HEALTH ECONOMICS
Health Econ. **25**: 1326–1340 (2016)
Published online 4 August 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hec.3220

PHYSICIAN PAYMENT CONTRACTS IN THE PRESENCE OF MORAL HAZARD AND ADVERSE SELECTION: THE THEORY AND ITS APPLICATION IN ONTARIO

JASMIN KANTAREVIC[a,b,c,d,*] and BORIS KRALJ[a]

[a]*Ontario Medical Association, Toronto, Canada*
[b]*Canadian Centre for Health Economics, Toronto, Canada*
[c]*Institute for Labor Studies, Bonn, Germany*
[d]*University of Toronto, Toronto, Canada*

ABSTRACT

We develop a stylized principal–agent model with moral hazard and adverse selection to provide a unified framework for understanding some of the most salient features of the recent physician payment reform in Ontario and its impact on physician behavior. These features include the following: (i) physicians can choose a payment contract from a menu that includes an enhanced fee-for-service contract and a blended capitation contract; (ii) the capitation rate is higher, and the cost-reimbursement rate is lower in the blended capitation contract; (iii) physicians sort selectively into the contracts based on their preferences; and (iv) physicians in the blended capitation model provide fewer services than physicians in the enhanced fee-for-service model. Copyright © 2015 John Wiley & Sons, Ltd.

Received 15 October 2014; Revised 23 June 2015; Accepted 29 June 2015

JEL Classification: I10; I12; I18

KEY WORDS: physician remuneration; moral hazard; adverse selection; Ontario

1. INTRODUCTION

In the early 2000s, Ontario launched a major primary healthcare reform. One of the unique features of this reform is that it introduced a menu of payment contracts, rather than a single contract, in which physicians could choose to participate. In contrast to the traditional fee-for-service (FFS) contract, under which physicians receive a fee for each service they provide, the new contracts blend prospective (per patient) and retrospective (per service) payments to varying degrees. Understanding the impact of this reform has been the focus of much recent empirical research (Glazier *et al*, 2009; Kantarevic *et al*., 2011; Kralj and Kantarevic, 2013; Li *et al*., 2014; Rudoler *et al*., 2014). However, this literature still lacks a unified framework that can explain seemingly unrelated empirical regularities and provide a rationale for the main features of the reform, such as the existence of a menu of contracts and the specific blend of prospective and retrospective elements in each contract.

In this paper, we take the first few steps needed to fill this gap. In the first part, we start by documenting some of the most salient regularities regarding the new payment contracts. These include the menu and shape of contracts, which are defined by the institutional framework, and the selection and incentive effects of contracts, which we estimate using difference-in-difference methodology with propensity score matching. Specifically, physicians in Ontario can choose to participate in two main types of contracts: an enhanced fee-for-service (EFFS hereafter) contract and a blended capitation contract. In both contracts, physicians receive a payment per patient and a payment per service; however, the payment per patient is lower, and the payment

*Correspondence to: Economics, Ontario Medical Association, Toronto, Canada. E-mail: jasmin.kantarevic@oma.org

per service is higher in the EFFS than in the capitation contract. Further, our empirical results suggest that physicians with a lower number of services tend to choose the capitation contract (the selection effect), and that the lower payment per service in the capitation contract tends to reduce further the number of services these physicians provide (the incentive effect). In the second part, we then develop a stylized principal–agent model with moral hazard and adverse selection in which these regularities arise as a solution to the problem of designing optimal payment contracts when neither physician type nor physician action is contractible.

The specific contribution of this paper is therefore to document a set of regularities about the physician payment reform in Ontario and to develop a relatively standard economic model to interpret this reform. More generally, the paper contributes to our understanding of how physicians respond to payment incentives. Traditionally, interest has centered on comparing the two main methods of payment – FFS and capitation – in terms of their impact on the access, quality, and cost of health care (Léger, 2008; McGuire, 2000; Scott, 2000; Zweifel *et al*., 2009). The conclusions from this literature are mostly based on a principal–agent model with moral hazard in which physician actions cannot be observed or verified (Ellis and McGuire, 1986, 1990; Ma, 1994; Ma and McGuire, 1997). More recent studies have also considered the adverse selection framework that focuses on unobserved heterogeneity among physicians such as ability, altruism, and efficiency (Chone and Ma, 2007; Makris and Siciliani, 2013). Lastly, there is a small but growing literature, to which our study contributes, that combines both moral hazard and adverse selection to study the design of optimal contracts and their impact on physician behavior (Allard *et al*., 2011, 2014; Jack, 2005).

Our study most closely parallels the analysis of the UK fundholding scheme by Jack (2005) and Dusheiko *et al*. (2006). Jack develops a model with moral hazard and adverse selection and concludes that the optimal menu of contracts includes a set of cost-sharing rules resembling that available to primary care physicians in the UK, whereas Dusheiko *et al*. use difference-in-difference methodology to study the incentive and selection effects of the fundholding scheme on provider behavior. The main difference in our paper is that we focus on access to care as the main policy objective, whereas Jack focuses on the cost of non-physician inputs and quality of care, and we also augment the difference-in-difference approach of Dusheiko *et al*. with propensity score matching.[1] Nevertheless, the analytical tools used are sufficiently similar to testify to their value in understanding the design of payment reforms and their impact on physician behavior in diverse healthcare environments.[2]

The rest of the paper proceeds as follows. In Section 2, we describe the payment contracts available to primary care physicians in Ontario and document selected regularities that we study in this paper. In Section 3, we set up the model and use it to explain these regularities. We conclude in Section 4.

2. REGULARITIES ABOUT PHYSICIAN PAYMENT CONTRACTS IN ONTARIO

Prior to the physician payment reform in Ontario, two main policy concerns were access to primary health care and the physician practice style. The issue with access was not only the chronic shortage of family physicians in the province, which was addressed subsequently in large part by increasing enrollment in medical schools, but also limited access to physicians during evening and weekends. The issue with the practice style was that almost all physicians practiced in the traditional FFS model, which was often criticized because of its excessive focus on the volume-based acute care and the lack of incentives to form teams with other physicians and healthcare providers (Léger, 2008; McGuire, 2000).

The reform addressed the two problems simultaneously by linking significant financial incentives to physician participation in the new models of care (Buckley and Sweetman, 2014; Rudoler *et al*., 2014, 2015). These

[1]A more detailed comparison with Jack (2005) is provided in Section 3.

[2]Allard *et al*. (2011, 2014) study how heterogeneous physicians choose between payment mechanisms and the impact of this choice on their treatment and referral behavior. We also study the selection and incentive effects of similar types of contracts, although on different outcomes, but we additionally examine the design of the optimal menu of contracts and their shape, which is taken as a given in these two studies.

DOI: 10.1002/hec

models centered on patient enrollment, comprehensive and preventative care, chronic disease management, after-hours access, and group practices with interdisciplinary teams, with significant premiums and bonuses linked to pay for performance targets.

As mentioned earlier, understanding the impact of this reform has been a focus of much recent empirical research. We contribute to this literature by developing a unified framework to study selected regularities of the reform, which we describe in the following sections.

2.1. Menu and shape of payment contracts

The first regularity we study is that the primary care reform introduced two main types of contracts rather than a single contract (Table I). The harmonized models, such as the Family Health Network and the Family Health Organization, are blended *capitation* models (capitation models hereafter). The non-harmonized models, such as the Family Health Groups and the comprehensive care model, are EFFS models. Currently, about two-thirds of primary care physicians in Ontario have chosen to participate in these two payment models.

The second regularity relates to the shape of payment contracts in EFFS and capitation models. In each model, the contract consists of two main elements: a payment per enrolled patient (the capitation rate) and a payment per service (the cost-reimbursement rate). Specifically, in the EFFS models, the capitation rate consists of a comprehensive care management (CCM) fee, equal in terms of annual value to about one regular office visit, whereas in the capitation models, the capitation rate includes both the CCM fee and an age–sex adjusted capitation payment, equal in terms of annual value to about five regular office visits (Table II). On the other hand, in the capitation models, physicians receive 15% of the FFS value of services (for core services provided in the capitation basket), whereas physicians in the EFFS models receive 100% of the FFS value services provided. Therefore, the capitation rate is higher in the capitation models, and the cost-reimbursement rate is higher in the EFFS models. In other words, the prospective payment is higher and the retrospective payment is lower in the capitation models than in the EFFS models.

These two regularities are similar to the fundholding scheme that was available to general practitioners in the UK from 1991 to 1999. In this scheme, physician practices were also able to choose from a menu of two contracts. In the fundholding contract, physician practices were given a budget from which to finance selected non-emergency hospital-delivered secondary care, and unused money could be used to purchase new equipment or other services (Dusheiko *et al*., 2006). The other contract was the status quo, in which physician practices neither bore the costs of secondary care directly nor appropriated any savings. Further, the shape of the contracts was also similar to Ontario's in the sense that the prospective payment (i.e., the budget) was larger and the cost-reimbursement rate lower for the fundholding contract than for the status quo contract. Therefore, in both the Ontario and UK payment models, there is a negative relationship between the prospective and retrospective payment elements across different contracts.

2.2. Selection and incentive effects of payment contracts

The last two regularities relate to the impact of payment contracts on physician behavior. Specifically, differences between the contracts may impact both the physician's choice of which contract to join (the selection effect) and his or her decision on how to practice (the incentive effect). Discerning these two effects is an

Table I. New primary care payment models in Ontario, April 2014

Payment model	Year of introduction	Physicians	% of family physicians
Harmonized (blended capitation)			
Family Health Network	2002	269	2
Family Health Organization	2007	4591	36
Non-harmonized (enhanced fee for service)			
Family Health Group	2003	2749	21
Comprehensive care model	2005	333	3

DOI: 10.1002/hec

Table II. Comparison of elements in patient enrollment models

	Harmonized models	Non-harmonized models
Compensation elements		
Fee-for-service billings (cost-reimbursement rate)	15%	100%
Capitation		
Comprehensive care management[a]	C$30	C$30
Age–sex adjusted capitation rate[b]	C$170	C$0
Incentives and bonuses[c]	Yes	Yes
Organizational elements		
Group size	≥3	≥ 3
Patient enrollment	Yes	Yes
After-hours requirement	Yes	Yes

[a]Approximate rate per patient per year as of 1 April 2014, which is then age–sex adjusted.
[b]Approximate gross rate per patient per year, as of 1 April 2014, which is then age–sex adjusted.
[c]Incentives and bonuses include preventative care bonuses (pap smears, mammograms, childhood immunizations, flu shots, and colorectal screening), special payments (obstetrical deliveries, hospital services, palliative care, prenatal care, and home visits), chronic disease management fees (diabetes and congestive heart failure), and incentives to enrol unattached patients.

empirical issue that has been discussed to some extent in the literature (Kralj and Kantarevic, 2013; Li *et al*., 2014), and in this paper, we confirm these findings using similar empirical strategies applied to a new data set.

2.2.1. Sample. The sample consists of a cohort of all 3641 physicians who participated in the Family Health Groups (the EFFS model) in the fiscal year 2006/2007. Of this cohort, 1563 physicians (43%) remained in the same model by fiscal year 2013/2014, whereas 2078 physicians (57%) switched to the Family Health Organizations (the blended capitation model).

2.2.2. Outcomes. The aspect of physician practice we study is volume based: the number of services and visits per day. Admittedly, this is not the only or even the most important aspect of physician practice. Nevertheless, it is still an important aspect in a healthcare system such as Ontario's in the late 1990s, which was characterized by a shortage of physicians, long wait times, and a significant number of patients with no regular family doctor. In addition, the quantitative aspect of physician practice tends to be accurately measured through administrative claims databases, such as the Ontario Health Insurance Plan claims database that we use in this study.

2.2.3. Empirical strategy. We wish to compare the outcomes of interest (visits and services) between the capitation and EFFS physicians. A simple comparison of outcomes in the fiscal year 2013/2014 is unlikely to produce an unbiased estimate of the incentive effect of contracts given that physician participation in the models is voluntary. This comparison can be improved if we compare changes in outcomes between 2006/2007 and 2013/2014 for the two groups of physicians: those who stayed in the EFFS model (the control group of stayers) and those who switched to the capitation model (the treatment group of switchers). Further improvement can be made if the comparison of changes in outcomes between the two groups is conditional on the covariates that are likely to vary between the two groups and that also determine their selection of the model. Specifically, we use the difference-in-difference matching estimator (Blundell *et al*., 2004; Dehejia and Wahba, 2002; Leuven and Sianesi, 2003; Nichols, 2007; Rosenbaum and Rubin 1983, 1985) that can be represented by the following general form:

$$A\hat{T}T = n^{-1}\sum_i \left\{ \Delta y_{it} - \sum_j w(i,j)\Delta y_{it} \right\}$$

where y denotes the outcome of interest, i and j denote, respectively, the treatment and control physicians in the region of common support, n is the number of physicians in the region of common support, and $w(i, j)$ are the matching weights obtained through propensity score matching. This empirical strategy will identify the incentive effect given two main assumptions. The first assumption, known as the conditional independence

DOI: 10.1002/hec

Table III. Summary statistics, fiscal year 2006/2007

	Full sample	Switchers[a]	Stayers[b,c] (all)	Stayers[d] (matched)
Number of physicians	3641	2078	1563	1350
Services per day	43.1	39.8	47.6*	39.6
Visits per day	29.6	27.9	31.9*	28.0
Average age	49.2	48.4	50.2*	48.9
Percent male	63.3	62.1	64.8	59.8
Percent in Toronto Central Region	12.3	12.8	11.5	12.6
Expected income gain (C$)	11,095	44,456	−33,257*	35,616

[a]Includes physicians who were in the Family Health Group as of the fiscal year 2006/07 but switched to the Family Health Organization by fiscal year 2013/14.
[b]Includes physicians who were in the Family Health Group model in both the 2006/07 and in the 2013/14 fiscal years.
[c]*Indicates that the difference from the FHO group is significant at 0.05 level with the two-tail *t*-test. The *t*-tests are based on a regression of each variable on the treatment indicator. Before matching, this is an un-weighted regression on the whole sample; after matching, the regression is weighted by using the propensity score weights obtained from the local linear regression model with the bi-weight kernel and a bandwidth of 0.2.
[d]Includes physicians who were in the Family Health Group model in the fiscal years 2006/07 and 2013/14 that were matched on the basis of the propensity score to the group of switchers.

assumption, is that depending on the propensity score,[3] the mean change in outcomes for the treatment and control group is identical. Although this is a strong assumption, its plausibility in our study derives from the fact that it only needs to hold after unobserved time-invariant individual characteristics that affect both treatment and outcomes have been settled. The second assumption, known as the common support assumption, requires a positive probability of observing control physicians at each level of covariates. In our estimation, we ensure that this assumption is met by excluding physicians whose propensity score falls outside the region of common support.

2.2.4. Estimation. The alternative difference-in-difference matching estimators differ in how they construct the matching weights. Because of its desirable properties (Fan, 1992, 1993), we use the local linear (LL) kernel as our baseline estimator, and we also provide the results using the alternative nearest neighbor and conventional kernel estimator. The LL kernel requires a specification of the kernel function and the bandwidth. In our baseline model, we use the bi-weight kernel and the bandwidth value of 0.1, and we also examine the robustness of our results to alternative specifications.[4] Lastly, we estimate the standard errors using bootstrapping with 200 replications.[5] The bootstrapping method is expected to work well for the kernel and LL kernel matching estimator, but it is in general not valid for the nearest neighbor (Abadie and Imbens, 2008). As a robustness check, we also estimate the standard errors using the methods described in Abadie and Imbens (2006, 2008).

2.2.5. Results. The descriptive statistics of the sample, as of 2006/2007 when all physicians practiced in the EFFS model, are presented in Table III.

These results clearly indicate that the stayers provided significantly more services and visits than the switchers. This evidence of selective sorting across models is further confirmed by the fact that the expected income gain for the switchers was about C$45,000, whereas it was about −C$33,000 for the stayers.[6] The table also indicates that the switchers were on average younger than the stayers, but there were no significant

[3]The propensity score in our study is defined as the probability that each physician switches to the blended capitation model, given a set of covariates that include age, sex, location, the expected income gain from switching (as of 2006/2007), and the outcomes of interest (services and visits, as of 2006/2007).
[4]For the bandwidth selection, we used Silverman's (1986) optimal plug-in selector.
[5]The optimal number of repetitions was selected with the three-step methodology developed by Andrews and Buchinsky (2000, 2001).
[6]The expected income gain is calculated by using the current profile of services and patients in the Family Health Groups and applying the payment rules from the Family Health Organization model.

DOI: 10.1002/hec

Table IV. Change in outcomes, 2006/2007 versus 2013/2014

Fiscal year	Services per day		Visits per day	
	Switchers[a]	Stayers[b]	Switchers[a]	Stayers[b]
2006/2007	39.8	47.6	27.9	31.9
2013/2014	29.7	43.1	20.2	27.8

[a]Includes physicians who were in the Family Health Group model as of the fiscal year 2006/2007 but switched to the Family Health Organization model by the fiscal year 2013/2014. $N = 2078$ physicians.
[b]Includes physicians who were in the Family Health Group model in the fiscal years 2006/2007 and 2013/2014. $N = 1563$ physicians.

differences between the two groups in terms of their gender composition or geographical distribution. This analysis generates the third empirical regularity regarding contracts: Physicians with lower volumes of services and visits tend to prefer the capitation models, whereas physicians with higher volumes tend to prefer the EFFS models. This result – that higher-productivity individuals sort into jobs in which the marginal return on their productivity is higher – is well documented in the labor economics literature (Lazear, 2000; Shearer, 2004).

Given this result, we provide a preliminary decomposition of the total effect of payment contracts into incentive and selection effects in Table IV.

The total difference in the number of services in 2013/2014 between the stayers and switchers was 13.4 services per day. Interpreting the same difference in 2006/2007 as the selection effect (because both groups of physicians practiced in the same EFFS model), we can decompose the total difference of 13.4 services per day into 7.8 services because of the selection effect and the remaining 5.6 services per day because of the incentive effect. A similar calculation for the number of visits suggests that the total difference of 7.6 visits per day can be decomposed into 4 visits per day because of the selection effect and 3.4 visits per day because of the incentive effect.

As discussed earlier, this comparison can be improved by matching the stayers and switchers based on their propensity to switch to the blended capitation model as of 2006/2007. The propensity score for each physician was calculated using the outcomes and covariates listed in Table III, and the resulting distribution is shown in Figure 1.

This figure shows that the empirical support of the two distributions is very similar, although, as expected, the switchers have a higher average probability of joining the capitation model than the stayers.[7]

Given the propensity scores, we then estimate a full difference-in-difference matching estimate of the incentive effect. The results are presented in the first row of Table V and indicate that physicians in the capitation model provide significantly fewer services and visits per day than they would if they practiced in the EFFS model.

These results, performed in Stata 12.0 with the psmatch2 command, were produced from an LL regression model with a bi-weight kernel, bandwidth of 0.1, and common support restriction. The remainder of rows in Table V shows that the results are quite robust to use of the alternative estimators (nearest neighbor and kernel), the alternative kernel functions (normal, uniform, Epanechnikov, and Tricube), alternative bandwidth values (0.05 and 0.20), and alternative trimming levels (0% and 10%). Lastly, the results using the nearest neighbor matching method developed by Abadie *et al*. (2004) that correct for bias in standard errors, performed in Stata 12.0 using the nnmatch command, are quite similar to those obtained with propensity score matching.[8]

[7]The propensity scores are estimated using the logit model. The model has a reasonably good fit. The likelihood ratio test clearly rejects the hypothesis that included variables that are jointly insignificant (the likelihood ratio chi-square statistic with 20 degrees of freedom is about 910, with the associated p-value < 0.000). In addition, McFadden's R^2 is about 0.18. Lastly, the model correctly predicts treatment for about 67% of sample physicians.
[8]Specifically, the estimated incentive effect for service per day is −5.87 (0.67) and −5.87 (0.44) when we use 1 and 10 neighbors, respectively, where the figures in parentheses represent heteroskedasticity-consistent standard errors. Similarly, the estimates are −4.44 (0.32) and −4.33(0.27), respectively, for the incentive effect for visits per day.

DOI: 10.1002/hec

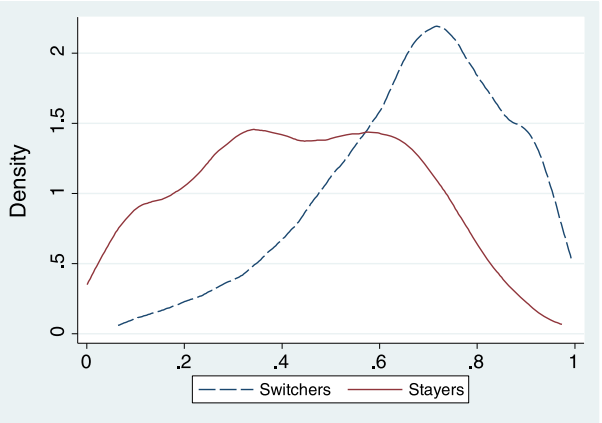

Figure 1. Distribution of estimated propensity scores

Therefore, this analysis generates the fourth empirical regularity regarding contracts: All else being equal, physicians provide more visits and services in the EFFS model than in the capitation model. Again, this finding has been well documented in the literature (Iversen and Lurås, 2012). Of particular interest here are the results for the UK fundholding scheme by Dusheiko *et al*. (2006), given its similarity to Ontario. Dusheiko *et al*. found empirical support for both the selection and the incentive effects of the fundholding contract. They assumed that physicians are heterogeneous with respect to a taste parameter and argued that this parameter determines both the admission rates for secondary care (because of the perceived gross benefit of admission) and the propensity to choose fund-holder status because of a fixed cost association with such status

Table V. Difference-in-difference matching estimates of incentive effects

	Services per day[b]	Visits per day[b]
Baseline model[a]	−5.67*** (0.45)	−4.10*** (0.29)
Alternative estimators		
Nearest neighbor (1 neighbor)	−5.06*** (0.58)	−3.87*** (0.36)
Nearest neighbor (10 neighbors)	−5.48*** (0.50)	−4.00*** (0.32)
Kernel	−5.62*** (0.42)	−4.08*** (0.28)
Alternative bandwidth values		
0.05	−5.49*** (0.47)	−4.02*** (0.30)
0.20	−5.77*** (0.45)	−4.14*** (0.29)
Alternative kernel functions		
Normal	−5.76*** (0.43)	−4.13*** (0.29)
Uniform	−5.77*** (0.46)	−4.15*** (0.30)
Epanechnikov	−5.72*** (0.45)	−4.12*** (0.29)
Tricube	−5.64*** (0.45)	−4.08*** (0.29)
Alternative trimming levels		
No trimming	−6.01*** (0.50)	−4.24*** (0.32)
10%	−5.64*** (0.43)	−4.06*** (0.28)

[a]The baseline model is the local linear regression model, with the bi-weight kernel and 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% of the treatment observations at which the propensity score density of the comparison observations is the lowest. The sample size for both dependent variables is 3428 physicians.
[b]Bootstrap standard errors in parentheses, using 200 bootstrap repetitions.
***Indicates significance at 1% level,
**significance at 5% level,
*significance at 10% level.

DOI: 10.1002/hec

(e.g., direct transaction cost). Empirically, they found that fundholding incentives reduced elective admission rates by 3% and accounted for 57% of the difference between fund-holder and non-fund-holder elective admissions, with 43% being a selection effect. This evidence is consistent with our study of Ontario in at least three ways. First, the authors convincingly demonstrate the importance of controlling for selection when estimating the behavioral consequences of payment contracts. Second, the results for the UK indicate that the volume of services (i.e., elective admission rates) is negatively related to the cost-reimbursement rate, which is consistent with our finding for Ontario. Lastly, the decomposition of incentive and selection effective is within the range of what we find for Ontario.

3. MODEL

3.1. Setup

To summarize, the empirical regularities that we wish to explain are as follows: (i) the existence of a menu of contracts; (ii) the lower capitation rate and the higher cost-reimbursement rate in the EFFS contract relative to the capitation contract; (iii) the capitation contract is more attractive to physicians with fewer services, and the EFFS is more attractive to physicians with a higher volume of services; and (iv) physicians in the capitation contract provide fewer services than physicians in the EFFS contract.

To explain these regularities, let us consider the classical problem of a payer (e.g., government) that wishes to design a payment contract for providers (e.g., physicians) to deliver healthcare services (e.g., patient visits). The number of patients treated by each provider is given exogenously and normalized to one. The number of services per patient, denoted by q, depends stochastically on the provider effort according to $q = e + \varepsilon$, where $e > 0$ denotes the provider effort (e.g., clinical hours) and ε is a mean-zero random variable (e.g., the stochastic component of patient demand for care).[9] All healthcare services are identical, and the price of each service (i.e., its value to the payer) is normalized to 1.

The payment contract w consists of two parts: a fixed payment a and a variable payment bq, with $0 \leq b \leq 1$. We refer to a as the capitation rate and to b as the cost-reimbursement rate. Although this contract is linear,[10] it is general enough to encompass the common types of payment contracts observed in practice, such as the pure FFS model ($a = 0$, $b = 1$), the pure capitation model ($a > 0$, $b = 0$), and the blended model ($a > 0$, $0 < b < 1$).

Provider utility is given by $U = w - c(e, \theta)$, where $\theta > 0$ is the provider type. For analytical convenience,[11] we assume that $c(e, \theta) = 0.5\theta e^2$. This cost function trivially satisfies the Spence–Mirrlees single-crossing property, because $c_{e\theta}(e, \theta) = e > 0$. Further, we assume that $\theta \in \{\theta_L, \theta_H\}$, with $\theta_L < \theta_H$. We refer to providers with $\theta = \theta_L$ as the low-cost type and to providers with $\theta = \theta_H$ as the high-cost type.[12] The proportion of the high-cost type is equal to α, and that of the low-cost type to $1 - \alpha$.

The payer's utility when contracting with a provider of type $i = \{L, H\}$ is equal to $V_i = q_i - w_i$.[13] We also assume that both the payer and providers are risk neutral, that all providers have an identical outside option equal to u,[14] and that the payer's outside option is 0.

[9]The stochastic component of services is introduced mainly to motivate the moral hazard problem; with deterministic services, the payer would be able to infer perfectly the provider effort.

[10]Our focus on the linear contracts is unlikely to be too restrictive, given that with more general nonlinear payment mechanisms, it is often the case that the optimal contract can be approximated by a set of linear contracts (e.g., Jack, 2005).

[11]The results generalize to the more general cost function $\theta c(e)$, where $c(.)$ is a strictly increasing and convex function. The results are available upon request.

[12]Because the Spence–Mirrlees condition holds, there is no loss of generality to consider a two-type model. A model with a continuum of types has the same qualitative implications. The results are available upon request.

[13]This specification of the social welfare function does not consider the shadow cost of public funding.

[14]We need not impose that u be positive, because it is conceivable that it can be negative because, for example, of student loans, cost of retraining, and mobility costs. The specific value of u will affect the value of capitation payments, which ensure provider participation, but not the cost-reimbursement rate.

DOI: 10.1002/hec

The timing of the contracting game is as follows. First, the nature determines the provider type $\theta \in \{\theta_L, \theta_H\}$, which is observed by the provider but not by the payer. Second, the payer offers a menu of two contracts (a_i, b_i) for $i = \{L, H\}$. Third, the provider either accepts one of the contracts or rejects both contracts. If the provider rejects both contracts, the game ends, and the provider receives its outside option u. If the provider accepts one of the contracts, it provides effort e that cannot be observed or verified by the payer.[15] Lastly, the nature determines ε, which then determines the number of services q_i and payoffs U_i and V_i.

The problem for the payer is to design a menu of contracts to maximize the benefit of the healthcare services provided to each patient, net of the payment to providers. Such a menu must satisfy three constraints: The contracts must be acceptable to providers; each provider must choose the contract that is designed for its type; and each contract must be compatible with the provider's optimal choice of effort. Specifically, the problem of designing an optimal payment contract can be stated as

$$Max\ E[V] = \alpha[e_H - a_H - b_H e_H] + (1-\alpha)[e_L - a_L - b_L e_L] \tag{1}$$

subject to

$$(PC_i) a_i + b_i e_i - 0.5\theta_i e_i^2 \geq u \tag{2}$$

$$(AS_i) a_i + b_i e_i - 0.5\theta_i e_i^2 \geq a_k + b_k e_k - 0.5\theta_k e_k^2 \tag{3}$$

$$(IC_i) e_i = argmax\ a_i + b_i \tilde{e}_i - 0.5\theta_i \tilde{e}_i^2 \tag{4}$$

for $i = \{L, H\}$ and $i \neq k$, where Equations (2)–(4) denote, respectively, the participation constraints, the adverse selection (or screening) constraints, and the incentive compatibility constraints.

Before analyzing the model, we wish to acknowledge some of the main features of healthcare markets not included in our stylized model. These include physician altruism, physician-induced demand, risk aversion, demand-side moral hazard, and risk selection, among others. All of these features are important for understanding healthcare markets, and we discuss their role at the end of this section.

3.2. First best

When the payer can observe and verify both the provider's effort and its type, the payment contracts must satisfy only the participation constraints. Furthermore, it is easy to verify that these constraints will bind at the optimum for each provider type. Using these constraints to substitute for the provider payment in the payer's expected utility yields

$$E[V] = \alpha\left[e_H - 0.5\theta_H e_H^2\right] + (1-\alpha)\left[e_L - 0.5\theta_L e_L^2\right] - u \tag{5}$$

Because this is a concave problem, the first-order condition for the effort level is both necessary and sufficient. Therefore, the first-best level of effort is given by

$$e_i^* = 1/\theta_i \tag{6}$$

for $i = \{L, H\}$. In this environment, it is not necessary to tie the provider's pay to the number of services because the provider's effort is verifiable and both parties are risk neutral. Therefore, substituting the first-best level in the participation constraint yields the optimal capitation rate

$$a_i^* = u + 1/2\theta_i \tag{7}$$

Therefore, in the full information environment, the low-cost type provides more effort and receives higher payment than the high-cost type ($e_L^* > e_H^*,\ a_L^* > a_H^*$).

[15] Alternatively, effort could be observed and verified but at a prohibitive cost. This is particularly the case if we think of effort as including not only the time component but also intensity and other characteristics.

DOI: 10.1002/hec

3.3. Moral hazard and adverse selection

When the payer cannot observe or verify either the provider's effort or its type, the contracts must satisfy all three types of constraints described in Equations (2)–(4). The provider's incentive compatibility constraint for each type i is given by the first-order condition from Equation (4), which can be written as

$$e_i = b_i/\theta_i \tag{8}$$

In the absence of adverse selection, the payer could induce the optimal level of effort from each provider by setting b_i equal to 1; from the participation constraint, it then follows that $a_i = u - 1/2\theta_i$. In this high-powered contract, the payer effectively 'sells the job' to the provider in exchange for a type-specific fee a_i, and the provider then fully internalizes the benefit of its effort. This efficiency result is not surprising given that the providers are risk neutral and there are no limited liability constraints. Further, the optimality of this FFS contract also follows because we abstract from physician agency, physician-induced demand, and demand-side moral hazard, all of which could play an important role in designing the optimal payment contract.

If, in addition to moral hazard, the payer cannot observe or verify the provider's type, the contract for each provider type must be such that each type chooses the contract designed for its type. This qualification is important because the optimal contract ($b_i^* = 1, a_i^* = u - 1/2\theta_i$) will, in general, fail to induce the appropriate sorting. Specifically, when offered a menu of contracts (b_i^*, a_i^*), both provider types will choose the contract designed for the high-cost type (b_H^*, a_H^*). To see this, note that for both provider types, choosing the contract designed for their type yields u. On the other hand, the high-cost type gains $u - 0.5\Delta\theta/\theta_H\theta_L < u$ if it chooses the contract designed for the low-cost type, and the low-cost type gains $u + 0.5\Delta\theta/\theta_H\theta_L > u$ if it chooses the contract designed for the high-cost type, where $\Delta\theta = \theta_H - \theta_L > 0$. Therefore, the low-cost type has an incentive to mimic the high-cost type, and the payer must design a menu of contracts different from (b_i^*, a_i^*).

By following a standard approach for solving adverse selection models (Bolton and Dewatripont, 2004; Laffont and Martimort, 2002), we assume that only the participation constraint for the high-cost type and the screening constraint for the low-cost type are binding and then verify ex post that the other two constraints are not binding at the optimum. Therefore, the relevant constraints are

$$a_H + b_H e_H - 0.5\theta_H e_H^2 \geq u \tag{9}$$

$$a_L + b_L e_L - 0.5\theta_L e_L^2 \geq a_H + b_H e_L(b_H) - 0.5\theta_L e_L^2(b_H) \tag{10}$$

At the optimum, both these constraints will be binding. In addition, by using the incentive compatibility constraints ($e_i = b_i/\theta_i$), we can express the payer's expected utility as

$$\begin{aligned} E[V] = \alpha\left[\frac{b_H}{\theta_H} - \frac{0.5b_H^2}{\theta_H}\right] + (1-\alpha)\left[\frac{b_L}{\theta_L} - \frac{0.5b_L^2}{\theta_L}\right] - u \\ - (1-\alpha)\left[\frac{0.5b_H^2\Delta\theta}{\theta_H\theta_L}\right] \end{aligned} \tag{11}$$

The first three terms represent the payer's expected utility in the full information environment, and the last term represents the expected information rent for the low-cost type. Therefore, the payer's problem entails a trade-off between productive efficiency and rent extraction.

Solving the first-order necessary and sufficient conditions for b_H and b_L and simplifying yields

$$b_L = 1 \tag{12}$$

$$b_H = \frac{\alpha\theta_L}{\alpha\theta_L + \Delta\theta(1-\alpha)} \in (0,1) \tag{13}$$

DOI: 10.1002/hec

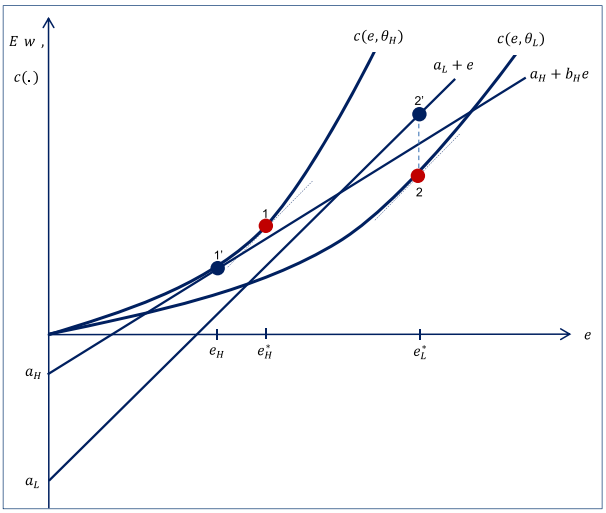

Figure 2. Optimal compensation contracts with moral hazard and adverse selection. Points 1 and 2 represent the efficient effort allocation in the full information environment, while 1′ and 2′ represent the optimal compensation contracts in the presence of both moral hazard and adverse selection

Therefore, the payer is able to induce the efficient level of effort from the low-cost type, and the cost-reimbursement rate for the high-cost type optimally trades off efficiency and rent extraction.[16] This result, known as the 'efficiency at the top,' is a standard result in contract theory (Bolton and Dewatripont, 2004; Laffont and Martimort, 2002).

The capitation payments can be derived from Equations (9), (10), (12), and (13) to obtain[17]

$$a_H = u - 0.5b_H^2/\theta_H \tag{14}$$

$$a_L = a_H - \frac{0.5\left(1 - b_H^2\right)}{\theta_L} < a_H \tag{15}$$

Lastly, it is straightforward to verify that the other two constraints are not binding.[18]

The main results of this section are illustrated in Figure 2. The equilibrium points 1 and 2 denote the full information environment, in which the marginal cost of effort for each provider type is equal to the (social) marginal benefit, which occurs at the point where the cost curves have a slope of 1. The equilibrium points 1′ and 2′ denote the environment with both moral hazard and adverse selection. In such an environment, the high-cost type exerts the inefficiently low level of effort, because $b_H < 1$ from Equation (13), whereas the low-cost type exerts the same efficient level of effort as in the full information environment, because $b_L = 1$ from Equation (12). At this equilibrium, the high-cost type receives its outside option u (normalized to 0 in the figure), whereas the low-cost type receives a rent equal to the vertical distance between its expected

[16]In the literature, it is usual to describe an FFS model as one in which the private marginal revenue exceeds the social marginal cost. The result that we obtain in this paper is directly because we extrapolate from other common assumptions about healthcare markets, such as physician altruism and physician-induced demand.

[17]The value of capitation payments in this model depends in a complex way on the providers' outside options and the distribution and variation of provider types, as shown in Equations (14) and (15). Therefore, the model does not imply that these capitation payments must necessarily be non-negative; further assumptions are needed to ensure this result, such as limited liability constraints.

[18]The participation constraint for the low-cost type is non-binding because, from Equations (10) and (14), we have that $U_L = u + 0.5\Delta\theta b_H^2/\theta_H\theta_L > u$. Further, the high-cost type earns u if it chooses the contract design for it, but it earns only $u - 0.5\Delta\theta\left(1 - b_H^2\right)/\theta_H\theta_L < u$ if it chooses the contract designed for the low-cost type.

DOI: 10.1002/hec

payment line and the cost curve at the efficient level of effort. Lastly, the capitation payment is higher for the high-cost type than for the low-cost type to ensure participation of both provider types.[19]

3.4. Discussion

The stylized model developed in Section 3.3 can account for all regularities related to the physician payment contracts in Ontario that we documented in Section 2. Specifically, the optimal contract design involves a menu of two contracts, where the prospective (a_i) and retrospective (b_i) elements are negatively related across the contract, the providers sort into the contract designed for their type (θ_i), and the provision of services varies between the contracts because of the differential cost-reimbursement rates (b_i). The link between the model implications and the regularities can be further clarified if we think of the high-cost physicians (θ_H) as those who switched to the blended capitation model and of the low-cost physicians (θ_L) as those who stayed in the EFFS model. All of the empirical regularities we documented could be understood in this unified framework as a solution to the problem of designing optimal payment contracts when neither physician type nor physician action is contractible.

Perhaps, the most important policy implication of this analysis is the recognition that under certain circumstances, it is optimal to offer healthcare providers a menu of contracts rather than a single 'best' contract. This approach, offering a menu of contracts, is analogous to the earlier literature in health economics that blended elements of FFS and capitation to overcome the unattractive features of each contract alone (Ellis and McGuire, 1986, 1990). The menu of contract similarly attempts to overcome the unattractive features of offering a single contract when providers are heterogeneous by tailoring each contract to the provider type. In designing such a contract, the policy makers should be aware of the inherent trade-off between efficiency and rent extraction, which must be resolved by designing contracts that are sufficiently different to induce each provider type to choose the contract designed for its type. This insight is relevant to policy makers concerned with either cost containment or quality improvement.

Although it is significant that this simple model can unify some observed regularities in healthcare markets, it must be recognized that these results are not entirely new, especially in the principal–agent literature. However, and to our best knowledge, the only other paper that employs this framework to explain the structure of physician payment contracts is Jack's (2005). Yet the results from Jack's model cannot be directly used to interpret the physician payment reform in Ontario. Specifically, Jack focuses on the quality and cost of health care as primary healthcare goals (rather than access), and the providers in his model differ in altruism (rather than the disutility of effort). Jack shows that the optimal contract can be approximated by a menu of the linear contracts of the form $\alpha(\theta) - \tau(\theta)c$, where $\alpha(.)$ is a fixed salary component, $1 - \tau(.)$ is a cost-reimbursement rate, θ is the degree of provider altruism, and c is the financial cost of providing treatment (e.g., the cost of labor services of other staff). In this model, $\alpha(\theta)$ and $\tau(\theta)$ are both increasing in θ, so that more altruistic physicians choose contracts with higher fixed payment and lower cost-reimbursement rate. Our model is similar to Jack's in two important ways. First, the optimal contract is a menu of contracts rather than a single contract. Second, the relationship between the fixed payment and the cost-reimbursement rate is negative across contracts. However, because Jack focuses on the quality and cost of health care and we focus on the access to health care, our contracts are linear in terms of number of services, whereas Jack's contracts are (at least approximately) linear in terms of the financial cost of services.

It is also important to realize that the optimality of a menu of contracts in our model is driven by heterogeneity between providers and that the exact source of this heterogeneity is not important for this result. In our model, providers are different with respect to their disutility of effort, but any of the alternative sources, such as risk aversion, altruism, and ability to induce demand, can serve this role as well. From this perspective, our omission of some important features of healthcare markets, such as risk aversion, altruism, and physician-induced demand, is not

[19]The capitation payments a_L and a_H are both negative in Figure 2 because we normalized the outside options to 0.

DOI: 10.1002/hec

critical for obtaining the result that a menu of contract is optimal. However, the exact source of heterogeneity between providers may have different implications for the shape of optimal contracts, as well as provider's behavioral responses to different contracts, and therefore, exploring this issue in future research is important.

An alternative approach to studying the implications of provider heterogeneity on a single dimension is to examine two or more sources of heterogeneity. For example, an extended model could consider an environment where physicians are different with respect to their disutility of effort, risk aversion, altruism, and ability to induce demand. A paper along these lines, although dealing with health insurance markets, is that of Einav *et al*. (2013) who examine selection on moral hazard. In their paper, the authors allow the policy holders to be heterogeneous in multiple ways, which allows them to distinguish between a standard adverse selection case (based on expected health risk and risk aversion) and between selection based on moral hazard (the policy holders' responsiveness to the price of insurance). In our model, however, the heterogeneity between providers relates to the cost of effort function, and this heterogeneity also influences the extent of moral hazard. Therefore, with a single dimension of heterogeneity, it is not possible to differentiate between the case of pure selection and selection based on moral hazard, even though our providers are forward looking and respond to the anticipated responsiveness to payment contracts. Nevertheless, this seems a promising area for future research.

Lastly, the patients in our model are identical, fully insured, and play a passive role in determining the quantity of medical services. Although such an assumption may serve as a first approximation in the universal coverage system in Ontario, there are at least two promising directions for future research. The first direction is to incorporate the insurance problem in the model and study the interplay between the optimal health insurance and the optimal menu of contracts for the providers (Ma and McGuire, 1997; Bardey and Lesur, 2006). The second direction is to recognize the heterogeneity among both patients and providers and study the strategic interaction between health insurance plans in a two-sided market framework in which the plans compete to attract both providers and patients (Bardey and Rochet, 2010; Bardey *et al*., 2014).

4. CONCLUSION

In this study, we analyzed the design of payment contracts aiming to maximize patient access to physician services when the payer has limited information about physician actions and his or her type. In such an environment, the optimal contract is a menu of contracts that blends FFS and capitation payments to varying degrees. This analysis draws attention to the potential benefit of screening, whenever there is unobserved heterogeneity among providers, and therefore of offering a menu of payment contracts rather than mandating a single contract. This point is well understood in the principal–agent literature, but it has yet to receive wider recognition in health care.

Although it is significant that the relatively standard economic model developed in this study can provide a unified explanation for some of the most important features of the primary care reform in Ontario, the model nevertheless falls short in fully explaining all complexities of the reform. For example, other features of the reform such as the introduction and impact of pay for performance bonuses, interdisciplinary teams, and preventative care incentives are not examined. In addition, the model abstracts from other characteristic features of healthcare markets, such as physician altruism, risk aversion, physician-induced demand, and risk selection. Extending the model to incorporate these features would help to address a richer set of questions and explain other features of the reform in Ontario and other jurisdictions.

ACKNOWLEDGEMENTS

We thank the editor, two anonymous referees, and participants in the 2015 Canadian Health Economics Study's Group annual conference in Toronto and the Canadian Centre for Health Economics seminar series at the University of Toronto.

DOI: 10.1002/hec

CONFLICT OF INTEREST

The authors have no conflict of interest.

REFERENCES

Abadie A, Drukker D, Herr JL, Imbens G. 2004. Implementing matching estimators for average treatment effects in Stata. *The Stata Journal* **4**(3): 290–311.

Abadie A, Imbens G. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**(1): 235–267.

Abadie A, Imbens G. 2008. On the failure of the bootstrap for matching estimators. *Econometrica* **76**(6): 1537–1557.

Allard M, Jelovac I, Léger PT. 2011. Treatment and referral decisions under different physician payment mechanisms. *Journal of Health Economics* **30**(5): 880–893.

Allard M, Jelovac I, Léger PT. 2014. Payment mechanism and GP self-selection: capitation versus fee for service. *International Journal of Health Care Finance and Economics* **14**(2): 143–160.

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

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

Bardey D, Lesur R. 2006. Optimal regulation of health system with induced demand and *ex post* moral hazard. *Annales D'Economie et de statistique* **83/84**: 279–293.

Bardey D, Rochet J-C. 2010. Competition among health plans: a two-sided market approach. *Journal of Economics and Management Strategy* **19**(2): 435–451.

Bardey D, Cremer H, Lozachmeur J-M. 2014. Competition in two-sided markets with common network externalities. *Review of Industrial Organization* **44**: 327–345.

Blundell R, Dias MC, Meghir C, Van Reenen J. 2004. Evaluating the employment impact of a mandatory job search assistance program. *Journal of the European Economics Association* **2**(4): 596–606.

Bolton P, Dewatripont M. 2004. *Contract Theory*. MIT Press: Cambridge, Massachusetts.

Buckley G, Sweetman A. 2014. Ontario's experiment with primary care reform. *University of Calgary, School of Public Policy Research Papers* **7**(11): 1–35.

Chone P, Ma CA. 2007. Optimal health care contracts under physician agency. *Annals of Economics and Statistics* **101/2**: 229–256.

Dehejia R, 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, Jacobs R, Smith P. 2006. The effect of financial incentives on gatekeeping doctors: evidence from a natural experiment. *Journal of Health Economics* **25**: 449–478.

Einav L, Finkelstein A, Ryan SP, Schrimpf P, Cullen MR. 2013. Selection on moral hazard in health insurance. *American Economic Review* **103**(1): 178–219.

Ellis RP, McGuire TG. 1986. Provider behavior under prospective reimbursement. *Journal of Health Economics* **5**: 129–151.

Ellis RP, McGuire TG. 1990. Optimal payment systems for health services. *Journal of Health Economics* **9**: 375–396.

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.

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

Iversen T, Lurås H. 2012. Capitation and incentives in primary care. In *The Elgar Companion to Health Economics*, Jones AM (ed.), (Second edn). Edward Elgar: North Hampton.

Jack W. 2005. Purchasing health care services from providers with unknown altruism. *Journal of Health Economics* **24**: 73–93.

Kantarevic J, Kralj B, 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.

Kralj B, Kantarevic J. 2013. Quality and quantity in primary care mixed payment models: evidence from family health organizations in Ontario. *Canadian Journal of Economics* **46**(1): 208–238.

Laffont J, Martimort D. 2002. *The Theory of Incentives: The Principal–Agent Model*. Princeton University Press: Princeton, New Jersey.

Lazear E. 2000. Performance pay and productivity. *American Economic Review* **90**(5): 1346–1361.

DOI: 10.1002/hec

Léger PT. 2008. Physician payment mechanisms. In *Financing Health Care: New Ideas for a Changing Society*, Lu M, Jonsson E (eds.), Wiley, Weinham: Germany; 149–176.
Leuven E, 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.
Li J, Hurley J, DeCicca P, Buckley G. 2014. Physician response to pay-for-performance – evidence from a natural experiment. *Health Economics* **23**(8): 962–978.
Ma CA. 1994. Health care payment systems: cost and quality incentives. *Journal of Economics and Management Strategy* **3**(1): 93–112.
Ma CA, McGuire T. 1997. Optimal health insurance and provider payment. *American Economic Review* **87**: 685–704.
McGuire T. 2000. Physician agency. In *Handbook of Health Economics*, Culyer AJ, Newhouse JP (eds.), vol. **1A**. North-Holland: Amsterdam; 461–536.
Makris M, Siciliani L. 2013. Optimal incentive schemes for altruistic providers. *Journal of Public Economic Theory* **15**(5): 675–699.
Nichols A. 2007. Causal inference with observational data. *The Stata Journal* **7**(4): 507–541.
Rosenbaum P, Rubin D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1): 41–55.
Rosenbaum P, Rubin D. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**(1): 33–38.
Rudoler D, Deber R, Barnsley J, Glazier RH, Laporte A. 2014. Paying for primary care: the factors associated with physician self-selection into payment models. *Canadian Centre for Health Economics* Working Paper No: 2014-06.
Rudoler D, Laporte A, Barneley J, Glazier R. H, Deber RB. 2015. Paying for primary care: a cross-sectional analysis of cost and morbidity distributions across primary care payment models in Ontario Canada. *Social Science and Medicine* **124**: 18–28.
Scott A 2000. Economics of general practice. In *Handbook of Health Economics*, Culyer AJ, Newhouse JP (eds.), vol. **1A**. North-Holland: Amsterdam.
Shearer B. 2004. Piece rates, fixed wages and incentives: evidence from a field experiment. *Review of Economic Studies* **71**: 513–534.
Silverman B. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall: London.
Zweifel P, Breyer F, Kiffman M. 2009. *Health Economics*. Springer-Verlag: Berlin.

DOI: 10.1002/hec