the unknown quantities with an estimate. For example, taking the Gaussian kernel (which is identical to the standard normal pdf), the rule of thumb expression for the optimal bandwidth value is$1.06\hat{\sigma}n^{-1/5}$, where *n* is the sample size. The results of implementing this method in our study are given in the table below.

Table A. Silverman's Rule of Thumb Bandwidth Selector with Gaussian Kernel

	Enrolled Diabetic Patients with DMI	Physicians with DMI
Sample size	3,598	3,655
St.Dev.	0.2829	0.4913
Optimal bandwidth	0.0583	0.1009

Appendix B: Selecting the Number of Bootstrap Repetitions

We closely follow Ham et al. (2011) in implementing the methodology developed by Andrews and Buchinsky (2000, 2001). Let θ be the ATT parameter identified by the matching estimator and let λ be its standard error. Further, let *B* denote the number of repetitions and *pdb* the measure of accuracy, which is the percentage deviation of the bootstrap quantity of interest based on bootstrap repetitions from the ideal bootstrap quantity of interest for which $B = \infty$. The magnitude of *B* depends on both the accuracy required and the data. If we required the actual percentage deviation to be less than *pdb* with a specified probability 1-τ, then the Andrews and Buchinsky propose a three-step method that takes *pdb* and τ as given and provides a minimum number of repetitions *B** to obtain the desired level of accuracy. We follow Ham et al. (2011) and set (*pdb*, τ)=(10,0.05). In the first step, we calculate the initial number of repetitions B_1= int(10,000∗$z^2_{1-\tau/2}$∗0.5/pdb^2), where $z_{1-\tau/2}$ is the 1-τ/2 quantile of standard normal distribution. In our case, B_1=193. In the second step, we use the bootstrap results $\{\hat{\theta}\colon \hat{\theta}_1, \dots, \hat{\theta}_{B1}\}$ to calculate ω=(2+γ_B)/4, where $\gamma_B = (B_1 - 1)^{-1} \sum_r^{B1} \left(\hat{\theta}_r - \mu_B\right)^4 / se_B^4 - 3$, where μ_B and se_B are the mean and standard deviation of $\{\hat{\theta}\colon \hat{\theta}_1, \dots, \hat{\theta}_{B1}\}$. The new number of repetitions is then calculated as B_2= int(10,000∗$z^2_{1-\tau/2}$∗ω/pdb^2). In the last step, the minimum number of repetitions is determined as *B**=max(B_1,B_2). The results of implementing this methodology in our study are given in the table below.

Table B. Andrews and Buchinsky (2000, 2001) Method for Selecting Bootstrap Repetitions

	Estimated B_2		Optimal Number B^*	
Matching Model	Enrolled Diabetic Patients with DMI	Physicians with DMI	Enrolled Diabetic Patients with DMI	Physicians with DMI
LLR	193	233	193	233
Kernel	191	183	193	193
NN	212	189	212	193

Note: LLR stands for Local Linear Regression and NN stands for Nearest Neighbour.

Appendix C: Expected Income Gain

In our analysis, we estimated the expected difference in income for a cohort of FHG physicians in 2006 between what they actually earned in 2006 and what they would hypothetically earned if they practiced in the FHO model. The actual base compensation for these physicians can be represented as $I_{FHG} = 1.1p_1q_1m + p_1q_1n + p_2q_2(m+n)$, where q_1 represents services eligible for the 10 percent comprehensive care premium, q_2 represents all other services, m is the number of enrolled patients, and n is the number of non-enrolled patients. In contrast, the hypothetical income for these physicians if they practiced in the FHO model can be represented as $I_{FHO} = Rm + 0.1p_1q_1m + p_2q_2(m+n) + \min\{p_1q_1n, z\}$, where R is the age-sex adjusted capitation rate, p_1 and q_1 are now the price and quantity of services included in the capitation basket, p_2 and q_2 are the price and quantity of services outside the basket, z is the hard cap on the basket services provided to non-enrolled patients, and, as before, m is the number of enrolled patients and n is the number of non-enrolled patients. To estimate this hypothetical base compensation in the FHO model, we used the actual profile of services provided to each patient in fiscal 2006 and the list of enrolled patients as of April 1, 2006. For R, we used the base rate of C$144.08 multiplied by the age-sex specific modifier for each enrolled patient. These modifiers include 19 five-year age categories for each sex and range from 0.44 for males 10-14 years of age to 2.71 for females over 90 years of age, with the provincial average standardized to 1. To identify q_1 and q_2, we used the list of over 100 fee codes specified in the FHO contract. Lastly, for z we used the actual value of C$47,500 that applied in fiscal 2006.