Labour Supply, Service Intensity and Contract Choice: Theory and Evidence on Physicians

Bernard Fortin, Nicolas Jacquemet, Bruce Shearer

2019-05

Labour Supply, Service Intensity and Contract Choice: Theory and Evidence on Physicians

Bernard Fortin
Nicolas Jacquemet
Bruce Shearer

Juin / June 2019

**Centre de recherche sur les risques
les enjeux économiques et les politiques publiques**

www.crrep.ca

ABSTRACT

We develop and estimate a structural model that incorporates service intensity and endogenous contract choice into the standard labour supply framework. We apply our model to data collected on specialist physicians working in Quebec (Canada). These physicians are typically paid a fee-for-service (FFS) contract. Our panel data set covers a period of policy reform which allowed physicians either to remain on FFS or to adopt a mixed remuneration (MR) contract, under which they receive a *per diem* as well as a reduced FFS. We estimate the preference parameters of physicians governing the choice of contract and their hours worked and services provided. We use our estimates to simulate labour supply elasticities, to predict (*ex ante*) the effects of contracts on physician behaviour, and to evaluate selection effects. The supply of services is reduced under a MR contract, suggesting incentives matter. The hours spent seeing patients is less sensitive to incentives than the supply of services. Our results suggest that a reform forcing all physicians to adopt the MR system would have had substantially larger effects on physician behaviour than were measured under the observed reform.

JEL Classification: C25, J22, J33, I10, J44.
Keywords: Labour Supply, Services, Contract Choice, Practice Patterns of Physicians, Discrete Choice Econometric Models.

Bernard Fortin : Département d'économique, Université Laval and CRREP, CIRANO and IZA.
Nicolas Jacquemet : Paris School of Economics and University Paris I Panthéon-Sorbonne.
Bruce Shearer : Département d'économique, Université Laval and CRREP, CIRANO and IZA.

The authors thank the *Collège des médecins du Québec* for making its survey data available and the *Régie de l'assurance maladie du Québec* and Marc-André Fournier for the construction of the database. This article was partly written while Fortin and Shearer were visiting the University Paris 1 Panthéon-Sorbonne. We thank participants at the Maurice Marchand Meeting in Health Economics (Lyon), the ADRES workshop on the Econometric Evaluation of Public Policies (Paris), the Canadian Economics Association (Montréal), the European Workshop on Econometrics and Health Economics (Thessalonique), the European Economic Association (Vienna) and the Econometric Society Winter Meeting (Chicago). We also thank seminar participants at CREST, the Free University of Amsterdam and Paris-Dauphine University. We are grateful to Michel Truchon as well as Bruno Crépon, Arnaud Dellis, Brigitte Dormont, Pierre-Yves Geoffard, Guy Laroque, Pierre-Thomas Léger, Pierre-Carl Michaud, and Marie-Claire Villeval for useful discussions and comments. We acknowledge research support from the Canadian Institute of Health Research (CIHR), le Fonds de recherche du Québec en société et culture (FRQSC), and the Canada Research Chair in Social Policies and Human Resources at the Université Laval.

1 Introduction

Physicians can affect their output at work through two basic margins: their hours spent at work and their volume of services per hour (McGuire, 2000). Yet, empirical studies of physician labour supply typically concentrate on either hours of work (*e.g.*, Showalter and Thurston, 1997; Ferrall, Gregory, and Tholl, 1998; Baltagi, Bratberg, and Holmas, 2005; Andreassen, Di Tommaso, and Strom, 2013; Kalb, Kuehnle, Scott, Cheng, and Jeon, 2018) or health care provision (Feldstein, 1970; Devlin and Sarma, 2008; Clemens and Gottlieb, 2014). Generalized models, which simultaneously analyse decisions over hours and services per hour (or service intensity) permit a more complete portrait of physician behaviour and allow for a richer policy evaluation environment. For instance, suppose that the government changes physician contracts, reducing incentives to perform services. If services per hour are negatively affected by this change while hours are not, then more physicians must be trained and hired to keep the supply of services constant and meet demand. Models that concentrate uniquely on hours of work will miss such effects. This issue is particularly relevant in a context of aging population. To date, little attempt has been made to analyse service intensity and hours of work together in empirical work.

In this paper, we develop and estimate a generalized physician labour supply model that combines reactions on both hours worked and services per hour. We provide a unified framework in which we can determine which margin is more sensitive to changes in incentives. Also, our model takes into account the particular nature of the environment in which physicians are working by distinguishing between clinical and non-clinical services. The latter includes for instance teaching and administrative activities (see Dumont, Fortin, Jacquemet, and Shearer, 2008).

We specify utility as a function of consumption, hours of work and service intensity.[1] Contracts are composed of an hourly wage rate and a piece rate per unit of service provided. The marginal return on an hour of work is thus endogenous and depends on service intensity. Similarly, the marginal return on service intensity depends on hours of work. These nonlinear prices are analogue to those obtained in quantity/quality models (Becker and Lewis, 1973). Assuming a continuous labour supply approach, some comparative static results are derived. In particular, we show that the compensated (Hicksian) supply curves of hours and services are positively sloped in the wage rate and the piece rate, respectively. In a more realistic model, the physician has the choice between two contracts: a fee-for-service (FFS) contract composed uniquely of a piece rate and a mixed remuneration (MR) contract composed of a wage rate per hour worked and a reduced piece rate. We show that this environment gives rise to a non-convex budget set, from which we derive an efficient budget constraint (the upper envelope of the contract-specific budget constraints).

We apply our model to the practice behaviour of specialist physicians working in the Province of

[1] In our model, the quality of a service (*e.g.*, in terms of its impact on a patient's health) is assumed the same regardless of how long the service takes to provide. This assumption is necessary given that no data are available on service quality. This important point will be discussed in detail later in the paper.

Quebec (Canada) between the years 1996-2002. All these physicians work within the Quebec public Health-Care System. Therefore, it is reasonable to assume that the parameters of each contract are exogenous for a physician. Our data contain information on individual physician labour supply (weekly hours spent seeing patients, weekly hours spent performing administrative tasks or teaching, and weeks worked per year) as well as the number of services provided by each physician per year and the fees for service and the *per diem* for each contract. The observation period also spans an important reform in physician compensation which we exploit to identify our model. Prior to 1999, most specialist physicians in Quebec (92%) were paid FFS public contracts, receiving a fee for each service provided. In 1999, the government introduced a non-mandatory mixed remuneration (MR) scheme, under which physicians received a (half) *per diem*, paid for 3.5 hours worked, and a reduced fee-for-service.

To estimate the model, we assume that preferences are (directly) independent of the compensation system. This implies that rational, unconstrained physicians will locate on the efficient budget constraint—the budget constraint that maximizes a physician's income for each possible combination of practice variables in his choice set. We derive the efficient budget constraint from our knowledge of the physician's contracts. We pay careful attention to the complications created by the institutional constraints imposed on these contracts within the Quebec Health-Care System (*e.g.*, income ceilings, regionally differentiated remuneration, and constraints on the choice of the compensation system at the individual level).[2] The simultaneous modelling of the allocation of time, work intensity and institutional constraints introduces strong nonlinearities into the budget constraint. To account for these nonlinearities in estimation, we discretize the choice set available to physicians (Zabalza, Pissarides, and Barton, 1980). This methodology is relatively free of restrictions (MaCurdy, Green, and Paarsch, 1990), imposing only that the marginal utility of income is positive (van Soest, 1995).

We then solve for the utility function parameters that generate the observed practice patterns as optimal choices along the efficient budget constraint. To account for selection we allow for heterogeneity in preferences (both observable and unobservable), estimating a mixed-logit model (McFadden and Train, 2000).[3] To minimize the effects of functional forms on our results, we use a flexible (quadratic) utility function. In order to limit computational time in estimation and to reduce the problem of heterogeneity in the nature of services provided, we restricted our sample to one speciality—pediatrics. This specialty provides high variability in the participation in MR—44% of pediatricians opted for MR in the year 2000 as compared with 31% for all specialities. The voluntary nature of the reform further complicates estimation, for the following reason. The decision to adopt MR was not individual specific, but determined at the department level within hospitals.[4]

[2]For simplicity sake and given the high average physicians' gross income, we assume that their (federal + provincial) income marginal tax rate fall in the highest tax bracket.

[3]These authors have shown that, under mild regularity conditions, any discrete choice model derived from random utility maximization has choice probabilities that can be approximated as closely as one pleases by a mixed-logit model.

[4]Members of each department (groups of specialists working in the same field) would vote on the adoption of MR;

Consequently, individual physicians could be constrained in their choice of a compensation system. Accounting for constraints on choice leads to a mixture of likelihoods wherein the probability of being constrained is estimated along with the other parameters.

Our results suggest that weekly hours elasticities are quite small while the (compensated) elasticities of service intensity and services with respect to the fee per service are much stronger, being estimated at about 0.303 and 0.373, respectively. Our results also suggest that the changes in incentives brought about by the 1999 reform significantly affected physician behaviour. Services completed decreased by 5.32% and non-clinical hours increased by 6.52%. What is more, service intensity decreased by 4.78% (less services per clinical hour). A mandatory reform, forcing all physicians to work under MR, would have reduced services by 9.03% and increased non-clinical hours by 12.04%. However, these larger effects have little to do with unobserved heterogeneity and selection. More important are the constraints placed on individual choice in the observed reform.

Simulations using our mixed-logit model estimates also allow us to evaluate selection effects on unobservables for the subgroups of physicians who prefer MR and FFS, respectively. Our results indicate that MR physicians provide fewer services and spend more time per service, on average, than do FFS physicians. The differences are not substantial. This confirms that when exogenous observables (gender and age) are controlled for, the selection problem on unobservables does not seem to be important.

The reform was also costly, increasing payments to physicians by over 9.95% . This is due to the large *per diem* that physicians were paid for working under MR. We investigate the effects of a constant-cost reform, under voluntary participation in MR. Under such circumstances, services provided would decrease relative to the FFS contract by 4.50%, the clinical hours worked by only 0.75% and service intensity by 3.78%.

The rest of the paper is organized as follows. Section 2 develops the basic model that we will use in this paper. Section 3 describes the institutional details of the FFS and MR systems and derives the physician's budget constraint. Section 4 presents our data and summary statistics. Section 5 adapts the model of Section 2 to the institutional details of the Quebec reform and develops our econometric model. Section 6 describe our empirical results, the policy simulations, and an analysis of selection related to the decision of participating in MR. Section 7 presents our conclusions.

2 A generalized model of labour supply

For expositional purposes, we first present a static model of continuous choice labour supply under linear contracts. Our model allows for decisions over hours of work and service intensity. Our goal is to motivate our empirical analysis and our estimation strategy within a simplified setting. Later we will adapt the model to fit the specific institutional details of physician labour supply in Quebec.

adoption required unanimous approval.

Preferences are represented by a continuous and twice-differentiable utility function

$$U(X, h, s), \tag{1}$$

where X is consumption, h is hours of work, and s is service intensity, that is, the number of services, S, performed per hour of work. h and s correspond to the physician's practice variables.[5] We suppose

$$U_X > 0,\ U_h < 0,\ U_s < 0, \tag{2}$$

where the latter inequality is justified by assuming that an increase in service intensity reduces on-the-job leisure, which is assumed a desirable good. From the definition of service intensity one has[6]:

$$S = sh. \tag{3}$$

The budget constraint is given by

$$X = wh + pS + y, \tag{4}$$

where X is the numéraire, w is the wage rate, p is the fee per unit of service and y is non-labour income. The variables w and p are treated as exogenous. This is consistent with the public health-care system in Quebec where physicians' remuneration parameters are determined at the government level. Note that the budget constraint, given by eq. (4), is general enough to account for many contracts of interest: setting $w = 0$ and $p > 0$ gives the FFS contract, setting $w > 0$ and $p = 0$ gives a fixed-wage contract, while setting $w > 0$ and $p > 0$ gives a mixed contract.

Four important remarks are in order concerning our model. First, as mentioned in the introduction, our approach omits the quality of services in the utility function (or alternatively, we assume that quality is fixed in our model). We acknowledge that this is a strong assumption, but it is necessary in the absence of data on quality. In Fortin, Jacquemet, and Shearer (2008, p.301), we showed that our utility function can be derived from a more general model in which services per hour enter utility for two reasons: first because it is an input of the work practice of the physician, and second because of ethical concerns, it is derived from the effects of quality on patients' health. However we cannot identify the structural effect on health without data on health outcomes. We therefore omit quality from the utility function in our specification.

[5]Our simple model considers only one kind of service and one kind of work hours. The empirical model will generalize the approach to two different kinds of both services (billable/non-billable) and hours of work (clinical/non clinical) so as to take into account the particularities of the reform.

[6]One interpretation of (3) is a (Cobb-Douglas) production function. In a more general model, this function could be written as $S = S(e, h; \mathbf{z})$, with e denoting effort and $\mathbf{z}$ denoting an exogenous vector of inputs that affect the marginal productivity of effort and hours worked. Absent information on effort and on $\mathbf{z}$ in our data set, we approximate the production function by eq. (3).

Second we assume complete and symmetric information. Therefore we ignore agency problems and moral hazard. This is a direct consequence of the fact that the utility function is assumed to depend on hours of work and the service intensity, both being observable variables.

Third, we suppose that the worker has complete control over his practice variables—freely choosing both his hours of work and his clinical services. This rules out constraints to supply or any demand shocks that might affect a physician's practice, allowing us to concentrate on the supply side of the medical market which considerably simplifies the empirical analysis.[7]

Fourth, in its most general form, our model combines traditional labour supply analysis with a piece-rate model, giving rise to non-linear (and endogenous) prices in the budget constraint. This can be seen by substituting (3) into (4), adding and subtracting psh, and rearranging. This gives $X = (w + ps)h + (ph)s + y^v$, where $y^v = y - phs$ is the *virtual* non-labour income. It follows that the marginal return to an hour of work, $w + ps$, depends on the physician's choice of service intensity—the number of services that can be performed in that hour. Similarly, the marginal return to service intensity, ph, depends on the physician's hours of work. Since service intensity changes the number of services performed per hour, the return to service intensity depends on the number of hours worked. These nonlinear prices are similar to those obtained in quantity/quality models (Becker and Lewis, 1973).

The nonlinear prices give rise to a non-convex budget set (see Appendix A.2). The second-order conditions for an interior solution require that the curvature of indifference surfaces be more pronounced than the curvature of the budget set. We assume this to be the case and denote the optimal solution $(X^\star, h^\star, s^\star)$. In Appendix A.3 we show that $(X^\star, h^\star, s^\star)$ is equivalent to $(X', h', S'/h')$ which maximizes the transformed utility:

$$u(X, h, S) \tag{5}$$

subject to

$$X - pS - wh = y,$$

where (5) is obtained by substituting $s = S/h$ directly into the utility function (1): $u(X, h, S) = U(X, h, S/h)$. Hence we can identify the parameters determining optimal behaviour using either program: Max (1) s.t. (3) and (4), or Max (5) s.t. (4). In most of the following, we concentrate on the transformed program. One advantage is that all arguments of the transformed utility are well-defined over the whole choice set; service intensity is not defined in (1) when hours are set at zero.

The non-linearities in the budget constraint complicate the comparative statics of the model. For example, an increase in non-labour income, y, will affect the worker's choices of service intensity and hours of work through two channels: the first is the standard income effect, the second

[7]This assumption seems reasonable within the context of a public health-care system such as in Quebec where long waiting lists for physicians' services render the demand side of the market relatively passive. Excess demand also reduces any incentive of physicians for demand inducement, which we also ignore.

is through its impact on the endogenous marginal returns to service intensity and hours of work. Some results are possible however. The fact that the budget constraint is linear in S and h implies that the expenditure function is concave in w and p. Hence, under the assumption that the transformed utility function is quasi-concave within the relevant region of analysis, we have

$$\frac{\partial \tilde{h}}{\partial w} \geq 0; \frac{\partial \tilde{S}}{\partial p} \geq 0, \tag{6}$$

where ˜ indicates that the partial is compensated. Notice however that one cannot sign cross-partial derivatives (or elasticities). Therefore,

$$\frac{\partial \tilde{h}}{\partial p} \underset{<}{\geq} 0. \tag{7}$$

Similarly, since service intensity is the ratio of services to hours worked, its compensated elasticity with respect to p is given by:

$$\tilde{\eta}_{s,p} = \tilde{\eta}_{S,p} - \tilde{\eta}_{h,p}, \tag{8}$$

which is unsigned. Indeed, while $\tilde{\eta}_{S,p} \geq 0$, the sign of $\tilde{\eta}_{h,p}$ is indeterminate. These results follow from a straightforward application of duality theory to the problem of maximizing (5) s.t. (4); we include a derivation in Appendix A.4 for completeness.

2.1 Endogenous Compensation Choice

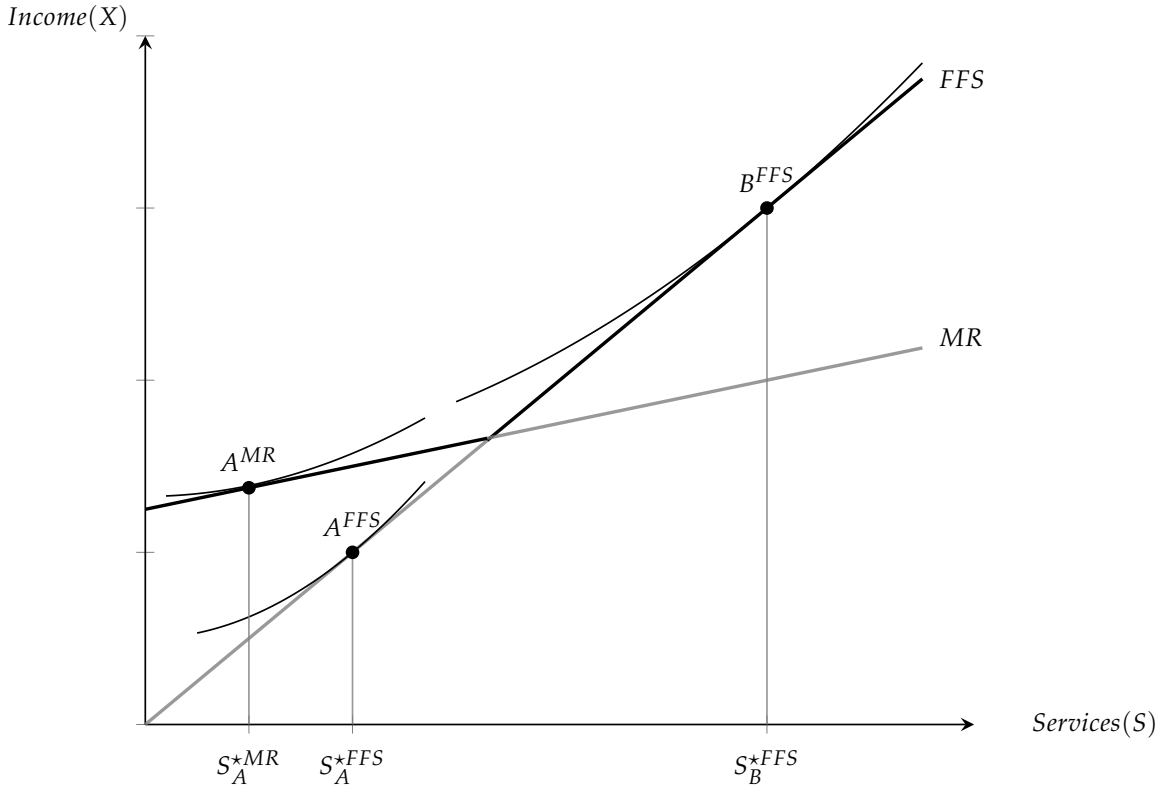
Introducing the choice of a compensation system complicates the analysis somewhat. We consider two cases: a fee-for-service (FFS), or piece rate, system ($X = pS$) and a mixed remuneration (MR) system ($X = wh + \alpha pS$), where $0 \leq \alpha < 1$ denotes the discount rate on the fee-for-service payment (setting $\alpha = 0$ gives a fixed-wage compensation system). To proceed we note that $U_X > 0$. Moreover, we assume that preferences are (directly) independent of the compensation system. This implies that rational workers will always select the compensation system that maximizes income for a given (h, S) combination. We therefore proceed in two steps: first we determine the *efficient budget constraint*, the upper envelope of X attainable from each value of (h, S). Assuming for simplicity zero non-labour income, we have

$$X(h, S; w, p, \alpha) = \max_{D \in \{0,1\}} [(1 - D)pS + D(wh + \alpha pS)], \tag{9}$$

where D is a dummy variable is equal to one when the worker participates in the MR system. Second, the worker solves his (transformed) program by choosing the $(X^\star, h^\star, S^\star)$ combination that maximizes his utility along the efficient budget constraint (9). The choice of a compensation system is then given by

$$D(h^\star, S^\star; w, p, \alpha) = \arg\max_{D \in \{0,1\}} [(1 - D)pS^\star + D(wh^\star + \alpha pS^\star)]. \tag{10}$$

Figure 1: Optimal Choices Along the Efficient Budget Constraint



Note. The Figure illustrates the endogenous selection into compensation schemes based on physicians preferences. Clinical hours are held constant.

This is illustrated in Figure 1 which considers the tradeoff between services and consumption (income), conditional on $h^{\star\ FFS}$, the optimal hours under the fee-for-service system.[8] The budget line FFS has slope p, the marginal monetary return to completing services under FFS; it passes through the origin because hours are not remunerated under FFS. The values of S, X chosen under FFS correspond to the optimal values $S^{\star\ FFS}$ and $X^{\star\ FFS}$. The line MR illustrates the tradeoff between services and income under MR, holding hours fixed at $h^{\star\ FFS}$. It cuts the y-axis at $wh^{\star\ FFS}$ and has slope equal to αp, reflecting the reduced fee-for-service payments received under MR.[9]

[8]Notice that optimal hours under FFS are not in general equal to zero, even though $w = 0$. This is due to the fact that the marginal return to hours includes both wage and the marginal effect of hours on FFS income ($= w + ps$). From the first-order conditions, we have $\frac{-U_h}{U_X} = w + ps$.

[9]Under a fixed-wage system ($\alpha = 0$) the monetary return to services provided is zero. A strict interpretation of the model under these circumstances would imply zero services. However, relatively straightforward extensions to the model would allow for positive services being allowed at the optimum. One possibility is to assume that $U_s > 0$ for $s < \bar{s}$, due for instance to concern of the physician for his patients. Another is to assume that monitoring allows for a minimum level of service intensity to be enforced.

The efficient budget constraint associated with the transformed program is given by the bold line.[10] It is piece-wise linear and non-convex; this raises well-known problems for optimization and labour supply estimation (Hausman, 1985) the choice set of workers (physicians), considering only a finite set of values for h, S (van Soest, 1995).

Figure 1 also illustrates potential problems of self-selection. Workers who have a preference for low service intensity levels (such as worker A, who chooses $S_A^{\star FFS}$ under FFS) will tend to choose MR, while those who have a preference for high service intensity levels (such as worker B, who chooses $S_B^{\star FFS}$ under FFS) will tend to choose FFS. A comparison of behaviour across compensation systems will potentially confound the effects of the compensation system with the differences in preferences. This is the case since the marginal return of services and the (virtual) non-labour income are endogenous, as they may vary at the optimum depending on the form of the indifference curves (compare A_1 and B_0). The econometric model must therefore allow for both observable and unobservable heterogeneity to take into account of possible selection bias.

3 Institutions: Physician Remuneration in Quebec

Health care is under provincial jurisdiction in Canada. Each province determines physician compensation systems and their level of pay. Within the Province of Quebec, physicians have traditionally been paid according to a fee-for-service compensation system.[11] Under this system, physicians receive a fee for each service provided. The fees paid are service specific, accounting for the difficulty and time intensiveness of the service provided. Our empirical work will account for these differences by constructing index numbers of services and prices.[12] In the present section, for expositional purposes, we take as given that there is one type of service, denoted S, and one fee, denoted p. The physician's budget constraint is then given by

$$\widehat{X}^{FFS} = pS. \tag{11}$$

In 1999, the government introduced a Mixed Remuneration scheme. Under this system, specialist physicians receive a wage (or *per diem*) for time spent at work in hospital (or other recognized health-care establishments). (Half) *per diem*s are paid for periods of 3.5 hours of work. To receive the *per diem*, a physician must explicitly declare the time period under which he is working under MR ("on the *per diem*"). During this period the physician is allowed to perform certain activities within a hospital. These activities include seeing patients, administrative services, and teaching; research activities are not covered. In practice, *per diems* of $\mathcal{D} = \$300$ are claimed in $\overline{d} = 3.5$ hour blocks, up

[10] To be exact, this is the efficient budget constraint conditional on $h^{\star,FFS}$, the unconditional efficient budget constraint varies with h as well as S.

[11] We ignore the cases of salaried physicians and physicians paid by vacation, which represents a small part (about 8%) of physicians before 1999 and a still smaller part afterwards.

[12] The construction of the index numbers for all prices and services is outlined in the next section and described in detail in the Appendix, Section A.1.

to 28 over a two-week period. The *per diem* was increased to $335 in 2003. Services provided during this period fall into two categories: billable services, denoted S^{B}, are remunerated at a reduced fee-for-service, αp, with $0 < \alpha \leq 1$,[13] and non-billable services, denoted S^{NB}, are not remunerated: $\alpha = 0$.[14]

Billable services must be further differentiated by whether or not they were performed while the physician was on the *per diem*. This is due to the fact that physicians working under MR do not necessarily spend all of their time under the *per diem*. Clinical services provided outside a *per diem* period are remunerated according to the same rate as for FFS physicians, p. We denote billable services that were performed outside of the *per diem* by S^{B}_{FFS}. Those performed under the *per diem* are denoted S^{B}_{MR}. Non-billable services, performed under *per diem* periods are not paid and hence not recorded, which raises econometric issues discussed later.

To calculate annual income under MR, let $\mathcal{N}$ denote the average number of *per diems* claimed per week throughout the year, and $\mathcal{W}$, the number of weeks worked during the year. Gross income under the MR system is then given by

$$\widehat{X}^{MR} = \mathcal{W}\mathcal{N}\mathcal{D} + \alpha p S^{\text{B}}_{MR} + p(S^{\text{B}}_{FFS} + S^{\text{NB}}_{FFS}). \tag{12}$$

The earnings derived from practice is thus given by[15]

$$\widehat{X} = D\widehat{X}^{MR} + (1 - D)\widehat{X}^{FFS}. \tag{13}$$

Table 1 provides a summary description of the compensation system that applies to specialist physicians in Quebec.

4 Data and Summary Statistics

Our data contains information on the labour supply behaviour and individual characteristics of physicians practicing in Quebec between 1996 and 2002. These data come from two sources. The first source of data is the time-survey conducted annually by the College of Physicians of Quebec. This survey provides information on the average number of hours per week spent seeing patients as

[13]The value of α is speciality and service specific. Its average over all billable services in our sample is approximately 0.3.

[14]The MR system is not applicable to work performed in private clinics; services provided in such clinics are generally billed on a FFS basis.

[15]Disposable income of physicians differs from (13) due to two factors. First, the government imposes income ceilings on physicians, beyond which the price paid for each service is reduced by 75%. Second, there is a regionally differentiated remuneration rate, designed to induce physicians to practice in remote areas of the province. These two points are taken into account in our econometric analysis.

Table 1: Remuneration of Quebec Physicians included in the sample

FFS	MR	
No fixed remuneration		- Earned for each 3.5 hours of work in hospital
Administrative/teaching activities	Half *per diem*:	- All kinds of practice eligible
uncompensated		- Limited to 28 every two weeks of work
	Billable	- Compensated at price αp during *per diem* hours
Clinical Services	Services :	- Compensated at price p outside *per diem* hours
compensated		
at price p	Non-billable	- Uncompensated during *per diem* hours
	Services :	- Compensated at price p outside *per diem* hours
Differentiated remuneration based upon individual characteristics		
Ceiling[a]		

[a]Except for emergency activities until 2001, and the whole hospital activities since 2001.
Note. The first two rows describe the way hours of work (*first row*) and services (*second row*) are remunerated under Fee-for-Service (*left-hand side*) and Mixed Remuneration (*right-hand side*). The last two rows describe some income policies that equally apply to both compensation schemes.

well as hours spent performing teaching and administrative duties. Since the MR reform occurred in the last quarter of 1999, we eliminated 1999 and 2000 from our empirical analysis, as these years correspond to a period of transition to the reform. Incorporating these data would introduce more dynamic complications to the model without any clear cut benefits to the estimation of the response of physicians to incentives. Also, we assume that annual weeks worked are exogenous and set at its average over the period (= 45) for each individual. Including the choice of the number of weeks worked in our model would have introduced an unnecessary complication. Preliminary results on this variable indicate that there is almost no variation in the data, no effect on the estimates, and the elasticities are tiny (Fortin, Jacquemet, and Shearer, 2010). Moreover, this allows us to include 2001 in our sample, a year for which the time-survey does not provide information on weeks worked. The survey also includes information on the personal characteristics of each physician, including age, gender, and specialization.

The second source of data is the Health Insurance Organization of Quebec (the RAMQ). This is a public-sector organization, responsible for paying physicians in the province. It therefore has administrative records containing information on the income and billing practices of each physician working in the province. Data on income and the number of services provided are available on a quarterly basis for every physician working in the province. Data from these two sources have been anonymously matched on the basis of physician billing numbers.

Typically, each physician provides a variety of different services, each remunerated at different rates. These rates reflect differing input requirements in terms of the physician's time and service intensity. To keep our estimation problem tractable, we aggregated these different services to form a quantity index of services provided, distinguishing only between billable and non-billable services.

We weighted the different types of services by the fee received for that service. This provides a control for the difficulty in providing the service.

Price variation is excluded from the index by holding price weights constant at the base year levels.[16] These weights are the base-year prices paid to FFS physicians; they are the same for both billable and non-billable services.[17] The price data for different services was also aggregated into indexes for billable and non-billable services, under FFS and MR. The price index for services provided under FFS, denoted p, was calculated as a Laspeyres price index. The average number of each type of service provided in the base year served as the weight for the price of that service. The index for services provided under MR, denoted αp, was similarly calculated by aggregating the fees paid for individual services under MR. Here we also used the average quantities of each service provided among FFS in the base year as weights. In this way, the MR price index excludes quantity variations due to MR switching. The precise calculations underlying all indexes are given in the Appendix A.1.

The empirical model that we estimate is numerically intensive, involving multidimensional integrals. In order to limit computational time we restricted the sample to one specialty: pediatrics. This specialty provides high variability in the participation in MR (58% of pediatricians opted for MR in the year 2001) and in the marginal incentives to perform services.[18] Focusing on one specialty also reduces the problem of heterogeneity in the nature of services provided. Summary statistics for the sample period are provided in Table 2. We divide the sample into Before MR Reform (1996 to 1998) and After MR Reform (2001-2002) and on the basis of physicians who remain under FFS or switch to MR (the panel is unbalanced due to flows in and out of the population of physicians over the period). A physician is considered to have switched to MR if he is paid (at least in part) under the MR system during the sample period. Note that patients are not informed about the compensation scheme that applies to the physician they see. Moreover, waiting lists were very long in Quebec during the sample period. As a result, it is unlikely that the changes in practice patterns that are observed upon the adoption of MR are due to changes in the patient mix that physicians face.

The top part of the table provides information on the professional practice behaviour of the physicians in our sample, disaggregated into the four categories considered. We focus on weekly hours of work (defined as the self-declared time spent in hospital), both in clinical medicine (providing services to patients, defined as *clinical hours*, h^c) and other activities (defined as the time spent working in hospital without seeing patients, like doing administration and teaching, and defined as non-clinical hours, h^o),[19] clinical services provided (both billable and non-billable), measured in

[16] To account for new services and services that become obsolete, we used two base years, producing a Linked Laspeyres index.

[17] This ensures that the difficulty weight applied to each service is independent of the manner in which the physician is paid.

[18] Dumont, Fortin, Jacquemet, and Shearer (2008) provides an extensive summary of MR across specialties.

[19] Note that administration covers the time physicians need to spend on following-up on patients treatment (updating

Table 2: Summary statistics on sampled physicians

	FFS physicians		MR physicians	
	Before	After	Before	After
	Observed practice			
Weekly Total Hours	43.09	41.92	48.64	46.73
	[13.01]	[12.83]	[12.67]	[10.62]
____ clinical (h^c)	38.69	38.85	41.38	39.02
	[12.79]	[11.62]	[13.73]	[12.62]
____ non clinical (h^o)	4.40	3.07	7.26	7.71
	[8.36]	[8.20]	[9.62]	[10.33]
Total Services[a]	167.00	167.94	141.81	122.19
	[66.83]	[72.88]	[56.16]	[72.24]
____ Non-billable[b] (S^{NB})	71.85	73.22	60.94	55.19
	[47.02]	[57.50]	[36.20]	[46.62]
____ Billable (S^B)	95.15	96.73	80.88	67.00
	[55.47]	[57.44]	[49.21]	[46.07]
Service intensity $\left(=\frac{S^{NB}+S^B}{h^c * W}\right)$	106.68	99.21	82.79	69.77
	[112.99]	[42.24]	[38.79]	[37.97]
Annual income[a] (X)	167.84	191.71	146.41	191.66
	[67.35]	[79.64]	[56.86]	[61.85]
	Sample characteristics			
Number of physicians	139	123	111	99
Number of observations	355	206	267	175
Gender (Male = 1]	0.66	0.65	0.52	0.55
	[0.47]	[0.48]	[0.50]	[0.50]
Age	49.89	52.70	43.07	47.26
	[11.17]	[11.04]	[10.04]	[10.03]

[a] In thousands of (1996) Can. Dollars.
[b] Lower bound for MR physicians after the reform. See below, Section 5.4.
Note. The upper part provides the average practice behavior of Quebec pediatricians included in our sample, split according to their choice of compensation scheme—FFS physicians are those who never adopt MR during the observation period, MR physicians are those who switch to MR—and the time period—before (1996-1998) and after (2001-2002) the reform. The bottom part of the Table summarizes individual characteristics. Standard errors appear in brackets.

thousands of (1996) Can. Dollars, service intensity (total clinical services per clinical worked hour), and annual income. We present the average and standard deviation of each variable. The bottom part of the table presents summary statistics on the demographic characteristics of physicians in each of the different categories.

These statistics on professional practice suggest both changes in behaviour subsequent to the introduction of MR (incentive effects), and the selection of physicians into contracts. Changes in

the patient record, heading a department, etc.). This does not include hospital administration like actually billing the patients, which falls under the responsibility of the hospital staff.

behaviour are evident from the Before and After columns among the MR physicians.[20] The average volume of services supplied by MR physicians decreased after switching to MR from 141.81 to 122.19. This change, in the order of 13.8%, is suggestive of a substantial reaction to incentives among those treated by the reform. Notice that this is composed of changes in both billable and non-billable services. The table shows that the supply of both types of services decreased after the introduction of the reform, although one must bear in mind that non-billable services are only partially observed after the reform.[21] There is a 17% decrease in billable services, from 80.88 to 67. This is a similar order of magnitude to the treatment effect (on the treated) among pediatricians calculated by Dumont, Fortin, Jacquemet, and Shearer (2008) among pediatricians (12.81%) using difference-in-difference techniques.

Other behavioural changes are also suggested by the table. MR physicians sharply decreased their service intensity from 82.79 to 69.77, a decrease of 15.7%. This decrease in services performed per hour is largely due to the change in services provided—weekly clinical hours worked changes relatively little with the reform (from 41.38 to 39.02, a decrease in the order of 5%). The increase in income among the FFS physicians that is observed after the introduction of MR is due to the government increasing the fees paid per service.[22] MR physicians' earnings increased much more (in percentage) than those of FSS physicians (31% *vs* 14%). This suggests that the introduction of the *per diem* offset any loss of earnings due to a reduction of services provided and hours worked.

Table 2 also points to potentially important selection effects on observables and/or inobservables in the data. There are notable differences in terms of clinical and non clinical weekly hours of work. Before the reform, MR physicians provided 7% more clinical hours and 65% more non clinical hours of work than FFS physicians. This latter result may be related to the endogenous decision to switch to MR, the non clinical hours being compensated under MR (by the *per diem*) but not under FFS. There is also evidence that physicians who eventually switched to MR were, on average, low "service intensity" physicians. MR physicians provided 15% fewer total services before the reform than FFS physicians. The difference in services leads to a substantial difference in annual income, pre-reform; MR physicians earned approximately 13% less income.

One important part of the explanation for these results is likely to be selection on observables (in particular, gender and age). Table 2 shows that before reform, 66% of FFS physicians were male, while only 52% of MR physicians were male. This indicates that the proportion of females who switched to MR (59%) is larger than that of males (38.6%). This is perhaps unsurprising since the female physicians work fewer hours and provide fewer services than do the male physicians in our sample. Thus female physicians had more incentive to adhere to the MR system. Also, before

[20] As the FFS physicians did not change compensation systems, it is to be expected that there is little variation in their behaviour after the reform.

[21] Recall, non-billable services are only paid under RM if they are performed outside of hospital. Therefore, part of the observed reduction is mechanical. The fact that the decrease in volume is relatively small, in spite of this, probably reflects the small payments that physicians receive for these, relatively minor, services.

[22] The reaction of FFS physicians to this increase is studied extensively in Some (2016).

reform, MR physicians are younger (43 years on average) than physicians who remained under FFS (50 years on average). This may partly be explained by the presence of preference habits that are likely to be stronger for older physicians.

While Table 2 provides a number of interesting statistics, more sophisticated econometric approaches are needed to address the selection issue. In particular, introducing unobserved heterogeneity in the model allows us to account for the selection on unobservables as well as to control for exogenous observable variables. Also, structural approaches are required to perform counterfactual reforms and ex-ante prediction. Dumont, Fortin, Jacquemet, and Shearer (2008) provide an empirical analysis of this reform via a difference-in-difference approach (see their Table 6). Their results are quite close to those in Table 2 and correspond to Average Treatment Effects on the Treated (ATT). Such unrestricted methods allow for the evaluation of the impact of the observed reform. However, ex-ante prediction seeks to predict the impact of, as yet, unobserved reforms. It requires the recovery of preference parameters to predict how physicians will react under alternative compensation systems.

5 Empirical Model

We now turn to developing our empirical model, adapting the theoretical model outlined in section 2 to the institutional details of the Quebec reform. We work with annual data and hence, specify preferences as a function of annual consumption, leisure and services, consistent with (5). We allow for two types of services: billable, denoted S^{B}, and non-billable, denoted S^{NB}. Recall that billable services are remunerated under both FFS and MR while non-billable services are remunerated only under FFS. Non-billable services will be supplied under MR if, for example, physicians gain utility from patient health (Arrow, 1963; Evans, 1974), or if such services are complements (or an input) in the production of billable services.[23] We allow for this possibility in estimating the model, treating the level of non-billable services observed under MR as a lower bound to the actual level supplied.

To account for the supply of time to administrative and teaching services under FFS, when they are not remunerated, we assume that they yield non-pecuniary benefits. For example, performing teaching tasks may increase influence and prestige.[24] To capture this in a simple and direct manner, we allow for two types of work in our model: clinical work, denoted by h^c, and non-clinical work, denoted by h^o, capturing time per week spent in administrative and teaching duties. We denote the total weekly hours by h^t (with $h^t = h^c + h^o$). Pure leisure is denoted by l.

[23] Fortin, Jacquemet, and Shearer (2008) provide the theoretical analysis of a model of physician behaviour in which utility depends on practice through the health produced, as the result of ethical concerns.

[24] An alternative would be to assume that these activities are complementary to billable services.

Physicians' preferences are represented by an annual utility function,[25]

$$u = u(X, h^o, l, S^{\mathrm{B}}, S^{\mathrm{NB}}) \tag{14}$$

defined over:

X (Annual income),
h^o (Weekly hours of administrative work and teaching)
l (Weekly hours of leisure outside of work),
S^{B} (Number of billable services supplied throughout the year),
S^{NB} (Number of non-billable services supplied throughout the year).

The weekly time constraint is given by:

$$l = T - h^c - h^o,$$

where $T = 24 \times 7 = 168$, the maximum amount of time available in a week. We also allow for differences in the marginal utility (or disutility) of billable and non-billable services.[26] The efficient budget constraint is obtained from the compensation system that maximizes net income, X, for each practice vector.

5.1 Discrete Alternatives

Given the non linearities in the efficient budget constraint after the MR reform, we follow the standard tradition in the empirical labour supply literature (van Soest, 1995; Hoynes, 1996) and discretize the physicians' choice set. For each variable describing the practice patterns of physicians, we consider a finite number of possible alternatives among which each physician can choose. We allow for $N_c = 4$ levels of clinical hours of work, $N_o = 4$ levels of non-clinical hours of work, $N_{S^{\mathrm{B}}} = 5$ levels of billable services, and $N_{S^{\mathrm{NB}}} = 5$ levels of non-billable services. The complete choice set of practice variables involves $dim(J) = N_c \times N_o \times N_{S^{\mathrm{B}}} \times N_{S^{\mathrm{NB}}} = 400$ alternatives. A single alternative, corresponding to one particular practice possibility, is a set of values: $j = \left\{c_j, o_j, S_j^{\mathrm{B}}, S_j^{\mathrm{NB}}\right\}$ respectively pointing to the ${c_j}^{th}$ level of discretized clinical hours of work, $c_j \in \{1, ..., N_c\}$, the ${o_j}^{th}$ level of discretized non-clinical hours of work, *etc*. The consumption under each alternative is computed through the efficient budget constraint, along which the physician maximizes utility.

[25]Recall that the number of weeks worked is set at 45 and therefore the number of weeks of leisure is equal to 6 in the utility function.

[26]This allows for the possibility that different types of services may be associated with different levels of difficulty and require different service intensity levels to complete the task. For example, an important element of non-billable services consists of follow-up visits by the physician, which check the progress of a patient after a particular treatment.

5.2 Choice Probabilities and the Utility Function

Let V_{ij} stand for the annual utility of physician i in alternative j. A standard assumption (McFadden, 1974) is to account for alternative-specific measurement errors on utility by decomposing V_{ij} into a deterministic component, u_j, and a random term which is independent across alternatives ϵ_{ij}. Thus,

$$V_{ij} = u_j + \epsilon_{ij}, \text{ where } \epsilon_{ij} \sim \text{ i.i.d. Gumbel (extreme value type I).}$$

Note that the random part of utility cannot be interpreted as reflecting unobservable heterogeneity since it is independent across alternatives; individual heterogeneity will be added in Section 5.4.1 below.

We specify utility as a quadratic function, which constitutes a second order approximation of any well-behaved utility function. While the approximation is not necessarily quasi-concave everywhere in its domain, this does not affect our empirical application; discretizing the choice set allows us to estimate the model without imposing quasi-concavity (van Soest, 1995). Appendix A.5 provides a step-by-step description of our econometric approach. Note finally that the discrete approach to estimating labour supply models requires the marginal utility of consumption to be positive at all chosen points along the budget constraint (van Soest, 1995). The quadratic utility function does not impose this condition. We will therefore check whether the latter is satisfied at the optimum for each individual and each period in our sample.

5.3 Identification

By influencing budget (but not preferences) parameters associated with various alternatives, the introduction of the MR reform helps identification. Thus, factors such as 1) the piece-rate difference between real billable FSS and billable MR services, 2) the piece-rate difference between real FSS and real (non-billable in MR) services, 3) the *per diem* difference between FFS and MR, affect the real monetary rewards associated with a number of alternatives. In a sense, our approach is similar to Blundell, Duncan, and Meghir (1998) who exploited tax reforms over time to identify a structural labour supply model. Changes over time in real income ceilings and in the consumer price index (for given levels of nominal piece-rates and other FFS parameters) also help parameter identification. Moreover, the structure of our model (our optimization approach, the introduction of both observed and unobserved heterogeneity, the functional form of the utility function, and the statistic distributions used) are all factors which make identification easier.[27]

While we do not formally prove that our model is identified, a simple (admittedly heuristic) identification demonstration is based on the proposition that, under weak regularity conditions, a parametric model is (locally) identified at a given vector θ if and only if the information matrix is non-singular at this point (Rothenberg, 1971). For each specification of our model (except a few

[27] An alternative identification strategy would be to use instruments that are correlated with the decision to participant in MR, but independent of preferences. Data limitation preclude us from applying these methods.

cases discussed later), the estimator that maximizes the likelihood function strongly converged towards the same parameter vector and gave rise to a non-singular estimated covariance matrix (the inverse of the information matrix). This suggests that our model is not only locally but globally identified.

5.4 Estimation issues

Several features of our data set necessitate modifications to the standard estimation methodology and likelihood function. First, since every combination of the discretized practice variables has to be considered as an alternative, the model allows for choices that contradict the technical constraint a physician faces. For example, a physician could theoretically choose to provide the highest available level of services while exerting zero hours of clinical work. Obviously such an alternative is not observed in our sample. For estimation purposes, we exclude those alternatives that are impossible in practice and, in concrete terms, never observed. We then estimate the model by reducing the choice set to the alternatives actually chosen in the sample: $J^C \subset J$, where $dim(J^C) = 314$. Note that this strategy leads us to use the same alternatives for estimation independent of the alternative that was chosen.

To account for the partial observability of non-billable services under MR, we integrate over all possible actual services that could have generated a given level of observed services. Let S_m^{NB} denote the level of non-billable services that is observed for a given physician (*i.e.* and delivered outside the *per diem* period). Since, for this observation, S_m^{NB} is a lower bound to the actual number of non-billable services provided, we observe S_m^{NB} whenever $S^{\text{NB}} \in \left\{ S_m^{\text{NB}}, S_{m+1}^{\text{NB}}, S_{m+2}^{\text{NB}}, \ldots, S_{N_{S^{\text{NB}}}}^{\text{NB}} \right\}$. What is more, since the different levels of non-billable services are mutually exclusive, the individual contribution to likelihood for an MR physician that chose the observable $\mathbf{z}_j, S_m^{\text{NB}}$ is obtained by summing over $S_l^{\text{NB}} \quad l \geq m$; *i.e.*,

$$P\left(\mathbf{z}_j, S_m^{\text{NB}}\right) = \frac{\exp(\boldsymbol{\gamma}'\mathbf{z}_j + \mathbf{z}_j'\boldsymbol{\beta}\mathbf{z}_j)}{\sum_{k=1}^{J^C} exp(u_k)} \sum_{l=m}^{N_{S^{\text{NB}}}} \exp(\gamma_{S^{\text{NB}}} S_l^{\text{NB}} + \mathbf{b}_A'.\mathbf{z}_j S_l^{\text{NB}} + \beta_{S^{\text{NB}}}(S_l^{\text{NB}})^2). \tag{15}$$

The traditional Logit probabilities are thus corrected for the uncertainty about the chosen alternative inside the chosen subset. The contribution to the likelihood of individual i is then given by

$$\mathcal{L}_{ij} = \left(\frac{exp(u_j)}{\sum_{k=1}^{J^C} exp(u_k)} \right)^{1-D_i} \left[P\left(\mathbf{z}_j, S_m^{\text{NB}}\right) \right]^{D_i}, \tag{16}$$

where D_i indicates whether a physician worked under MR ($D_i = 1$) or FFS ($D_i = 0$).

5.4.1 Heterogeneity in Preferences

We account for observable heterogeneity in preferences and productivity in the model, allowing the estimated coefficients of the utility function presented in (37) in Section A.5.1 to depend on individual characteristics. In particular, we allow the linear coefficient terms, γ, and the quadratic coefficient terms, $\boldsymbol{\beta}$, to be linear functions of age and gender

$$\gamma_i^k = \gamma_0^k + \gamma_1^k \times Age_i + \gamma_2^k \times DMale_i \quad k = \{o, l, L, S^{\text{B}}, S^{\text{NB}}, X\}, \tag{17}$$

$$\beta_i^k = \beta_0^k + \beta_1^k \times Age_i + \beta_2^k \times DMale_i \quad k = \{o, l, L, S^{\text{B}}, S^{\text{NB}}, X\}, \tag{18}$$

where $DMale$ is a dummy variable indicating male physicians.

We account for unobservable heterogeneity by adding normally distributed random terms to the functions in (17). Define

$$\widetilde{\boldsymbol{\gamma}}_{\boldsymbol{i}} = (\widetilde{\gamma}_i^o, \widetilde{\gamma}_i^l, \widetilde{\gamma}_i^{\text{B}}, \widetilde{\gamma}_i^X)$$

to be the vector of random coefficients, where

$$\widetilde{\gamma}_i^{k'} = \gamma_0^{k'} + \gamma_1^{k'} \times Age_i + \gamma_2^{k'} \times DMale_i + \eta_i^{k'} \quad k' = o, l, S^{\text{B}}, X.$$

We assume that $\eta_i^{k'} \sim N(0, \sigma_{k'})$ and that the η's are mutually independent, and independent of ϵ_j, $\forall j$. Conditional on the $\widetilde{\boldsymbol{\gamma}}_{\boldsymbol{i}}$'s, the contributions to the likelihood are given by

$$l_{ij}(\widetilde{\boldsymbol{\gamma}}_{\boldsymbol{i}}, \boldsymbol{\beta}_{\boldsymbol{i}}) = \left(\frac{exp(u_{ij})}{\sum_{k=1}^{J^C} exp(u_{ik})} \right)^{1-D_i} \left[P_{ij}\left(\mathbf{z}_j, S_m^{\text{NB}}\right) \right]^{D_i}, \tag{19}$$

where the utility index now depends on i to incorporate both observed and unobserved heterogeneity. The unconditional probabilities correspond to the mixed logit specification. One potentially important advantage of the latter is that it does not impose the independence of irrelevant alternative (IIA) condition as does the standard multinomial logit.

The estimation of this model requires the computation of a large number of four-dimensional integrals. To calculate these integrals we rely on simulated Maximum Likelihood estimator. The estimator derived from this specification is asymptotically equivalent to an exact ML estimator given that $\sqrt{r}$, where r is the number of draws, rises faster than the size of the sample (Gourieroux and Monfort, 1993).[28] The panel dimension of our data is taken into account by allowing for repeated choices by each physician. In this context the random draws generating individual heterogeneity are constant across time for a given physician, but vary across physicians. The errors, ϵ_{ij}, are iid.

[28] The draws were generated using Halton sequences that produces lower simulation variance for given r (Train, 1999).

5.4.2 Calculating income

We identify the utility-function parameters by restricting the observed decisions to be optimal choices. This requires calculating the utility associated with each alternative available to a physician; *i.e.*, each $j \in J$. Since each physician is only observed in one state in a given period, and since different states imply different income levels, estimation requires calculating the counterfactual income levels for each of the unobserved states. To do so, we rely on our discussion of the budget constraint presented in Section 3. In particular we use equations (11) to (13) to predict the income for each alternative. Recall that the only level of income we observe is the one earned by each physician in his/her chosen alternative, under his/her current compensation scheme.

Recall from our discussion in Section 3, we aggregated services provided into two types: billable, denoted S^{B}, and non-billable, denoted S^{NB}. Under FFS, both types of services are paid at the same aggregate price, denoted p_t. Consumption in alternative j, in year t, under FFS is then given by

$$X_{j,t}^{FFS} = p_t(S_j^{\text{B}} + S_j^{\text{NB}}). \tag{20}$$

Calculating consumption under MR based on (12) is somewhat more complex, since payment for services depends on whether the service is provided under the *per diem* or not (recall that billable services provided under the *per diem* were paid a lower fee, αp). Appendix A.5.2 provides full details about how we account for these issues.

5.4.3 Constrained Choice

Recall that the actual choice of a compensation system was not individual specific. Rather, members of specialist departments within each hospital determined the compensation system by vote, only adopting the MR system if the vote was unanimously in favour. This raises the possibility that some physicians may be constrained in their choice of a compensation system and, hence, not be located on the efficient budget constraint.[29] However only those physicians who prefer MR are potentially constrained; those who prefer FFS are ensured their unconstrained choice since the voting rule is unanimous. This implies that physicians who are observed on sections of the efficient budget constraint under MR are not constrained. Physicians observed under FFS can be divided into two groups: those who are observed in an alternative j for which $X_j^{MR} > X_j^{FFS}$ are constrained. Those who select alternatives for which $X_j^{MR} < X_j^{FFS}$ are potentially constrained.[30]

To account for constraints on choices, let ψ denote the probability that a physician is constrained from attaining the efficient budget constraint.[31] We then define the following observed regimes:

[29] We do see a number of physicians (30 in 2002) who are paid FFS contracts when they would earn higher income under MR, for the same practice variables.

[30] The unanimity rule applies to movement back from MR to FSS. However, in practice, no department switched from MR to FFS in our sample.

[31] An alternative approach to modelling the choice restriction would directly introduce the work contract into the utility

$\mathcal{R}_1$ the physician is observed FFS when only FFS is available (*i.e.*, pre-reform observations);

$\mathcal{R}_2$ the physician is observed MR when MR dominates;

$\mathcal{R}_3$ the physician is observed FFS when MR dominates;

$\mathcal{R}_4$ the physician is observed FFS when FFS dominates.

We disregard the case of physicians observed MR while FFS dominates which is ruled out by assumption.[32]

Let D_{ij} indicate the presence of physician i in regime $\mathcal{R}_j, \forall j \in \{1,2,3,4\}$. A constrained physician selects his optimal labour supply alternative along the FFS budget constraint rather than the efficient budget constraint denoted Eff. We therefore redefine utility to account for the relevant budget constraint. Let $u_{ij}^{\mathcal{B}}$ denote the utility derived by physician i from alternative j when income is computed under budget constraint $\mathcal{B} \in \{FFS, Eff\}$ and let

$$P^{\mathcal{B}}\left(\mathbf{z}_j, S_m^{\text{NB}}\right) \tag{21}$$

denote the probability of observing a given alternative $\left(\mathbf{z}_j, S_m^{\text{NB}}\right)$ for an MR physician, from (15). The individual contribution to the likelihood function is given by

$$\begin{aligned} l_i(\widetilde{\boldsymbol{\gamma}}_i, \boldsymbol{\beta}_i) = & \left[\frac{exp(u_{ij}^{FFS})}{\sum_{k\in J} exp(u_{ik}^{FFS})}\right]^{D_{i1}} \\ & \times \left[(1-\psi)\, P^{Eff}\left(\mathbf{z}_j, S_m^{\text{NB}}\right)\right]^{D_{i2}} \\ & \times \left[\psi\, \frac{exp(u_{ij}^{FFS})}{\sum_{k\in J} exp(u_{ik}^{FFS})}\right]^{D_{i3}} \\ & \times \left[\psi\, \frac{exp(u_{ij}^{FFS})}{\sum_{k\in J} exp(u_{ik}^{FFS})} + (1-\psi)\, \frac{exp(u_{ij}^{Eff})}{\sum_{k\in J} exp(u_{ik}^{Eff})}\right]^{(1-D_{1i}-D_{2i}-D_{3i})}. \end{aligned} \tag{22}$$

function so that choices provide the different utility depending on whether they are completed under MR or FFS. One problem with this approach is that it multiplies by two the number of alternatives. More importantly, it is *ad hoc* as it does not take into account the nature of the institutional constraint.

[32]There are only 10 observations that fall into this category; they are classified in $\mathcal{R}_2$. One interpretation of this case is that these physicians make optimization errors. These regimes only apply to observations post-reform, which are restricted to 2002 observations in our sample. To ease exposition, we thus restrict here to contributions to the likelihood of 2002 observations (i.e., conditional on the contribution of the same physician in all years before), and ignore the dependency on time in the notation.

The likelihood function reflects the fact that the constraints on behaviour only apply to regimes $\mathcal{R}_2$—$\mathcal{R}_4$ since $\mathcal{R}_1$ occurs before the reform. Physicians in regime $\mathcal{R}_2$ are unconstrained which occurs with probability $(1-\psi)$. The physicians in regime $\mathcal{R}_3$ are constrained which occurs with probability ψ. The physicians in regime $\mathcal{R}_4$ can be either constrained or unconstrained.[33]The main source of empirical identification of this probability thus is the share of physicians whose utility maximizing alternative is paid MR while they are observed in a FFS one.

6 Results

6.1 Parameter Estimates

We estimated three versions of the quadratic utility function: first, with observed heterogeneity (*i.e.*, with age and gender); second, with observed and unobserved heterogeneity (a random term being added to all the linear terms of the parameters, γ^k), and third, with observed heterogeneity on all parameters, but unobserved heterogeneity restricted to a subset of the linear parameters. Each case incorporates constrained choice of the compensation system—the contribution to the likelihood of observation i, conditional on $\widetilde{\gamma}_i$, is given by (22).[34] The results are presented in Table 3.

The first column presents results when observed heterogeneity is introduced into the linear and quadratic terms for non-clinical hours worked per week (h^o), hours of leisure per week (l), non-billable services (S^{NB}), billable services (S^{B}) and income (X). These coefficients are permitted to vary with age and gender. The second column introduces unobserved heterogeneity. In this specification, a random term is added to the parameters on the linear terms (in addition to being functions of age and gender). The standard error of this error term is reported accordingly. Finally, the third column is the same as the second one but in which we parsimoniously assume that some linear parameters are not random.

Recall that the discrete approach to estimating labour supply models requires the marginal utility of consumption to be positive at all chosen points along the budget constraint. This requirement is satisfied for 76% of observations in the model with observed heterogeneity. It is satisfied for 57% of the observations in the model with both observed and unobserved heterogeneity on all linear parameters. After experimentation, we found that a higher percentage (= 83%) is reached by a model in which unobserved heterogeneity is restricted to the linear parameters of non-clinical hours (h^o), leisure (l), and billable services (S^{B}). In the interests of selecting a model that best fits the data, while respecting theoretical restrictions, we concentrate on this last version.

[33]The probability of being constrained is fixed and does not depend on observable variables such as age and gender. This assumption has been made for parsimony, but also reflects the fact that preliminary results showed that the estimated coefficients of these variables were not significant.

[34]We also estimated the model without taking account of constrained choice. The results for these specifications were generally less satisfactory than those presented. However, we used these results to simulate the impact of making MR system individual rather than group free (see section 6.3).

6.2 Preference parameters

Note that we choose this specification in spite of the fact that the likelihood function increased substantially (from -4386 to -4290) upon assuming that all linear parameters are random (version 2). This reflects the tradeoff between fitting the sample data and estimating economic models. Our selection criteria is in the spirit of a more general strategy of selecting the best fitting model among the set of models for which the theoretical restrictions are not rejected. Moreover, as a robustness check, we performed the simulation of the treatment effects using the latter, full specification. The results are very similar to those we present here.[35] Interestingly, whatever the specification considered, the probability, ψ, that a physician is constrained from attaining the efficient budget constraint is very high (=0.46 in all specifications) and highly significant. This suggests that introducing a reform allowing physicians to choose their compensation system individually may have a strong effect on their behaviour. We will return to these issues in section 6.3.3.

Table 4 compares the predictions with observed choices for the year 2002 (the last post-reform year in our sample). The middle column of the table gives the average predicted value of the different choice variables of the model. The last column gives the average observed values of these same variables. On the whole, the model's fit is very good. The average (combined) hours worked per week, average clinical hours worked, average number of billable services, average service intensity (time spent per clinical service), and average income are all matched very closely by the model's predictions. The model has more trouble matching the number of non-clinical hours worked.[36]

6.3 Policy simulations

Estimation of a structural model allows us to simulate the impact of different compensation policies on physician behaviour. Different compensation policies imply different budget constraints, which in turn affect the probabilities of selecting different practice alternatives. Given knowledge of the preference parameters we simply calculate the (expected) predicted behaviour on the basis of the revised budget constraint. We compute bootstrapped standard errors of this predicted behavior, obtained by repeated random draws of the model parameters from their estimated distributions and by recalculating predicted behavior for each draw.

Table 3: Preference Parameters

	Homogenous MNL		Mixed Logit: on linear terms		Mixed Logit: preferred specification					
					All physicians		MR physicians		FFS physicians	
	Coef.	(St.d.)	Coef.	(St.d.)	Coef.	(St.d.)	Coef.	(St.d.)	Coef.	(St.d.)
γ^o	3.65e-01***	(0.071)	7.78e-01***	(0.140)	8.25e-01***	(0.157)	8.17e-01***	(0.139)	8.32e-01***	(0.135)
σ^o	—	—	1.40e-01***	(0.035)	1.50e-01***	(0.023)	1.50e-01***	(0.025)	1.50e-01***	(0.026)
$\gamma^o \times Male$	—	—	2.40e-01***	(0.056)	2.21e-01***	(0.056)	2.21e-01***	(0.055)	2.21e-01***	(0.054)
$\gamma^o \times Age$	—	—	-8.58e-02***	(0.028)	-7.82e-02***	(0.027)	-7.82e-02***	(0.028)	-7.82e-02***	(0.023)
γ^l	6.87e-01***	(0.067)	8.57e-01***	(0.228)	1.09e+00***	(0.258)	1.09e+00***	(0.262)	1.08e+00***	(0.272)
σ^l	—	—	9.33e-02***	(0.016)	1.06e-01***	(0.015)	1.06e-01***	(0.015)	1.06e-01***	(0.016)
$\gamma^l \times Male$	—	—	1.25e-02	(0.123)	-2.06e-02	(0.121)	-2.06e-02	(0.121)	-2.06e-02	(0.129)
$\gamma^l \times Age$	—	—	7.33e-02*	(0.046)	4.91e-02	(0.052)	4.91e-02	(0.054)	4.91e-02	(0.056)
$\gamma^{S^{NB}}$	4.00e-02***	(0.014)	1.31e-02	(0.039)	5.77e-02**	(0.025)	5.77e-02***	(0.019)	5.77e-02***	(0.022)
$\gamma^{S^{NB}} \times Male$	—	—	1.76e-02	(0.015)	1.36e-02	(0.011)	1.36e-02	(0.011)	1.36e-02	(0.011)
$\gamma^{S^{NB}} \times Age$	—	—	3.92e-03	(0.009)	-2.81e-03	(0.004)	-2.81e-03	(0.003)	-2.81e-03	(0.003)
γ^{S^B}	6.59e-02***	(0.014)	6.19e-02*	(0.039)	1.88e-01***	(0.021)	1.89e-01***	(0.026)	1.82e-01***	(0.039)
σ^{S^B}	—	—	8.35e-02***	(0.010)	9.33e-02***	(0.011)	9.33e-02***	(0.011)	9.33e-02***	(0.013)
$\gamma^{S^B} \times Male$	—	—	5.20e-03	(0.025)	2.53e-02	(0.026)	2.53e-02	(0.024)	2.53e-02	(0.023)
$\gamma^{S^B} \times Age$	—	—	1.84e-02***	(0.006)	3.14e-03	(0.005)	3.14e-03	(0.007)	3.14e-03	(0.005)
γ^x	4.40e-02***	(0.012)	1.65e-01***	(0.048)	3.88e-02***	(0.012)	3.88e-02***	(0.014)	3.88e-02***	(0.013)
σ^x	—	—	3.58e-02***	(0.005)	—	—	—	—	—	—
$\gamma^x \times Male$	—	—	-6.46e-03	(0.019)	-2.46e-03	(0.014)	-2.46e-03	(0.013)	-2.46e-03	(0.014)
$\gamma^x \times Age$	—	—	-1.18e-02	(0.010)	—	—	—	—	—	—
β^o_l	-2.87e-03***	(0.001)	-3.37e-03***	(0.001)	-3.92e-03***	(0.001)	-3.92e-03***	(0.001)	-3.92e-03***	(0.001)
$\beta^o_{S^{NB}}$	8.13e-05	(0.000)	-2.04e-04	(0.000)	-8.13e-05	(0.000)	-8.13e-05	(0.000)	-8.13e-05	(0.000)
$\beta^o_{S^B}$	-8.72e-04***	(0.000)	-1.48e-03***	(0.000)	-1.60e-03***	(0.000)	-1.60e-03***	(0.000)	-1.60e-03***	(0.000)
β^o_x	-1.38e-04	(0.000)	5.98e-05	(0.000)	-8.57e-05	(0.000)	-8.57e-05	(0.000)	-8.57e-05	(0.000)
$\beta^l_{S^{NB}}$	-2.27e-04**	(0.000)	-1.70e-04	(0.000)	-4.03e-04***	(0.000)	-4.03e-04***	(0.000)	-4.03e-04***	(0.000)
$\beta^l_{S^B}$	-2.83e-04***	(0.000)	-3.99e-04**	(0.000)	-7.27e-04***	(0.000)	-7.27e-04***	(0.000)	-7.27e-04***	(0.000)
β^l_x	-1.95e-04**	(0.000)	-3.55e-04**	(0.000)	—	—	—	—	—	—
$\beta^{S^{NB}}_{S^B}$	1.07e-04***	(0.000)	3.20e-05	(0.000)	4.59e-05	(0.000)	4.59e-05	(0.000)	4.59e-05	(0.000)
$\beta^{S^{NB}}_x$	-1.24e-04***	(0.000)	1.03e-06	(0.000)	—	—	—	—	—	—
$\beta^{S^B}_x$	-7.19e-05***	(0.000)	9.74e-05*	(0.000)	—	—	—	—	—	—
β^o	-6.96e-04	(0.001)	-1.29e-02***	(0.004)	-1.13e-02***	(0.004)	-1.13e-02***	(0.004)	-1.13e-02***	(0.003)
$\beta^o \times Male$	—	—	-4.64e-03***	(0.001)	-3.94e-03***	(0.001)	-3.94e-03***	(0.001)	-3.94e-03***	(0.001)
$\beta^o \times Age$	—	—	2.32e-03***	(0.001)	1.85e-03***	(0.001)	1.85e-03***	(0.001)	1.85e-03***	(0.001)
β^l	-2.47e-03***	(0.000)	-3.36e-03***	(0.001)	-4.26e-03***	(0.001)	-4.26e-03***	(0.001)	-4.26e-03***	(0.001)
$\beta^l \times Male$	—	—	3.16e-05	(0.001)	1.63e-04	(0.000)	1.63e-04	(0.001)	1.63e-04	(0.001)
$\beta^l \times Age$	—	—	-2.51e-04*	(0.000)	-1.60e-04	(0.000)	-1.60e-04	(0.000)	-1.60e-04	(0.000)
$\beta^{S^{NB}}$	-3.15e-05*	(0.000)	-1.50e-04	(0.000)	-1.43e-04*	(0.000)	-1.43e-04**	(0.000)	-1.43e-04**	(0.000)
$\beta^{S^{NB}} \times Male$	—	—	3.90e-06	(0.000)	-2.82e-05	(0.000)	-2.82e-05	(0.000)	-2.82e-05	(0.000)
$\beta^{S^{NB}} \times Age$	—	—	-7.40e-06	(0.000)	1.53e-05	(0.000)	1.53e-05	(0.000)	1.53e-05	(0.000)
β^{S^B}	-1.25e-04***	(0.000)	-3.42e-04***	(0.000)	-5.79e-04***	(0.000)	-5.79e-04***	(0.000)	-5.79e-04***	(0.000)
$\beta^{S^B} \times Male$	—	—	1.69e-04	(0.000)	1.87e-04*	(0.000)	1.87e-04**	(0.000)	1.87e-04**	(0.000)
$\beta^{S^B} \times Age$	—	—	-9.88e-05***	(0.000)	-6.17e-05**	(0.000)	-6.17e-05	(0.000)	-6.17e-05***	(0.000)
β^x	—	—	-3.14e-04***	(0.000)	-1.06e-04***	(0.000)	-1.06e-04***	(0.000)	-1.06e-04***	(0.000)
$\beta^x \times Male$	—	—	2.47e-05	(0.000)	2.21e-05	(0.000)	2.21e-05	(0.000)	2.21e-05	(0.000)
$\beta^x \times Age$	—	—	1.51e-05	(0.000)	—	—	—	—	—	—
ψ	4.58e-01***	(0.067)	4.55e-01***	(0.066)	4.54e-01***	(0.067)	4.54e-01***	(0.067)	4.54e-01***	(0.067)
LL	-5098.1		-4290.2		-4385.6		-4390.26		–4386.86	

Table 4: Model Fit

	Observed Total	Predicted 2002	Observed 2002
Weekly Total Hours	44.77	45.93	43.88
____ clinical (h^c)	39.50	39.70	38.78
____ non clinical (h^o)	5.27	6.23	5.10
Total Services[a]	147.42	145.57	144.38
____ Non-billable (S^{NB})	64.27	61.61	63.59
____ Billable (S^{B})	83.15	83.96	80.78
Service Intensity $\left(e = \frac{S^{\mathrm{NB}}+S^{\mathrm{B}}}{h^c * W}\right)$	82.93	81.49	82.74
Annual Income[a] (X)	143.71	149.18	142.76

[a]Thousands of (1996) Can. Dollars.
Note. The cells display the average practice behavior (in terms of practice variables) observed over the whole sample period (*first column*) and in 2002 (*last column*). The *second column* reports the average practice behavior predicted by the model in 2002.

6.3.1 Elasticities through policy simulations

Table 5 provides results on elasticities of practice variables with respect to non-labour income, hourly wage rate, and fee per service.[37] The second column provides our benchmark; it is computed as the average practice choice simulated from the estimated model against a simplified budget constraint, broadly representative of the prevailing case before the reform. We assume an hourly wage rate equal to \$10, the full fee under FFS on all clinical services, and an exogenous non-labour income equal to \$10,000.[38] We remove all the other parameters that may affect a physician's budget constraint (*e.g.*, income ceilings and regionally differentiated remuneration). The physician's budget is thus linear in (w, p, y) with all arguments strictly positive. As the MR reform involved substantial changes in the fee per service and wage parameters, for comparison sake, we also performed our elasticity simulations based on large (50%) percentage changes in each of these parameters. Similarly, the computation of the income elasticity, $\varepsilon_{k/y}$, for each practice variable, k, is based on the variation in practice induced by a 50% increase in non-labour income. Also, we use Slutsky decompositions of uncompensated elasticities into compensated and total income elasticities:

[35]In spite of the difference in fit, the simulated treatment effects from the two models are quite similar. For example, under the observed reform, total services decrease by 6.5% (5.32%), weekly hours increase by and 0.34% and 0.24% and services intensity decreases by 4.78% and 6.25% under the preferred model and full specification, respectively.

[36]Caution should be exercised in interpreting the statistics over non-billable services. Recall that the recorded volume of non-billable services is a lower bound to the actual volume of services completed. We calculate the observed volume as the recorded volume divided by the (estimated) probability that the physician provides additional services.

[37]The reader should bear in mind that an important difference between the elasticity simulations and the actual reform is that, under the actual reform, the *per diem* (hourly wage) simultaneously becomes positive.

[38]We add small positive hourly wage and non-labour income to the observed FFS contract in order to allow us to simulate elasticities at the benchmark.

Table 5: Elasticity of practice variables

	Ref.	Non-labour income		Hourly wage rate				Service piece-rate			
		$\Delta\, y$	$\varepsilon_{k/y}$	$\Delta\, W$	$\varepsilon_{k/w}$	$\widetilde{\varepsilon}_{k/w}$	$\frac{whW}{y}\varepsilon_{k/y}$	$\Delta\, IP$	$\varepsilon_{k/IP}$	$\widetilde{\varepsilon}_{k/IP}$	$\frac{PA}{y}\varepsilon_{k/y}$
Weekly Total Hours	45.62	45.56	-0.003	45.63	0.000	0.053	-0.053	44.76	-0.038	0.052	-0.090
	(1.06)	(1.06)	(0.0000)	(1.06)	(0.0000)	(0.0004)	(0.0004)	(1.07)	(0.0002)	(0.0006)	(0.0013)
_____ clinical (h^c)	39.77	39.69	-0.004	39.77	-0.000	0.065	-0.065	38.87	-0.045	0.065	-0.110
	(0.90)	(0.89)	(0.0000)	(0.90)	(0.0000)	(0.0008)	(0.0008)	(0.88)	(0.0005)	(0.0008)	(0.0024)
_____ non clinical (h^o)	5.86	5.86	0.002	5.86	0.000	-0.029	0.030	5.89	0.012	-0.039	0.051
	(0.59)	(0.58)	(0.0002)	(0.59)	(0.0000)	(0.0480)	(0.0482)	(0.66)	(0.0230)	(0.0513)	(0.1420)
Total Services[a]	150.18	148.61	-0.021	150.11	-0.001	0.367	-0.368	131.07	-0.254	0.373	-0.627
	(4.02)	(4.04)	(0.0000)	(4.02)	(0.0000)	(0.0088)	(0.0088)	(5.95)	(0.0047)	(0.0124)	(0.0258)
_____ Non-billable (S^{NB})	65.12	63.88	-0.038	65.06	-0.002	0.667	-0.669	50.69	-0.443	0.697	-1.140
	(2.90)	(2.90)	(0.0001)	(2.90)	(0.0000)	(0.0287)	(0.0289)	(4.62)	(0.0126)	(0.0437)	(0.0841)
_____ Billable (S^{B})	85.06	84.72	-0.008	85.04	-0.000	0.137	-0.137	80.38	-0.110	0.124	-0.234
	(2.99)	(2.97)	(0.0000)	(2.99)	(0.0000)	(0.0014)	(0.0014)	(2.94)	(0.0014)	(0.0013)	(0.0042)
Service intensity $\left(=\frac{S^{NB}+S^{B}}{h^c * W}\right)$	75.53	74.88	-0.017	75.50	-0.001	0.303	-0.304	67.44	-0.214	0.303	-0.518
	(1.50)	(1.51)	(0.0000)	(1.50)	(0.0000)	(0.0061)	(0.0061)	(2.61)	(0.0029)	(0.0090)	(0.0178)
Annual income[a] (X)	142.36	145.20	0.040	142.49	0.002	-0.701	0.703	184.26	0.589	-0.610	1.199
	(3.57)	(3.59)	(0.0000)	(3.57)	(0.0000)	(0.0076)	(0.0076)	(7.80)	(0.0089)	(0.0075)	(0.0216)

[a] Thousands of (1996) Can. Dollars.

Note. Elasticities of practice variables simulated fromÂ estimated preferences. In the reference situation, physicians are paid the full fee under FFS on all clinical services, an hourly wage rate equal to \$10 and an exogenous non-labour income equal to \$10,000. Elasticities are computed from a 50% change in each parameter of the resulting budget constraint—for each parameter, the first column displays predicted average behaviour from the updated budget constraint. Bootstrapped standard errors appear in parenthesis.

$\varepsilon_{k/w} = \widetilde{\varepsilon}_{k/w} + wh^t\frac{\mathcal{W}}{y}\varepsilon_{k/y}$ and $\varepsilon_{k/p} = \widetilde{\varepsilon}_{k/p} + \frac{pS}{y}\varepsilon_{k/y}$, and where $\mathcal{W}$ is set at 45 weeks of work, to compute the wage and fee per service compensated elasticities of each practice variable.[39]

Results from the second panel of Table 5 indicate that physicians' average clinical weekly hours of work, and the volume of (billable and non-billable) services are negatively affected (with $p < 0.01$) by an increase in non-labour income. However, non-clinical hours of work increase with non-labour income (with $p < 0.01$). This may partly be explained by the fact that this activity yields important non-pecuniary benefits to the physician and that these benefits are normal goods. Overall, the simulated elasticities are modest (in absolute value) though, ranging between -0.003 for weekly hours of work and -0.021 for services. Moreover, physicians' service intensity, as measured by the volume of services provided (in 1996 Can. dollars) per clinical hour of work, decreases with non-labour income but very slightly, with an elasticity of -0.017 (with $p < 0.01$).

Results from the third panel indicate that uncompensated own wage elasticity of total weekly hours is close to 0. This suggests that physicians' labour supply curve for weekly hours is essentially vertical. Moreover, the compensated own wage elasticity is positive, although quite small, being estimated at 0.053 (with $p < 0.01$). Our results also indicate that services and hours of work are net complements, as cross compensated wage elasticity of services is positive (= 0.367, with $p < 0.01$).

The last panel provides results regarding elasticities with respect to changes in the fee for service (piece rate). The own uncompensated service elasticity is negative and equal to -0.254, with $p < 0.01$. Thus, the labour supply curve for services is backward-bending. Interestingly, the negative effect of an increase in the fee per service is much larger (in absolute value) on non-billable services

[39] This is an approximation since the choice set is discrete and the variations in wage and fee per service are not infinitesimal.

($= -0.443$) than on billable services ($= -0.110$). The compensated own service elasticity is positive as expected and quite large and significant ($= 0.373$). Notice also that the compensated elasticity of weekly hours of work with respect to fee per service is positive but small ($= 0.052$). As expected, a compensated increase in the fee per service induces the physician to spend less time in non-clinical (teaching and administrative) activities and more time to perform clinical services, but again these effects are small (-0.039 and 0.065, respectively). These results suggest that compensated changes in the fee for service have a positive and significant impact on physicians' behaviour—especially on the volume of their services and their service intensity.

Our results on elasticities suggest that physicians (pediatricians) react to incentives in the directions predicted by the theory. The compensated own elasticities are all positive and the effects of non-labour income are all negative on weekly hours of work (except on non-clinical hours) and on services. However, the compensated and uncompensated, weekly hours elasticities with respect to wage and the fee for service are very small. This result is consistent with those reported in studies focusing on hours of work supplied by physicians who are not self-employed: for example, Sloan (1975); Noether (1986); Saether (2005) found that the wage elasticities are modest or non-signifiant in this context. The (compensated) service (and service intensity) elasticities with respect to wage and the fee for service are considerably larger (in absolute value) than those on hours. Finally, we note that the incentive effects on services provided are generally much larger (in absolute value) than are those on hours worked. This observation, which will reappear as a common theme throughout our policy simulations, has possibly important implications for the quality of services provided and is consistent with the results of Dumont, Fortin, Jacquemet, and Shearer (2008).[40]

6.3.2 The Observed Reform

We begin our analysis of different reforms by simulating the effects of the observed policy—the introduction of the MR system as a constrained choice on the part of physicians. We compare predicted behaviour under FFS (the first column of Table 6) to that under the MR system, taking account of the probability of being constrained. The budget constraint under MR is then the mixture of the constrained budget constraint and the unconstrained (efficient) budget constraint. The results are given in the second column of Table 6, labelled "Group Free MR". Note that the results correspond to the average treatment effects (ATE) of the reform and are not to be confounded with those of the descriptive statistics in Table 2 which approximate the Average Treatment Effects on the Treated (ATT). These results are instructive in many ways. First, notice the reform increased the number of weekly hours worked very slightly, by 0.34%. Moreover, this is entirely due to increases in non-clinical hours which rose by 6.52%; clinical hours in fact decreased by 0.57%. This suggests that the *per diem* incorporated into the MR payment system did induce physicians to spend more

[40] Shearer, Somé, and Fortin (2017) measures the elasticity of hours and services with respect to price changes at a less aggregated level, among FFS physicians. He also found that services are more sensitive to price changes than hours.

Table 6: Treatment effects of MR

	FFS	Group Free MR		Individual Free MR		Mandatory MR	
		Practice	Variation	Practice	Variation	Practice	Variation
Weekly Total Hours	45.77	45.93	0.34%	46.06	0.62%	46.08	0.68%
	(1.06)	(0.978)	(0.000)	(0.937)	(0.000)	(0.935)	(0.000)
____ clinical (h^c)	39.92	39.70	-0.57%	39.51	-1.04%	39.53	-0.99%
	(0.92)	(0.866)	(0.000)	(0.932)	(0.000)	(0.931)	(0.000)
____ non clinical (h^o)	5.85	6.23	6.52%	6.55	11.96%	6.55	12.04%
	(0.62)	(0.572)	(0.006)	(0.725)	(0.018)	(0.718)	(0.017)
Total Services[a]	153.75	145.57	-5.32%	138.76	-9.75%	139.86	-9.03%
	(4.16)	(4.374)	(0.000)	(5.351)	(0.001)	(5.601)	(0.001)
____ Non-billable (S^{NB})	67.97	61.61	-9.36%	56.31	-17.15%	57.26	-15.75%
	(3.05)	(3.050)	(0.001)	(3.716)	(0.002)	(3.839)	(0.002)
____ Billable (S^{B})	85.78	83.96	-2.12%	82.45	-3.88%	82.60	-3.71%
	(3.03)	(2.961)	(0.000)	(3.008)	(0.000)	(3.045)	(0.000)
Service intensity $\left(=\frac{S^{NB}+S^{B}}{h^c * W}\right)$	77.02	73.34	-4.78%	70.24	-8.80%	70.76	-8.13%
	(1.55)	(1.720)	(0.000)	(2.373)	(0.001)	(2.542)	(0.001)
Annual income[a] (X)	135.68	149.18	9.95%	160.41	18.23%	158.21	16.61%
	(3.66)	(4.109)	(0.000)	(4.474)	(0.000)	(4.454)	(0.000)

[a] Thousands of (1996) Can. Dollars.
Note. Average practice behaviour predicted by the model in 2002 depending on whether physicians are paid according to: a mandatory Fee-for-Service (*first column*) ; the Mixed Remuneration scheme chosen conditionally on group agreement (*second column*) ; an MR system freely chosen on an individual basis (*third column*) ; or a mandatory MR (*last column*). The percentage variation provided for each compensation scheme takes FFS as a benchmark. Bootstrapped standard errors appear in parenthesis.

time on administrative and teaching activities. The reform also had important effects on the volume of services provided. Physicians reduced their supply of services in the order of -5.32%. This reflects physicians responding to (large) monetary incentives.

As with the elasticities, we see that services are more sensitive than hours seeing patients. The MR compensation system reduced the marginal payment for services received by physicians (on average by 30%) and hence the marginal benefit to their completion. This substitution effect is accentuated by the negative income effect on the volume of services associated with the higher annual income received by MR physicians. Indeed, the physician annual income increased on average by nearly 10%. This reflects the large *per diem* payments that MR physicians received, independent of the number of services provided. The fact that the reform was expensive also raises the question as to whether or not it could have been enacted for lower cost. We return to this point below in Section 6.3.5. Our results show that service intensity decreased (by 4.78%) with the reform which suggests that physicians spent more time with their patients under MR.

6.3.3 Mandatory MR Reform

Given the voluntary nature of the observed reform, a natural question to address is how a mandatory MR reform would affect behaviour. We address this within the context of our model by simulating optimal choices along the MR budget constraint. We then compare the resulting predicted behaviour to that under FFS. The results are presented in the fourth column of Table 6. They suggest that a mandatory reform would have had considerable effects on services provided (a decrease of 9.03% relative to FFS) and non-clinical hours (an increase of 12.04% relative to FFS); these are much larger than under the observed reform. Physicians would also spend more time with patients – services per hour worked seeing patients would decrease by 8.13% relative to FFS. The cost of the program would also be significantly affected (an increase of 16.61% relative to FFS).

The mandatory reform changes two things vis-à-vis the observed reform: first, it removes the choice of the compensation system and second it removes constraints on an individual's choice of a MR compensation system. To decompose the overall effect into its component parts, we simulated the observed voluntary reform, removing the constraint on choice. We set $\psi = 0$, allowing physicians to choose their compensation system individually along the efficient budget constraint. The subsequent predicted behaviour is compared to behaviour under FFS. The results are given in the the third column of Table 6, labelled "Individual Free MR." They are close to the results from the mandatory reform. This suggests that constraints on choice are the most important factor in explaining the difference between the actual and mandatory reforms. Even though workers who switched to MR were low-productivity physicians, many high-productive physicians—who would have reacted strongly to the change in compensation system— would have switched to MR if they had not been constrained in their choice. Physicians who are currently observed under FFS can (on average) find a practice pattern under MR that provides them with higher income and that they prefer, but they are constrained from choosing it. Geometrically, this suggests that the line *MR* in Figure 1 should be shifted upward so that a large number of pediatricians would choose the MR system if they were free to do so.[41]

6.3.4 Selection Effects

Related to the previous subsection, our mixed-logit model allows us to evaluate the consequences of selection on unobservables on the effect of various reforms. To this end, we use our model to recover preferences that are specific to the sub-population of physicians choosing each of the two compensation schemes. MR preferences are computed by drawing the random preference parameters—non-clinical hours (h^o), leisure (l), and billable services S^{B}—from the distribution of preferences, conditional on working under MR. We thus calculate the average vector of utility parameters for this

[41] This is consistent with the fact that pediatrics is one of the specialities that pushed very hard on government to introduce the MR system.

Table 7: Treatment effects: MR-specific preferences

	FFS	Group Free MR		Individual Free MR		Mandatory MR	
		Practice	Variation	Practice	Variation	Practice	Variation
Weekly Total Hours	46.67	46.82	0.32%	46.94	0.58%	46.97	0.63%
	(1.26)	(1.168)	(0.000)	(1.119)	(0.000)	(1.117)	(0.000)
clinical (h^c)	40.24	39.98	-0.64%	39.77	-1.18%	39.79	-1.13%
	(1.05)	(0.985)	(0.000)	(1.025)	(0.000)	(1.024)	(0.000)
non clinical (h^o)	6.43	6.84	6.32%	7.17	11.59%	7.18	11.66%
	(0.83)	(0.730)	(0.005)	(0.816)	(0.016)	(0.812)	(0.016)
Total Services[a]	151.37	142.65	-5.76%	135.39	-10.56%	136.38	-9.90%
	(4.69)	(5.485)	(0.000)	(7.118)	(0.001)	(7.364)	(0.001)
Non-billable (S^{NB})	69.17	62.33	-9.88%	56.64	-18.11%	57.50	-16.88%
	(3.17)	(3.342)	(0.001)	(4.473)	(0.003)	(4.650)	(0.004)
Billable (S^{B})	82.21	80.32	-2.29%	78.75	-4.20%	78.89	-4.04%
	(3.36)	(3.438)	(0.000)	(3.638)	(0.000)	(3.672)	(0.000)
Service intensity $\left(=\frac{S^{NB}+S^{B}}{h^c * W}\right)$	75.23	71.35	-5.15%	68.09	-9.49%	68.56	-8.87%
	(1.61)	(2.060)	(0.000)	(3.116)	(0.002)	(3.263)	(0.002)
Annual income[a] (X)	133.44	148.03	10.94%	160.18	20.04%	158.16	18.53%
	(4.13)	(4.891)	(0.000)	(5.533)	(0.001)	(5.474)	(0.001)

[a] Thousands of (1996) Can. Dollars.
Note. Average practice behaviour predicted by the model in 2002 depending on whether MR physicians are paid according to: a mandatory Fee-for-Service (*first column*) ; the Mixed Remuneration scheme chosen conditionally on group agreement (*second column*) ; an MR system freely chosen on an individual basis (*third column*) ; or a mandatory MR (*last column*). The percentage variation provided for each compensation scheme takes FFS as a benchmark. Bootstrapped standard errors appear in parenthesis.

subgroup, using the approach suggested by Train (2009, p.258-264).[42] The FFS-specific preferences are computed in a similar way, accounting for constraints on choice. Choosing FFS can result either from a strict preference for this compensation a scheme, or a constrained choice due to the unanimity rule required to switch to the MR system (with a probability ψ). Since ψ has been estimated to be 46%, this possibility is important and should not be ignored. Based on compensation-specific preferences (reported in the last two columns of the results table 3), we perform simulations of the various reforms for each subgroup separately. The comparison of the resulting changes in practice patterns with those observed in Table 6 provides an evaluation of the importance of selection effects on unobservables.

Tables 7 and 8 provide results of simulations using MR- and FFS-specific preferences, respectively. In our analysis, we focus on selection effects in the observed Group Free reform. As expected, one observes that in average MR-specific preferences lead to less services supplied (142.65) and lower service intensity (71) than FFS-specific preferences (145.7 and 74.04, respectively). More-

[42] Appendix 7 of the paper provides details about the algorithm used to simulate the impact of the reforms on the subgroup of MR physicians.

Table 8: Treatment effects: FFS-specific preferences

	FFS	Group Free MR		Individual Free MR		Mandatory MR	
		Practice	Variation	Practice	Variation	Practice	Variation
Weekly Total Hours	45.01	45.18	0.38%	45.33	0.70%	45.35	0.76%
	(1.11)	(1.034)	(0.000)	(0.990)	(0.000)	(0.994)	(0.000)
____ clinical (h^c)	39.58	39.38	-0.51%	39.22	-0.93%	39.24	-0.87%
	(0.95)	(0.966)	(0.000)	(1.039)	(0.000)	(1.045)	(0.000)
____ non clinical (h^o)	5.43	5.80	6.86%	6.11	12.57%	6.12	12.67%
	(0.64)	(0.596)	(0.005)	(0.696)	(0.016)	(0.689)	(0.016)
Total Services[a]	153.70	145.79	-5.15%	139.20	-9.44%	140.30	-8.72%
	(5.56)	(5.221)	(0.000)	(5.666)	(0.001)	(5.934)	(0.001)
____ Non-billable (S^{NB})	67.41	61.26	-9.12%	56.14	-16.72%	57.10	-15.29%
	(3.24)	(3.036)	(0.001)	(3.651)	(0.002)	(3.826)	(0.002)
____ Billable (S^{B})	86.29	84.52	-2.05%	83.05	-3.75%	83.20	-3.58%
	(4.44)	(4.275)	(0.000)	(4.179)	(0.000)	(4.195)	(0.000)
Service intensity $\left(=\frac{S^{NB}+S^{B}}{h^c * W}\right)$	77.66	74.04	-4.67%	70.99	-8.59%	71.51	-7.91%
	(1.93)	(1.877)	(0.000)	(2.427)	(0.001)	(2.590)	(0.001)
Annual income[a] (X)	135.68	148.46	9.42%	159.10	17.26%	156.81	15.58%
	(4.94)	(4.604)	(0.000)	(4.689)	(0.001)	(4.611)	(0.001)

[a] Thousands of (1996) Can. Dollars.
Note. Average practice behaviour predicted by the model in 2002 depending on whether FFS physicians are paid according to: a mandatory Fee-for-Service (*first column*) ; the Mixed Remuneration scheme chosen conditionally on group agreement (*second column*) ; an MR system freely chosen on an individual basis (*third column*) ; or a mandatory MR (*last column*). The percentage variation provided for each compensation scheme takes FFS as a benchmark. Bootstrapped standard errors appear in parenthesis.

over, the impact (in %) of the reform on the volume of services performed based on MR-specific preferences is negative and a little stronger in absolute value (-5.76%) than based on FFS-specific preferences (-5.15%).

Given the small differences between figures for MR- and FFS-specific preferences, this suggests that the problem of selection on unobservables does not seem to be very important in our data, when one controls for exogenous observable variables (gender and age). This is consistent with our simulation results corresponding to the mandatory MR reform: the constraints on MR choice imposed by the observed Group Free reform seem to be much more important than selection in explaining the considerable impact of the mandatory MR reform on physicians' behaviour. Our results are also consistent with the descriptive statistics shown in Table 2. While our simulations do not seem to indicate important selection on unobservables, that on observables is likely to be substantial.

Table 9: Practice under a cost-preserving wage under mandatory MR

	FFS	Constant cost	% Var.	Variable cost	% Var.
Weekly Total Hours	45.77	46.35	1.27%	46.08	0.68%
_____ clinical (h^c)	39.92	39.63	-0.75%	39.53	-0.99%
_____ non clinical (h^o)	5.85	6.73	15.01%	6.55	12.04%
Total Services[a]	153.75	146.83	-4.50%	139.86	-9.03%
_____ Non-billable (S^{NB})	67.97	62.68	-7.77%	57.26	-15.75%
_____ Billable (S^B)	85.78	84.15	-1.90%	82.60	-3.71%
Service intensity $\left(=\frac{S^{NB}+S^B}{h^c * W}\right)$	77.02	74.11	-3.78%	70.76	-8.13%
Annual income[a] (X)	135.68	135.68	0.00%	158.21	16.61%
Per Diem (3.5 hours)	–	187.96$		300$	

[a] Thousands of (1996) Can. Dollars.
Note. Average practice behaviour predicted by specification 3 of the model (accounting for observed and partially unobserved heterogeneity) in 2002 depending on whether physicians are paid according to: a mandatory Fee-for-Service (*first column*), the Mandatory Mixed Remuneration scheme, associated to a *per diem* that maintain health care costs at a constant level (*second* and *third* columns), and the Mandatory Mixed Remuneration scheme, associated to the actual *per diem* (*fourth* and *fifth* columns).The *third* and *fifth* columns provide the percentage variation in practice induced by the change.

6.3.5 The Effects of (Constant-Cost) Contracts

One striking feature that is highlighted by the simulations in Sections 6.3.2 and 6.3.3 is the cost of the MR contract; the large *per diem* paid to physicians caused incomes to increase by over 9% in all versions of the reform investigated in Table 6. It is therefore of interest to investigate whether alternative contracts could achieve similar results at lower costs. To do so we concentrate on constant-cost contracts, that is, contracts that keep annual payments to physicians equal to those observed pre-reform (under FFS). We restrict attention to a mandatory reform, forcing all physicians to work under MR. This allows us to evaluate the impact of switching from a FFS to a mandatory MR scheme on physician practice, at constant aggregate costs.

To investigate physician behaviour under constant-cost contracts, we fix the fee-for-service paid under MR at the levels observed in the actual MR contract, but allow the *per diem* to be determined endogenously at a level that holds expected costs constant.[43] The results are given in Table 9 (we replicate the simulation results of the (variable-cost) mandatory MR from the two last columns of Table 6 for ease of comparison). The *per diem* paid to physicians in this case would be $53.70 per hour, or $187.96 per 3.5 hour period (compared to $300 in the observed contract), a reduction of 37%. By construction, physicians' annual income growth would be zero relative to FFS under a constant-cost reform as compared to 16.61% under the variable cost scheme. Moreover, some

[43]For a given *per diem*, we calculate the implied probabilities of different practice alternatives. This implies an expected cost (income) which we compare to the cost under FFS. The numerical procedure iterates over the *per diem* until convergence is achieved.

physician behaviours would change substantially as compared to the variable-cost mandatory MR system. In particular, the volume of services provided would not decrease by as much as under the observed reform. Total services would decrease by 4.50% relative to FFS rather than by 9.03% under the variable-cost reform. This reflects the presence of a smaller income effect. As under the variable-cost reform, the level of clinical hours of work would not be much influenced by the constant-cost system, given it small (compensated and uncompensated) elasticities. The service intensity would therefore decrease (by 3.78%), but much less than under the variable-cost reform (by 8.13%). These results reinforce the importance of incentives in determining the supply of services. In contrast, hours worked are relatively insensitive to incentives.

7 Conclusion

We have developed and estimated a structural labour supply model that incorporates service intensity into the standard consumption/leisure tradeoff. This generates endogenous prices since service intensity affects the opportunity cost of leisure and hours worked affect the marginal return to service intensity. Allowing for choice among alternative contracts adds further non linearities as rational individuals locate on the efficient budget constraint. We have applied our model to analyse the response of physicians to changes in their compensation system, identifying parameters from the differing incentives between fee-for-service contracts and mixed-remuneration contracts as observed in the Province of Quebec. Discretizing the choice set of physicians allows us to take account of non linearities in an empirically tractable manner.

We have used our estimates to simulate the effects of alternative policies and compensation systems, both on physician behaviour and costs. Our results suggest that incentives significantly affect physicians' service intensity and the volume of services provided. The observed MR reform led to a 5.32% reduction in the volume of services provided and to a 4.78% decrease in the service intensity. The effect on weekly hours was much less pronounced: hours spent at work increased by 0.34%. A counterfactual mandatory MR reform would have a substantially larger effect on behaviour: services would decrease by 9.03% and service intensity would decrease by 8.13%. The cost per physician would increase by over 16.61%, largely due to the large *per diem* offered to physicians, \$300 per 3.5 hours. A constant-cost (mandatory) reform, setting the *per diem* to \$187.96 dollars per 3.5 hours would generate substantially smaller effects on physician behaviour: services would decrease by 4.50%, and service intensity would decrease by 3.78%. Also, when controlling for gender and age, our analysis provides little evidence about selection on unobservables in the choice of the compensation system.

Our results have implications for the empirical application of labour supply models and data gathering. They demonstrate the importance of extending traditional models to incorporate changes on service intensity, at least in the health-care sector. The physicians in our sample adjust their behaviour much more in terms of the volume of services and service intensity than in terms of time

spent at work. Ignoring such changes would vastly misrepresent the effects of policies on the supply of health services. Future work will benefit from additional data sets that incorporate information on both labour supply and service intensity. Extending data sets to include information on health outcomes would also be helpful. Matched physician-patient data sets, allowing researchers to follow patients through time would allow researchers to compare health outcomes based on the payment system of physicians permitting further advances in measuring the quality of health care.

Our paper also raises some modelling issues for physician labour supply and measuring treatment effects. In developing our model we have assumed that physicians exercise complete control over their practice environment, choosing both the number of services to supply and hours to work, given exogenously determined prices. This makes sense within the context of publicly provided health-care systems. Yet in market based systems the number of services provided and their prices are subject to market forces. Extending the model to account for demand-side factors would allow applications in market-oriented health care systems. We also ignore general-equilibrium effects in our model. General-equilibrium effects would occur if, for example, there is a transfer of activities between physicians who chose MR and those who remained on FFS. Economists have begun to extend structural models to account for general-equilibrium effects in policy evaluation (see, for example, Lise, Seitz, and Smith, 2004). We leave this for future work.

References

ANDREASSEN, L., M. DI TOMMASO, AND S. STROM (2013): "Do medical doctors respond to economic incentives?," *Journal of Health Economics*, 32(2), 392–409.

ARROW, K. J. (1963): "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 53(5), 941–973.

BALTAGI, B. H., E. BRATBERG, AND T. H. HOLMAS (2005): "A panel data study of physicians' labor supply: the case of Norway," *Health Economics*, 14(10), 1035–1045.

BECKER, G. S., AND H. G. LEWIS (1973): "On the Interaction between the Quantity and Quality of Children," *Journal of Political Economy*, 81(2), S279–S288.

BLUNDELL, R., A. DUNCAN, AND C. MEGHIR (1998): "Estimating Labor Supply Responses Using Tax Reforms," *Econometrica*, 66(4), 827–861.

CLEMENS, J., AND J. D. GOTTLIEB (2014): "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?," *American Economic Review*, 104(4), 1320–49.

DEVLIN, R.-A., AND S. SARMA (2008): "Do Physician Remuneration Schemes Matter? The Case of Canadian Family Physicians," *Journal of Health Economics*, 25(7), 1168–1181.

DUMONT, E., B. FORTIN, N. JACQUEMET, AND B. SHEARER (2008): "Physicians' multitasking and incentives: Empirical evidence from a natural experiment," *Journal of Health Economics*, 27(6), 1436–1450.

EVANS, R. (1974): "Modeling the economic objectives of the physician," in *Health economics symposium, Proceedings of the First Canadian Conference 4-6 Sept.*, ed. by R. Fraser, pp. 33–46. Queen's University Industrial Relations Centre, Kingston (Ont.).

FELDSTEIN, M. S. (1970): "The Rising Price of Physician's Services," *Review of Economic and Statistics*, 52(2), 121–133.

FERRALL, C., A. W. GREGORY, AND W. G. THOLL (1998): "Endogenous Work Hours and Practice Patterns of Canadian Physicians," *The Canadian Journal of Economics*, 31(1), 1–27.

FORTIN, B., N. JACQUEMET, AND B. SHEARER (2008): "Policy Analysis in the health-services market: accounting for quality and quantity," *Annals of Economics and Statistics*, 91-92, 293–319.

——— (2010): "Labour Supply, Work Effort and Contract Choice: Theory and Evidence on Physicians," IZA Discussion Paper No. 5188.

GOURIEROUX, C., AND A. MONFORT (1993): "Simulation-based inference : A survey with special reference to panel data models," *Journal of Econometrics*, 59(1-2), 5–33.

HAUSMAN, J. A. (1985): "The Econometrics of Nonlinear Budget Sets," *Econometrica*, 53(6), 1255–1282.

HOYNES, H. W. (1996): "Welfare Transfers in Two-Parent Families: Labor Supply and Welfare Participation Under the AFDC-UP Program," *Econometrica*, 64(2), 295–332.

KALB, G., D. KUEHNLE, A. SCOTT, T. C. CHENG, AND S.-H. JEON (2018): "What factors affect physicians' labour supply: Comparing structural discrete choice and reduced-form approaches," *Health Economics*, 27(2), e101–e119.

LISE, J., S. SEITZ, AND J. SMITH (2004): "Equilibrium Policy Experiments and the Evaluation of Social Programs," *NBER WP*, (10283).

MACURDY, T., D. GREEN, AND H. PAARSCH (1990): "Assessing Empirical Approaches for Analyzing Taxes and Labor Supply," *Journal of Human Resources*, 25(3), 415–490.

MCFADDEN, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, pp. 105–142. New York Academic Press, New York (NJ).

MCFADDEN, D., AND K. TRAIN (2000): "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15(5), 447–470.

MCGUIRE, T. G. (2000): "Physician Agency," in *Handbook of Health Economics*, ed. by A. J. Culyer, and J. P. Newhouse, vol. 1A, pp. 461–536. North-Holland, Amsterdam.

NOETHER, M. (1986): "The Growing Supply of Physicians: Has the Market Become More Competitive?," *Journal of Labor Economics*, 4(4), 503–537.

ROTHENBERG, T. J. (1971): "Identification in Parametric Models," *Econometrica*, 39(3), 577–591.

SAETHER, E. (2005): "Physicians' Labour Supply: The Wage Impact on Hours and Practice Combinations," *Labour*, 19(4), 673–703.

SHEARER, B., N. H. SOMÉ, AND B. FORTIN (2017): "Physicians' Response to Incentives: Evidence on Hours of Work and Multitasking," mimeo.

SHOWALTER, M. H., AND N. K. THURSTON (1997): "Taxes and labor supply of high-income physicians," *Journal of Public Economics*, 66(1), 73–97.

SLOAN, F. A. (1975): "Physician Supply Behavior in the Short Run," *Industrial and Labor Relations Review*, 28(4), 549–569.

TRAIN, K. E. (1999): "Halton Sequences for Mixed Logit," *University of Berkeley, Working Paper*.

——— (2009): *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edn.

VAN SOEST, A. (1995): "Structural Models of Family Labor Supply: A Discrete Choice Approach," *Journal of Human Resources*, 30(1), 63–88.

ZABALZA, A., C. PISSARIDES, AND M. BARTON (1980): "Social security and the choice between full-time work, part-time work and retirement," *Journal of Public Economics*, 14(2), 245–276.

A Appendices

A.1 Indexes

Quantities: Let p_a^t stand for the price of the service a at time t and $S_{a,i}^t$ for the number of a-type services a physician i provided at time t. The annual level of services S_i^t is then measured as:

$$\begin{cases} S_i^t = \sum_a S_{a,i}^t \; p_a^{1996} & \text{if } 1996 \leq t < 2000, \\ S_i^t = \sum_a (S_{a,i}^t \; p_a^{2000}) \; \dfrac{\sum_a S_{a,i}^{2000} \; p_a^{1996}}{\sum_a S_{a,i}^{2000} \; p_a^{2000}} & \text{if } 2000 \leq t \leq 2002. \end{cases} \tag{23}$$

The same price are used for weighting billable and non-billable services. The variable S_i^t in (23) then stands for either non-billable services, $S_i^t = S_i^{\text{NB}\,t}$, or billable ones, $S_i^t = S_i^{\text{B}\,t}$, aggregated using the same price levels.

Prices: For the same reasons, the weights used for price indexes are the average level of services provided by FFS physicians. This avoids incorporating into price measures the effect of the variations in services due to switching to MR. Let $\overline{S}_a^t$ denote the average level of billable services of type a provided by all the FFS physicians belonging to the specialty considered. The price index of services is then given by:

$$\begin{cases} p^t = \dfrac{\sum_a \overline{S}_a^{1996} \; p_a^t}{\sum_a \overline{S}_a^{1996} \; p_a^{1996}} & \text{if } 1996 \leq t < 2000, \\ p^t = \dfrac{\sum_a \overline{S}_a^{2000} \; p_a^t}{\sum_a \overline{S}_a^{2000} \; p_a^{2000}} \dfrac{\sum_a \overline{S}_a^{1996} \; p_a^{2000}}{\sum_a \overline{S}_a^{2000} \; p_a^{2000}} & \text{if } 2000 \leq t \leq 2002. \end{cases} \tag{24}$$

Once again, we hold constant the weights used for measuring the price index under MR, PF^t, since it is calculated using the average billable services provided by FFS physicians, at MR reduced prices.

A.2 Non-Convex Budget Set

Let the budget set be given by

$$XM = \left\{ (X, s, h) \in R^3 : X - psh - wh \leq y, s \geq 0, X \geq 0 \right\}.$$

Let (X^o, s^o, h^o) and (X^1, s^1, h^1) be on the frontier of XM with $h_1 > h_o, s_1 > s_o$ and

$$X^o - ps^o h^o - wh^o = X^1 - ps^1 h^1 - wh^1 = y. \tag{25}$$

For $\lambda \in (0, 1)$, define

$$(X^3, s^3, h^3) = \lambda (X^1, s^1, h^1) + (1 - \lambda)(X^o, s^o, h^o). \tag{26}$$

Convexity requires

$$X^3 - ps^3 h^3 - wh^3 \leq y \tag{27}$$

or,

$$\lambda X^1 + (1-\lambda)X^o - p\left[\lambda s^1 + (1-\lambda)s^o\right]\left[\lambda h^1 + (1-\lambda)h^o\right] - w\left[\lambda h^1 + (1-\lambda)h^o\right] \leq y. \tag{28}$$

But

$$y = \lambda X^1 + (1-\lambda)X^o - \left[\lambda p s^1 h^1 + (1-\lambda)p s^o h^o\right] - \left[\lambda w h^1 + (1-\lambda)w h^o\right]. \tag{29}$$

So (28) can be written

$$\begin{aligned}
-p\left[\lambda s^1 + (1-\lambda)s^o\right]\left[\lambda h^1 + (1-\lambda)h^o\right] &\leq -p\left[\lambda s^1 h^1 + (1-\lambda)s^o h^o\right] \\
\left[\lambda s^1 + (1-\lambda)s^o\right]\left[\lambda h^1 + (1-\lambda)h^o\right] &\geq \lambda s^1 h^1 + (1-\lambda)s^o h^o \\
\lambda(\lambda-1)s^1 h^1 + \lambda(\lambda-1)s^o h^o &\geq \lambda(\lambda-1)s^1 h^o + \lambda(\lambda-1)s^o h^1 \\
s^1 h^1 + s^o h^o &\leq s^1 h^o + s^o h^1 \\
\left(s^1 - s^o\right)\left(h^1 - h^o\right) &\leq 0.
\end{aligned} \tag{30}$$

But this contradicts $h_1 > h_o, s_1 > s_o$, so the budget set is not convex.

A.3 Equivalence

Let $(X^\star, h^\star, s^\star)$ be the unique optimal vector that maximizes $U(X,h,s)$ subject to $(X,h,s) \in XM$ where

$$XM = \left\{(X,s,h) \in R^3 : X - pS - wh \leq y, s \geq 0, X \geq 0, S = sh\right\}.$$

Then $(X^\star, h^\star, s^\star)$ satisfies

$$\begin{aligned}
U(X^\star,h^\star,s^\star) > U(X',h',s') \quad & \forall (X',h',s') \in XM, (X',h',s') \neq (X^\star,h^\star,s^\star) \iff \\
U(X^\star,h^\star,S^\star/h^\star) > U(X',h',S'/h') \quad & \forall (X',h',s') \in XM, (X',h',s') \neq (X^\star,h^\star,s^\star) \iff \\
u(X^\star,h^\star,S^\star) > u(X',h',S') \quad & \forall (X',h',s') \in XM, (X',h',s') \neq (X^\star,h^\star,s^\star).
\end{aligned} \tag{31}$$

A.4 Comparative Statics

To perform the comparative statics, we use the transformed utility function $u = u(X,h,S)$. We assume an interior solution and the quasi-concavity of $u(X,h,S)$ over the relevant region. Let the expenditure function $m(w,p,\bar{u})$ be the solution to the standard dual program $\min_{\{X,h,S\}} X - wh - pS$ subject to $\bar{u} - u(X,h,S) \leq 0$. In our case, the expenditure function yields the minimum amount of non-labour income needed to get $\bar{u}$ for given w and p. Then, from Shephard's Lemma,

$$\begin{aligned}
\frac{\partial m(w,p,\bar{u})}{\partial w} &= -\tilde{h}(w,p,\bar{u}), \\
\frac{\partial m(w,p,\bar{u})}{\partial p} &= -\tilde{S}(w,p,\bar{u}).
\end{aligned} \tag{32}$$

Also, from the concavity of the expenditure function,

$$\begin{aligned}
\frac{\partial m(w,p,\bar{u})^2}{\partial w^2} &= -\frac{\partial \tilde{h}(w,p,\bar{u})}{\partial \hat{w}} \leq 0, \\
\frac{\partial m(w,p,\bar{u})^2}{\partial p^2} &= -\frac{\partial \tilde{S}(w,p,\bar{u})}{\partial \hat{p}} \leq 0,
\end{aligned} \tag{33}$$

which demonstrates the inequalities (6).

Moreover, since the concavity of the expenditure function imposes no restrictions on the signs of the cross derivatives in wage and price, one has

$$\frac{\partial m(w,p,\bar{u})^2}{\partial w \partial p} = \quad -\frac{\partial \tilde{h}(w,p,\bar{u})}{\partial p} \underset{<}{>} 0, \tag{34}$$

which demonstrates (7). Finally,

$$\tilde{s}(w,p,\bar{u}) = \quad \frac{\tilde{S}(w,p,\bar{u})}{\tilde{h}(w,p,\bar{u})}, \tag{35}$$

from which

$$\tilde{\eta}_{s,p} = \tilde{\eta}_{S,p} - \tilde{\eta}_{h,p}, \tag{36}$$

where $\tilde{\eta}_{i,p}$ is the compensated elasticity of i with respect to p, for $i = s, S, h$, which demonstrates (8).

A.5 Econometric Methodology

A.5.1 The Basic Empirical Model

Let V_{ij} stand for the annual utility of physician i in alternative j that is decomposed into a deterministic component, u_j, and a random term which is independent across alternatives ϵ_{ij}. Thus,

$$V_{ij} = u_j + \epsilon_{ij}, \text{ where } \epsilon_{ij} \sim \text{i.i.d. Gumbel (extreme value type I).}$$

The deterministic component of the quadratic utility function to be maximized is given by [44]

$$u_j = \boldsymbol{\gamma}' \mathbf{z}_j + \mathbf{z}_j' \boldsymbol{\beta} \mathbf{z}_j + \gamma_{S^{\text{NB}}} S_j^{\text{NB}} + \boldsymbol{b}_{S^{\text{NB}}}' \mathbf{z}_j S_j^{\text{NB}} + \beta_{S^{\text{NB}}} (S_j^{\text{NB}})^2, \tag{37}$$

where the practice characteristics that are fully observable are denoted by

$$\boldsymbol{z}_j = (h_j^o,\ T - h_j^o - h_o^c,\ S_j^{\text{B}},\ X_j)',$$

and those for which we observe a lower bound to the actual number performed, S_j^{NB},[45] and where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_o & \beta_o^l & \beta_o^{S^{\text{B}}} & \beta_o^y \\ \beta_o^l & \beta_l & \beta_l^{S^{\text{B}}} & \beta_l^y \\ \beta_o^{S^{\text{B}}} & \beta_l^{S^{\text{B}}} & \beta_{S^{\text{B}}} & \beta_{S^{\text{B}}}^y \\ \beta_o^y & \beta_l^y & \beta_{S^{\text{B}}}^y & \beta_y \end{pmatrix}; \boldsymbol{\gamma} = \begin{pmatrix} \gamma_o \\ \gamma_l \\ \gamma_{S^{\text{B}}} \\ \gamma_y \end{pmatrix}; \boldsymbol{b}_{S^{\text{NB}}} = \begin{pmatrix} \beta_o^{S^{\text{NB}}} \\ \beta_l^{S^{\text{NB}}} \\ \beta_{S^{\text{NB}}}^{S^{\text{B}}} \\ \beta_{S^{\text{NB}}}^y \end{pmatrix}.$$

[44] In what follows, the individual index i is neglected, where possible, for clarity.

[45] Recall that MR physicians do not spend all of their time on the *per diem*. When they perform non-billable services off the *per diem*, they are paid for them as a FFS physician would be and we observe the transactions. When they perform non-billable services on the *per diem* they are not paid for them and we do not observe the transactions. The number of non-billable services that are observed for MR physicians is therefore a lower bound to the number of such services actually performed.

To begin, we consider the case for which S^{NB} is fully observable. A physician chooses alternative j if: $V_{ij} \geq V_{ik}, \ \forall k \neq j$. The individual contribution to the likelihood function is the probability of this event occurring, *i.e.*,

$$\mathcal{L}_{ij} = P\left[V_{ij} \geq V_{ik}, \forall k \neq j\right] = P\left[\epsilon_{ij} \geq u_k - u_j + \epsilon_{ik}, \forall k \neq j\right] = \frac{exp(u_j)}{\sum_{k=1}^{J} exp(u_k)}. \tag{38}$$

A.5.2 Calculation of MR earnings

A number of issues arise in calculating gross income under the MR system (see eq. (12)). First, a physician's income depends on the number of *per diems* claimed. As this is unknown, we must approximate it. To do so, we assume that each MR physician works the maximum number of *per diems* possible for a given number of hours worked, the remainder of his time is then allocated to FFS.

We estimate the number of (half) *per diems* worked during a week by

$$\widehat{\mathcal{N}} = \frac{\min\left\{floor\left(\frac{2\times(h^c+h^o)}{\overline{d}}\right), 28\right\}}{2}, \tag{39}$$

where $\overline{d}$ is the number of hours per *per diem* and 28 represents the maximum number of *per diem*s that a physician can claim over a two-week period.

Second, recall that we distinguish between billable services provided under the *per diem*, denoted S^{B}_{FFS}, for which the physician is paid a discounted fee, αp, and those provided outside of the *per diem*, denoted S^{B}_{MR}, for which the physician is paid the regular fee, p. Given that we do not observe whether or not a given service was remunerated under the *per diem*, we use θS^{B} and $(1-\theta)S^{\text{B}}$ to estimate S^{B}_{MR} and S^{B}_{FFS}, respectively. Here θ is the proportion of time spent under the *per diem*, estimated as the share of total hours worked in a week under the *per diem* and given by

$$\hat{\theta} = \frac{\overline{d}\,\widehat{\mathcal{N}}}{h^c + h^o}. \tag{40}$$

Hence we attribute billable services to MR and FFS in the same proportion as we attribute hours worked to MR and FFS.

Consumption in alternative j, in year t, under MR is then given by[46]

$$X^{MR}_{j,t} = 46\widehat{\mathcal{N}}_j\mathcal{D}_t + p_t S^{\text{NB}}_j + \hat{\theta}_j \alpha p_t(S^{\text{B}}_j) + (1-\hat{\theta}_j)p_t S^{\text{B}}_j, \tag{41}$$

[46]Note that the fact that we only observe a lower bound to S^{NB} does not affect our calculations of income. This is because the observed lower bound represents the exact number of non-billable acts performed outside of the *per diem* period where they were remunerated. The unobservable part of S^{NB} is provided within the *per diem* period and does not affect income.

where $\widehat{\mathcal{N}_j}$ is the number of (half) *per diems* worked in alternative j, $\mathcal{D}_t$ is the payment per (half) *per diem* in year t, and $\hat{\theta}_j$ is the estimated share of total hours worked in a week in alternative j attributed to the *per diem*.

We accounted for government imposed income ceilings and regional income differentials as discussed in footnote (15). The actual provisions governing regional remuneration rate calculations involve a wide variety of individual characteristics—such as city of practice – not included in the data set. However, our data contains each physician's quarterly income before and after the correction for the regionally differentiated remuneration rate. We therefore approximate the actual regionally-differentiated remuneration rate facing physician i, and denoted τ_i, as the ratio of the two reported levels of income over the whole sample period.

The actual level of income ceilings during the period is publicly available from government authorities in charge of physician compensation. However, these ceilings depend on the establishment in which the services were provided, information that is not available to us.[47]To take account of these exceptions in a tractable manner we calculate the average percentage of time that pediatricians spent in establishments where income ceilings were applied. The relevant ceiling for physician i, is then taken to be the actual income ceiling adjusted for the average percentage of time spent in establishments where the cap applies.

With these elements in hand, the actual consumption in each alternative is predicted according to equations (20) and (41).[48] To convert consumption into real terms we deflate actual (nominal) consumption in each alternative using the price index provided by *Statistics Canada*. The average inflation rate for the whole period is 1.92%. Overall, our strategy for approximating consumption in each alternative proved to be a precise predictor of the observed income of physicians included in our sample.[49]

[47]For example, emergency services were excluded from the capped income prior to 2001.

[48]For simplicity, we ignore income taxes in our analysis. However, since most physicians in our sample period have a yearly income implying the highest marginal (provincial + federal) tax rate and since there has been no tax reform over our period, the marginal tax rate is likely to be constant for most physicians.

[49]A regression of physicians' observed income on their predicted income yielded a R^2 of 0.83, with a coefficient of 0.97 (standard error = 0.005) and a non significant intercept.

A.6 Algorithm to Simulate Selection Effects

Let

$$d_j^{MR} = \begin{cases} 1 & \text{if } X_j^{MR} > X_j^{FFS}; \\ 0 & \text{else .} \end{cases}$$

To simulate the average $\boldsymbol{\kappa}$ in the sub-group who would select MR, where $\boldsymbol{\kappa}$ is the (unknown) vector of the parameters of the individual utility functions:

1. Draw $\boldsymbol{\kappa}_r$ from $g(\boldsymbol{\kappa})$, its asymptotic distribution.
2. Calculate $Pr(j|\boldsymbol{\kappa}_r)$ for each alternative $j = \{1, 2, \ldots, J\}$
3. Calculate $Pr(MR|\boldsymbol{\kappa}_r) = \sum_{j=1}^{J} Pr(j|\boldsymbol{\kappa}_r) d_j^{MR}$
4. Repeat (1)—(3) R times, saving $\boldsymbol{\kappa}_r$ and $Pr(MR|\boldsymbol{\kappa}_r)$ at each iteration
5. Calculate
$$\omega^r = \frac{Pr(MR|\boldsymbol{\kappa}_r)}{\sum_{r=1}^{R} Pr(MR|\boldsymbol{\kappa}_r)}, \quad r = 1, 2, \ldots, R.$$
6. Calculate $\hat{\boldsymbol{\kappa}}_{MR} = \sum_{r=1}^{R} \omega^r \boldsymbol{\kappa}_r$, the average beta for the sub-group in the population who would select MR.

We can then use $\hat{\boldsymbol{\kappa}}_{MR}$ to do simulations for the sub-group who prefer MR. We can then do the same for FFS physicians to calculate selection effects. It is easy to extend the algorithm to account for constraints on MR physicians choices associated with Group Free MR reform (*i.e.*, the introduction of ψ).