DISCUSSION PAPER SERIES

IZA DP No. 5188

Labour Supply, Work Effort and Contract Choice: Theory and Evidence on Physicians

Bernard Fortin
Nicolas Jacquemet
Bruce Shearer

September 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Labour Supply, Work Effort and Contract Choice: Theory and Evidence on Physicians

Bernard Fortin
Université Laval, CIRPÉE, CIRANO and IZA

Nicolas Jacquemet
Paris School of Economics and University Paris I Panthéon-Sorbonne

Bruce Shearer
Université Laval, CIRPÉE, CIRANO and IZA

Discussion Paper No. 5188
September 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

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

IZA Discussion Paper No. 5188
September 2010

ABSTRACT

Labour Supply, Work Effort and Contract Choice: Theory and Evidence on Physicians[1]

We develop and estimate a generalized labour supply model that incorporates work effort into the standard consumption-leisure trade-off. We allow workers a choice between two contracts: a piece rate contract, wherein he is paid per unit of service provided, and a mixed contract, wherein he receives an hourly wage and a reduced piece rate. This setting gives rise to a non-convex budget set and an efficient budget constraint (the upper envelope of contract-specific budget sets). We apply our model to data collected on specialist physicians working in the Province of Quebec (Canada). Our data set contains information on each physician's labour supply and their work effort (clinical services provided per hour worked). It also covers a period of policy reform under which physicians could choose between two compensation systems: the traditional fee-for-service, under which physicians receive a fee for each service provided, and mixed remuneration, under which physicians receive a *per diem* as well as a reduced fee-for-service. We estimate the model using a discrete choice approach. We use our estimates to simulate elasticities and the effects of *ex ante* reforms on physician contracts. Our results show that physician services and effort are much more sensitive to contractual changes than is their time spent at work. Our results also suggest that a mandatory reform, forcing all physicians to adopt the mixed remuneration system, would have had substantially larger effects on physician behaviour than those observed under the voluntary reform.

JEL Classification: C25, J22, J33, I10, J44

Keywords: labour supply, effort, contracts, practice patterns of physicians, discrete choice econometric models, mixed logit

Corresponding author:

Bernard Fortin
Département d'économique
Université Laval
Local 2182, Pavillon J.-A. DeSève
1025, Avenue des Sciences-Humaines
Québec (Québec) G1V 0A6
Canada
E-mail: Bernard.Fortin@ecn.ulaval.ca

1 Introduction

Empirical labour-supply models have played an important role in economic policy analysis. Simulations from these models have been used to measure and predict the effects of tax, welfare and social policies on labour-force participation and hours worked in the economy. Examples include Heckman (1974); Hausman (1980); Hoynes (1996); Blundell, Duncan, and Meghir (1998); see Blundell and MaCurdy (1999) for a survey. The parallel, but related, empirical contracting literature has focussed on the effect of monetary incentives on effort and productivity (Paarsch and Shearer, 2000; Shearer, 2004), paying attention to the endogenous choice of a compensation system on the part of agents (Lazear, 2000) and heterogeneous responses to incentives (Chiappori and Salanié, 2003). Generalized models, which incorporate effort decisions into traditional labour-supply models, allow for a richer policy evaluation environment, particularly in settings where productivity is not proportional to time spent at work. Such models can be used to measure and predict the effects of wages and contracts on hours worked and productivity, capturing responses that may be missed in traditional models; they also permit cost and welfare comparisons across contracts.[2] Yet, to date, little attempt has been made to combine contracting models with labour-supply models in empirical work.[3] This paper addresses this issue. We develop and estimate a generalized labour-supply model using a unique data set on physicians' practice behaviour.

Physician behaviour is well-suited to analysis within a generalized model. Physicians can affect both their time spent at work and the volume of services that they provide while at work through effort (McGuire, 2000). Recent estimates of time-based labour supply elasticities have been provided by Showalter and Thurston (1997) and Baltagi, Bratberg, and Holmås (2005).[4] Estimates of the effects of compensation policies on physician behaviour have been provided by Gruber and Owings (1996) and Devlin and Sarma (2008). Our approach allows us to measure these effects within the context of a parsimonious economic model and to measure the effects of contracts on the time spent

[1]The authors thank the *Collège des médecins du Québec* for making its survey data available and the *Régie d'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 CIRPÉE Workshop on Applied Micro-Econometrics (Québec), 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 and Marie-Claire Villeval for useful discussions and comments. We acknowledge research support from the Canadian Institute of Health Research (CIHR) and the Canada Research Chair in Social Policies and Human Resources at the Université Laval.

[2]Ferrall and Shearer (1999); Margiotta and Miller (2000); Paarsch and Shearer (2009); Copeland and Monnet (2009) are among those who have implemented empirical contracting models to investigate the welfare properties of different contracts and the cost of moral hazard within the firm.

[3]One important exception is Dickinson (1999) who analyses a model incorporating on- and off-the-job leisure, that is tested using controlled laboratory experiments.

[4]The effect of incentives on physicians' choice of location has also been analysed within the context of the labour supply model (Bolduc, Fortin, and Fournier, 1996)).

per service, a suggested measure of health-care quality (Ma and McGuire, 1997).[5] It also allows us to generalize treatment-effect estimates to predict the effects of compensation policies *ex-ante*[6] and, in some cases, to make welfare comparisons across contracts.

We develop a simple model which we use to motivate and develop our general approach. We specify utility as a function of consumption, hours of work and effort (measured by the volume of services produced per hour of work). 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 effort. Similarly, the marginal return on effort depends on hours of work. These nonlinear prices are similar to those obtained in quantity/quality models (Becker and Lewis, 1973). Some comparative static results are derived; 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 worker has the choice between two contracts: one composed uniquely of a piece rate and another 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 Quebec (Canada) between the years 1996-2002. All these physicians work within the Quebec public Health-Care System. 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. 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 fee-for-service (FFS) public contracts, receiving a fee for each service provided. In 1999, the government introduced a mixed remuneration (MR) scheme, under which physicians received a *per diem*, paid per hour worked, and a reduced fee-for-service. A notable aspect of the reform was its voluntary nature; from the time the mixed compensation system was introduced, two sub-samples of physicians are observed: those who adopt the MR system and those who remain under the FFS system. We exploit this change in the compensation system to identify the physicians' preference parameters.

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

[5]Recent empirical work suggests that compensation policies do influence physician behaviour in these directions (Dumont, Fortin, Jacquemet, and Shearer, 2008).

[6]See Marschak (1953); Heckman and Vytlacil (2001), and Todd and Wolpin (2008) for a discussion of *ex-ante* policy evaluation.

the institutional constraints imposed on these contracts within the Quebec Health-Care System (*e.g.*, income ceilings, regionally differentiated remuneration, constraints on the choice of the compensation system at the individual level). 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). To minimize the effects of functional forms on our results, we use a flexible (quadratic) utility function, which can be viewed as a second-order approximation to the true 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.[7] 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.

Elasticities and the effects of policy reforms are simulated through changing appropriate parameters of the budget constraint and allowing physicians to re-optimize. Our results suggest that labour supply (weekly hours and weeks) elasticities are quite small while the (compensated) elasticities of effort and services with respect to the fee per service are much stronger, being estimated at about 0.3 and 0.4 respectively. Non-clinical hours (spent on administrative and teaching activities), that are not remunerated under a FFS contract, are quite sensitive to compensated changes in the fee per service, with an elasticity of -0.4. Our results also suggest that the changes in incentives brought about by the 1999 reform strongly affected physician behaviour. Services completed decreased by 9% and non-clinical hours increased by 30%. What is more, work effort decreased, suggesting that the quality of care may have increased (more time spent per service). A mandatory reform, forcing all physicians to work under MR, would have reduced services by 15% and increased non-clinical hours by 50%. However, these larger effects are not due to unobserved heterogeneity and selection, but rather to the constraints placed on individual choice in the observed reform.

[7] Members of each department (groups of specialists working in the same field) would vote on the adoption of MR; adoption required unanimous approval.

The reform was also costly. Payments to physicians increase by over 9%. This is due to the large *per diem* that physicians were paid for working under MR. We investigate the welfare effects of constant-cost contracts under voluntary participation in MR. Under such circumstances services provided would decrease relative to the fee-for-service contract by 6.47%. Yet the welfare implications are inconclusive: clinical hours worked would decrease by only 1.72% and time spent per service would increase by 4.83%.

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 fee-for-service and mixed remuneration 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 Québec reform and develops our empirical model. Section 6 describe our empirical results and the policy simulations. Section 7 presents our conclusions.

2 A generalized model of labour supply

We present a static model of labour supply behaviour under linear contracts. Our model allows for decisions over hours of work and effort. 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.

Preferences are represented by the strictly quasi-concave and twice-differentiable utility function

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

where X is consumption, h is hours of work, and e is effort. We assume

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

Effort is determined by the number of services, A, performed per hour of work; we have[8]

$$A = eh. \tag{3}$$

In some settings, time spent per service can be taken as a measure of quality of services – Ma and McGuire (1997) have suggested such a measure in the case of physicians. This amounts to $= 1/e$

[8]One interpretation of (3) is a (Cobb-Douglas) production function. In a more general model this function could be written as $A = A(e, h; \bar{\mathbf{z}})$, with e denoting a measure of effort such as the ratio of *effective* hours of work to hours at work and $\bar{\mathbf{z}}$ denoting an exogenous vector of inputs that affect the marginal productivity of effort and hours worked. This could be decomposed into two sub-matrices: $\overline{\mathbf{z_1}}$ containing physical capital (*e.g.*, hospital equipment) and $\overline{\mathbf{z_2}}$ containing human capital, or personal, characteristics ((*e.g.*, age, experience). Such a specification would also allow for more complex substitution patterns between effort and hours (perhaps due to fatigue).

in our model, changes in which are a valid measure of changes in time spent providing services if on-the-job leisure is fixed.[9]

The budget constraint is given by

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

where X is consumption (the price of which being normalized to one), w is the wage rate, p is the fee per unit of service and y is non-labour income. Variation in w and p are treated as exogenous.[10] Note that the budget constraint, (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 mixed contract, while setting ($w > 0$ and $p = 0$) gives a fixed-wage contract.

We assume complete and symmetric information. This is a direct consequence of the fact that effort is the ratio of two observable variables: hours worked and services completed.[11] What is more, we assume that the worker has complete control over his practice variables – freely choosing both his hours of work and his clinical services.[12]

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 peh, and rearranging. This gives $X = (w + pe)h + (ph)e + y^v$, where $y^v = y - phe$ is the virtual non-labour income. It follows that the marginal return to an hour of work ($w + pe$) depends on the physician's choice of effort – the number of services that can be performed in that hour. Similarly, the marginal return to effort (ph) depends on the physician's hours of work. Since effort changes the number of services performed per hour, the return to effort 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.3). The second-order conditions for optimal behaviour 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 $(X^\star, h^\star, e^\star)$. In Appendix A.4 we show that $(X^\star, h^\star, e^\star)$ is equivalent to $(X', h', A'/h')$ which maximizes the transformed utility:

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

[9]This ensures that workers with higher values of $1/e$ do not simply take longer breaks between services.

[10]This is consistent with the public health-care system in Quebec. See Feldstein (1970) for an analysis that incorporates the price-setting behaviour of physicians in a market-based environment.

[11]Agency problems and moral hazard could be introduced by incorporating random elements into the production function, perhaps due to differences in the difficulty of tasks or in the marginal utility of on-the-job leisure.

[12]Within the context of models of physician behaviour 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. It 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.

subject to

$$X - pA - wh = y,$$

where (5) is obtained by substituting $e = A/h$ directly into the utility function (1): $u(X, h, A) = U(X, h, A/h)$. 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; effort is not defined in (1) when hours are set to 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 effort and hours of work through two channels: the first is the standard income effect, the second is through its impact on the endogenous marginal returns to effort and hours of work. Some results are possible, however. In particular, the fact that the budget constraint is linear in A and h implies that the expenditure function is concave in w and p; hence,

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

where˜indicates that the partial is compensated. Notice however that concavity of the expenditure function does not allow us to sign cross-partial derivatives; hence,

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

Similarly, since effort is the ratio of services to hours worked, its compensated partial derivative with respect to p depends on

$$\frac{\partial \tilde{A}}{\partial p} - \tilde{A}\frac{\partial \tilde{h}}{\partial p} \tag{8}$$

which is also unsigned. These results follow from a straight-forward application of duality theory to the problem of maximizing (5) s.t. (4); we include a derivation in Appendix A.5 for completeness.[13]

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 ($X = pA$) and a mixed compensation (MR) system ($X = wh + \alpha pA$), where $\alpha < 1$ 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,

[13]Similar results are shown in Edlefsen (1981, 1983) using the Hessian matrix from the original maximization problem. Edlefsen (1981) also shows that a compensated increase in the fee per unit of service will increase both effort and hours of work under a FFS system ($p > 0$ and $w = 0$) or a MR system where w is small, if leisure at work (or $-e$) and leisure outside of work (or $-h$) are both net substitutes with respect to consumption. Using a somewhat different model, Dickinson (1999) also finds that the effect of a compensated increase in the piece-rate on hours of work is ambiguous.

we assume that preferences are (directly) independent of the compensation system. This implies that rational workers will always select that compensation system that maximizes income for a given (h, A) 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, A). Assuming for simplicity zero non-labour income, we have

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

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

$$D(h^\star, A^\star; w, p, \alpha) = \arg\max_{D \in \{0,1\}} (1 - D)pA^\star + D(wh^\star + \alpha pA^\star) \tag{10}$$

evaluated at the optimal levels of $A^\star, h^\star$.[14]

This is illustrated in Figure 1 which considers the tradeoff between services and consumption (income), conditional on $h^{\star,FFS}$, optimal hours under the fee-for-service system.[15] 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 A, X chosen under FFS correspond to the optimal values $A^{\star,FFS}$ 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.[16]

The efficient budget constraint associated with the transformed program is given by the bold line.[17] It is piece-wise linear and non convex; this raises well-known problems for optimization and

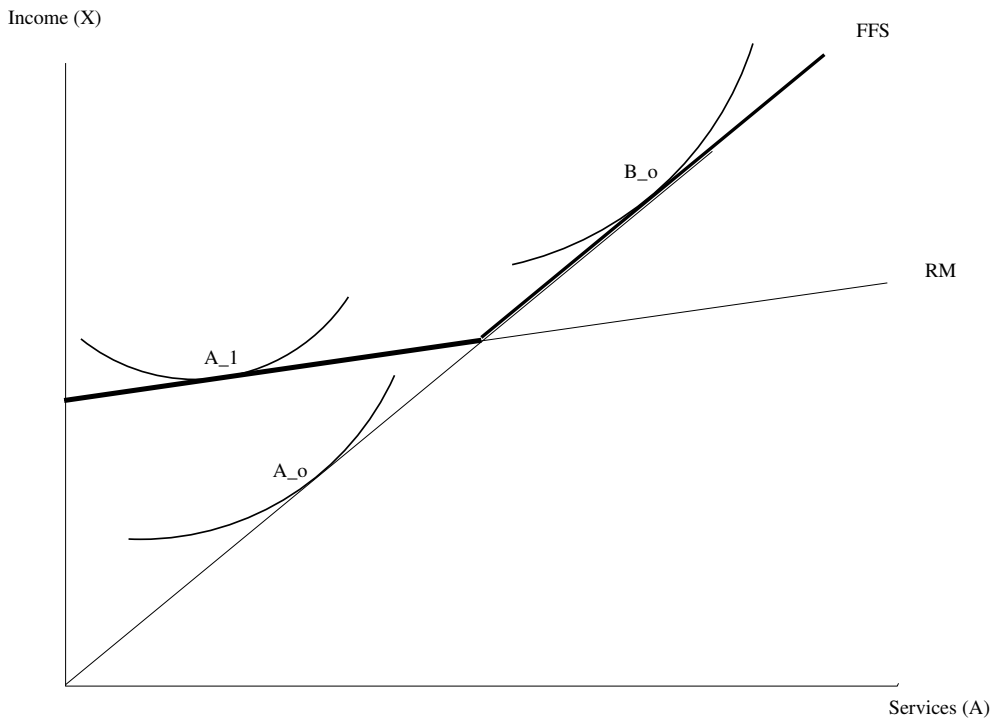
[14]The empirical model that we estimate will be adjusted to account for the institutional details of specialist physician pay in Quebec, including income ceilings and regional payment differentials. This renders the efficient budget constraint more complex, with more segments (which helps identification of the model), yet its derivation follows the same basic steps.

[15]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 ($= w + pe$). The first-order condition for hours is given by $\frac{-U_h}{U_X} = w + pe$.

[16]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 $U_e > 0$ for $e < \bar{e}$, due for instance to concern of the physician for his patients. Another, is to assume that monitoring allows for a minimum level of effort to be enforced; see, for example, Lin (2003). We do not elaborate on these possibilities here in order to keep our model simple and since the fixed wage contract is not observed in our empirical data. In any case, our econometric model does not impose $U_e < 0$.

[17]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 A.

Figure 1: Optimal choices along the efficient budget constraint



labour supply estimation (Hausman, 1985; MaCurdy, Green, and Paarsch, 1990). Our approach will be to discretize the choice set of workers (physicians), considering only a finite set of values for h, A (van Soest, 1995).

Figure 1 also illustrates potential problems of self-selection. Workers who have a preference for low effort (or high service quality) levels (such as worker A) will tend to choose MR, while those who have a preference for high effort levels (such as worker B) will tend to choose FFS.[18] A comparison of behaviour across compensation systems will potentially confound the effects of the compensation system with the differences in preferences. The econometric model must therefore allow for both observable and unobservable heterogeneity to take into account of possible selection bias.

[18] Again, the figure loses some exactness by collapsing a 3-dimensional problem into two dimensions. While an individual who chooses A_o under FFS and prefers A_1 will definitely switch to MR (since utility under MR must be at least as high as A_1), the converse is not necessarily true; an individual who prefers B_o may still switch to MR at $h^{\star,MR}$.

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 Québec, physicians have traditionally been paid according to a fee-for-service compensation system.[19] 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.[20] In the present section, for expositional purposes, we take as given that there is one type of service, denoted A, and one fee, denoted p. The physician's budget constraint is then given by

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

In 1999, the government introduced a Mixed Remuneration (MR) scheme. Under this system, specialist physicians receive a wage (or *per diem*) for time spent at work. (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 (or other recognized health-care establishments). 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 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 A^{BS}, are remunerated at a reduced fee-for-service, αp, $0 < \alpha < 1$; and non-billable services, denoted A^{NBS}, are not remunerated: $\alpha = 0$.[21] As with p, the discount factor α is specialty and service specific.

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 A^{BS}_{FFS}. Those performed under the *per diem* are denoted A^{BS}_{MR}. Non-billable services, performed under *per diem* periods are not paid and hence not recorded.

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

[19] 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.

[20] 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.

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

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	*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 raw*) and services (*second raw*) are remunerated under Fee-for-Service (*left-hand side*) and Mixed Remuneration (*right-hand side*). The last two raws describe some income policies that equally applies to both compensation schemes.

under the MR system is then given by

$$\widehat{X}^{MR} = \mathcal{W}\,\mathcal{N}\,\mathcal{D} + \alpha p\; A_{MR}^{BS} + p\;(A_{FFS}^{BS} + A_{FFS}^{NBS}). \tag{12}$$

3.1 Net Income

The government also imposes income ceilings on physicians, beyond which the fee-for-service is reduced by 75%. As well, there is a regionally differentiated remuneration rate, designed to induce physicians to practice in remote areas of the province. The regional rate depends on the geographic location of the practice, as well as the physician's specialty. Income ceilings apply to the gross differentiated income.

To calculate net income, define τ as the differentiated remuneration rate faced by the physician in the region where he practices ($\tau > 1$ denotes a favoured region). Let C denote the ceiling cap above which services are discounted at a 0.75 rate.[22] The potential income in the absence of any ceiling cap can thus be written as

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

Moreover, the net income, X, is given by:

$$X = min[\widehat{X}, C] + max\left[0.25\,[\widehat{X} - C], 0\right] + \tau\widehat{X}. \tag{14}$$

[22]We only observe whether the physician participated in the MR compensation system during a given year. We do not observe the exact partition of time between FFS and MR compensation system during that year.

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[23] as well as hours spent performing teaching and administrative duties. The survey also includes the number of weeks worked per year for the years 1996, 1997, 1998 and 2002. Due to the exclusion of weeks worked from the survey in 1999-2001, we exclude these observations from the empirical analysis. Since the MR reform occurred in the last quarter of 1999, we end up eliminating the 3 years immediately following the reform. The survey also includes information on the personal characteristics of each physician, including: age, gender, specialization and location.

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. We matched data from these two sources 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 effort. 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.[24] Price variation is excluded from the index by holding price weights constant at the base year levels.[25] These weights are the base-year prices paid to FFS physicians; they are the same for both billable and non-billable services[26] The price data for differ-

[23]Patients can be seen either in hospitals or in private clinics. Physicians working private clinics are paid the public-sector fee schedule; they work under FFS when they see patients in private clinics.

[24]For example, consider the case of a pediatrician who, in a morning, completes 4 primary visits and 6 follow-up visits at the hospital. This would count for 10 services if all services were treated equally. Yet, primary visits require an initial interview and a complete diagnosis and generally last for 45 minutes. Control visits, on the other hand, typically last only 20 minutes. Indeed, primary visits are compensated at 47 $ each, while the price falls to 16.50 $ for control visits.

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

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

Table 2: Summary statistics on sampled physicians

	FFS physicians				MR physicians			
	Before MR Reform		After MR Reform		Before MR Reform		After MR Reform	
	Mean	Standard Error	Mean	Standard Error	Mean	Standard Error	Mean	Standard Error
Observed practice								
Weekly Total Hours	43.09	13.01	41.69	12.71	48.64	12.67	46.52	10.03
_____ clinical	38.69	12.79	38.79	11.33	41.38	13.73	37.96	12.26
_____ non clinical	4.40	8.36	2.90	7.22	7.26	9.62	8.55	11.22
Weeks	46.03	5.37	47.07	2.24	46.29	5.35	46.78	1.93
Total Services[a,b]	167.00	66.83	169.80	73.42	141.81	56.16	115.01	75.42
_____ Non-billable[a]	71.85	47.02	73.39	55.63	60.94	36.20	52.00	47.65
_____ Billable	95.15	55.47	96.40	56.64	80.88	49.21	63.01	46.31
Effort $\left(=\frac{TotalServices}{Clinical\ Hours\ of\ Work}\right)$	106.82	109.81	96.68	42.21	81.70	40.00	64.93	38.73
Annual income[a]	167.84	67.35	195.08	79.02	146.41	56.86	190.28	62.57
Sample characteristics								
Number of physicians	139	–	105	–	111	–	92	–
Number of observations	355	–	105	–	267	–	92	–
Sex (Male=1)	0.66	0.47	0.68	0.47	0.52	0.50	0.55	0.50
Age	49.89	11.17	53.57	10.70	43.07	10.04	48.30	10.11

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

[b] Lower bound for MR physicians after the reform. See below, Section 5.3.

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 (2002) the reform. The bottom part of the Table summarizes individual characteristics.

ent services was also aggregated into indexes for billable and non-billable services, under both 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 multi-dimensional integrals. In order to limit computational time we restricted the 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) and in the marginal incentives to perform services (the average discount factor, α is equal to 30%).[27] Focusing on one speciality also reduces the problem of heterogeneity in the

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

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 (2002) and on the basis of physicians who remain under FFS or switch to MR. 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.

The top part of the table provide information on professional practice of physicians in our sample, as disaggregated into the four categories considered. We focus on weekly hours of work, both in clinical medicine (providing services to patients) and other activities (administration and teaching), annual weeks of work, clinical services provided (both billable and non-billable) in thousands of (1996) Can. Dollars, effort (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 demographic characteristics of each of the categories.

To summarize, while the practice patterns of MR and FFS physicians in terms of weeks of work are very similar, there is some difference 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 suggests the presence of a potential selection bias problem related to the decision to switch to MR, the non clinical hours being compensated under MR but not under FFS. There is a substantial difference in terms of clinical services provided; MR physicians provided 15% fewer total services before the reform. This also highlights the presence of a potentially important selection problem in the analysis of the impact of the MR system on practice behaviour; physicians who eventually switched to MR were, on average, low "productivity" physicians. The difference in services leads to a substantial difference in annual income, pre-reform; MR physicians earned approximately 13% less income. Results show that while before reform, 66% of FFS physicians were male, 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, 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.

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 Québec 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 A^{BS}, and non-billable, denoted A^{NBS}. 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.[28] 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.[29] 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 weekly total hours by h^t (with $h^t = h^c + h^o$). Pure leisure is denoted by l.

Physicians' preferences are represented by an annual utility function,

$$U = U(X, h^o, L, l, A^{BS}, A^{NBS}) \tag{15}$$

defined over:

X (Annual income),
h^o (Weekly hours of administrative work and teaching)
L (Weeks of leisure during the year),
l (Weekly hours of leisure outside of work),
A^{BS} (Number of billable services supplied throughout the year),
A^{NBS} (Number of non-billable services supplied throughout the year).

The usual time constraints imply that:

$$L = 52 - \mathcal{W}$$
$$l = T - h^c - h^o,$$

where $T = 24 \times 7 = 168$, the maximum amount of time available in a week. We allow leisure during workweeks, l, and leisure during non-work weeks, L, to be imperfect substitutes.[30] We also allow for differences in the marginal utility (or disutility) of billable and non-billable services.[31] The efficient budget constraint is obtained from the compensation system that maximizes net income, X, for each for each practice vector, $(h^o, L, l, A^{BS}, A^{NBS})$; this is given by equations (11) to (14).[32]

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

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

[30]Imperfect substitution between these two types of leisure is supported by empirical evidence (Hanoch, 1980; Blank, 1988).

[31]This allows for the possibility that different types of services may be associated with different levels of difficulty and require different effort 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.

[32]To compute the efficient budget constraint, one must also make assumptions about how a MR physician allocates his

Table 3: Sample distribution regarding discretized practice variables

Weekly Total Hours (h^t)				Weeks (W)		Total Services (A)			
clinical (h^c)		non clinical (h^o)				Non-billable (A^{NBS})		Billable (A^{BS})	
5	6.23%	0	45.30%	30	3.42%	0	9.77%	0	6.84%
30	33.94%	4	40.29%	50	96.58%	30000	27.11%	20000	12.82%
45	52.01%	20	10.62%	.	.%	60000	27.35%	50000	23.20%
70	7.81%	40	3.79%	.	.%	90000	26.74%	100000	42.86%
.	.%	.	.%	.	.%	180000	9.04%	190000	14.29%

Note. For each practice variable considered in the analysis, the left-hand side column provides the discretized levels used in the estimation, the right-hand side column reports the distribution of the sample across this set.

5.1 Discrete Alternatives

Given the non linearities in the efficient budget constraint after the MR reform, we follow recent tradition in the empirical labour supply literature (van Soest, 1995; Saether, 2005) 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 levels of clinical hours of work, N_o levels of non-clinical hours of work, N_w levels of weeks of work, N_{BS} levels of billable services and N_{NBS} levels of non-billable services. Thus the complete choice set of practice variables involves $dim(J) = N_c \times N_o \times N_w \times N_{NBS} \times N_{BS}$ alternatives. A single alternative, corresponding to one particular practice possibility, is a set of values: $j = \{c_j, o_j, w_j, NBS_j, BS_j\}$ 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.

An important step in implementing the empirical specification is the determination of the partition of the choice variables, defining each alternative. The identification of preference parameters which replicate the data suggests that this partition should replicate the actual distribution of choices observed in the data. Yet this has to be traded off with computational costs; each additional point along any given demand function induces an exponential increase in the dimension of the choice set. These concerns led us to an asymmetric partition of the choice variables. Variables which display greater heterogeneity in actual choices are more finely partitioned than those with lesser heterogeneity. The sample distribution around the chosen partition is given in Table 3; it gives $dim(J) = 800$ alternatives.

hours of work and billable services in and outside of the *per diem* periods. These issues and other details concerning the calculation of physicians' income are discussed in section 5.3.2

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 5.3.1 below.

Following Keane and Moffitt (1998), we specify utility as a quadratic function, which constitutes a second order approximation of any well-behaved utility function. We differentiate between practice characteristics that are fully observable, denoted

$$Z_j = \left[(h_j^o), (52 - \mathcal{W}_j), (T - h_j^o - h_o^c), (A_j^{BS}), (X_j)\right]',$$

and those for which we observe a lower bound to the actual number performed, A_j^{NBS}.[33]

To begin, we consider the case for which A^{NBS} is fully observable. The deterministic component of utility is given by[34]

$$u_j = \gamma' Z_j + Z_j' \beta Z_j + \gamma_{NBS} A_j^{NBS} + B_{NBS}' Z_j A_j^{NBS} + \beta_{NBS} (A_j^{NBS})^2, \tag{16}$$

where

$$\beta = \begin{pmatrix} \beta_o & \beta_o^L & \beta_o^l & \beta_o^{BS} & \beta_o^y \\ \beta_o^L & \beta_L & \beta_L^l & \beta_L^{BS} & \beta_L^y \\ \beta_o^l & \beta_L^l & \beta_l & \beta_l^{BS} & \beta_l^y \\ \beta_o^{BS} & \beta_L^{BS} & \beta_l^{BS} & \beta_{BS} & \beta_{BS}^y \\ \beta_o^y & \beta_L^y & \beta_l^y & \beta_{BS}^y & \beta_y \end{pmatrix}; \gamma = \begin{pmatrix} \gamma_o \\ \gamma_L \\ \gamma_l \\ \gamma_{BS} \\ \gamma_y \end{pmatrix}; B_{NBS} = \begin{pmatrix} \beta_o^{NBS} \\ \beta_L^{NBS} \\ \beta_l^{NBS} \\ \beta_{NBS}^{BS} \\ \beta_{NBS}^y \end{pmatrix}.$$

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.*,

$$\mathcal{L}_{ij} = P\left[V_{ij} \geq V_{ik}, \ \forall k \neq j\right] = P\left[\epsilon_{ij} \geq u_k - u_j + \epsilon_{ik}, \ \forall k \neq j\right] = \frac{e^{u_j}}{\sum_{k=1}^{J} e^{u_k}}. \tag{17}$$

[33] 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.

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

5.3 Estimation issues

Several features of our data set necessitate slight modifications to the 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 whereas 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 $dim(J^C) = 640$. Note that this strategy leads us to use the same alternatives for estimation independent of the alternative that was chosen. This uniform conditioning property (McFadden, 1978) has been shown to ensure consistent estimation.

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 A_m^{NBS} denote the level of non-billable services that is observed for a given physician (*i.e.*, delivered outside the *per diem* period). Since, for this observation, A_m^{NBS} is a lower bound to the actual number of non-billable services provided, we observe A_m^{NBS} whenever $A^{NBS} \in \left\{A_m^{NBS}, A_{m+1}^{NBS}, A_{m+2}^{NBS}, \ldots, A_{N_{NBS}}^{NBS}\right\}$. 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 Z_j, A_m^{NBS} is obtained by summing over $A_j^{NBS} \quad j \geq m$; *i.e.*,[35]

$$P\left(Z_j, A_m^{NBS}\right) = \frac{\exp(\gamma' Z_j + Z_j' \beta Z_j)}{\sum_{k=1}^{J^C} e^{u_k}} \sum_{l=m}^{N_{NBS}} \exp(\gamma_{NBS} A_l^{NBS} + B_A'.Z_j A_l^{NBS} + \beta_{NBS}(A_l^{NBS})^2). \tag{18}$$

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

$$\mathcal{L}_{ij} = \left(\frac{e^{u_j}}{\sum_{k=1}^{J^C} e^{u_k}}\right)^{1-D_i} \left(P\left(Z_j, A_m^{NBS}\right)\right)^{D_i}, \tag{19}$$

35

$$P\left(Z_j, A_m^{NBS}\right) = P\left(Z_j, A_m^{NBS}\right) \cup P\left(Z_j, A_{m+1}^{NBS}\right) \cup \ldots \cup P\left(Z_j, A_{N_{NBS}}^{NBS}\right) = \sum_{l=m}^{N_{NBS}} \exp[u(Z_j, A_l^{NBS})] / \sum_{k=1}^{J^C} e^{u_k},$$

and hence

$$P\left(Z_j, A_m^{NBS}\right) = \sum_{l=m}^{N_{NBS}} \exp\left(\gamma' Z_j + Z_j' \beta Z_j + \gamma_{NBS} A_j^{NBS} + B_{NBS}' Z_j A_j^{NBS} + \beta_{NBS}(A_j^{NBS})^2\right) / \sum_{k=1}^{J^C} e^{u_k},$$

the manipulation of which gives the result.

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

5.3.1 Heterogeneity in Preferences

We account for observable heterogeneity in the model, allowing the estimated coefficients to be functions of individual characteristics. In particular, we allow the linear coefficient terms, γ, and the quadratic coefficient terms, β, to be linear functions of age and gender; we write

$$\begin{aligned} \gamma_i^k &= \gamma_0^k + \gamma_1^k \times Age_i + \gamma_2^k \times DMale_i \quad k = \{o, l, L, BS, NBS, X\}, \\ \beta_i^k &= \beta_0^k + \beta_1^k \times Age_i + \beta_2^k \times DMale_i \quad k = \{o, l, L, BS, NBS, X\}, \end{aligned} \tag{20}$$

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

We account for unobservable heterogeneity by adding normally distributed random terms to the functions in (20) (with the exception of γ_i^{NBS}).[36] Define

$$\widetilde{\gamma}_i = (\widetilde{\gamma}_i^o, \widetilde{\gamma}_i^L, \widetilde{\gamma}_i^l, \widetilde{\gamma}_i^{BS}, \widetilde{\gamma}_i^X)$$

to be the vector of random coefficients, where

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

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

$$l_{ij}(\widetilde{\gamma}_i, \beta_i) = \left(\frac{e^{u_{ij}}}{\sum_{k=1}^{J^C} e^{u_{ik}}} \right)^{1-D_i} \left(P_{ij}\left(Z_j, A_m^{NBS}\right) \right)^{D_i}, \tag{21}$$

where the utility index now depends on i to incorporate both observed and unobserved heterogeneity. The unconditional probabilities correspond to the mixed logit specification:

$$\mathcal{L}_{ij} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} l_{ij}(\widetilde{\gamma}_i, \beta_i) \phi(\eta_o) \phi(\eta_L) \phi(\eta_l) \phi(\eta_{BS}) \phi(\eta_X) d\eta_o d\eta_L d\eta_l d\eta_{BS} d\eta_X, \tag{22}$$

where ϕ denotes the normal density function.

The estimation of this model requires the computation of a large number of five-dimensional integrals. To calculate these integrals we rely on simulation methods, approximating (22) by the average value of $l_{ij}(\widetilde{\gamma}_i, \beta_i)$ over r random draws in the distribution of each η_k.[37] The simulated Maximum Likelihood estimator derived from this specification is asymptotically equivalent to an exact ML estimator given that $\sqrt{r}$ rises faster than the size of the sample (Gourieroux and Monfort, 1993).

[36] The model did not converge when unobserved heterogeneity was added to γ_i^{NBS}, possibly do to the fact that we only observe a lower bound to the number of non-billable services.

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

5.3.2 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 counter-factual 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 (14) to predict the income for each alternative.

Recall from our discussion in Section 3, we aggregated services provided into two types: billable, denoted A^{BS}, and non-billable, denoted A^{NBS}. 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

$$X_{j,t}^{FFS} = p_t(A_j^{BS} + A_j^{NBS}). \tag{23}$$

Calculating consumption under MR 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. A number of issues arise in calculating (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 *per diems* worked during a week by

$$\widehat{\mathcal{N}} = \frac{\min\left\{floor\left(\frac{2\times(h^c+h^o)}{\overline{d}}\right), 28\right\}}{2}, \tag{24}$$

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 A_{FFS}^{BS}, for which the physician is paid a discounted fee, αp, and those provided outside of the *per diem*, denoted A_{MR}^{BS}, 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 θA^{BS} and $(1-\theta)A^{BS}$ to estimate A_{MR}^{BS} and A_{FFS}^{BS}, 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

$$\hat{\theta} = \frac{\overline{d}\,\widehat{\mathcal{N}}}{h^c + h^o}. \tag{25}$$

For each random parameter, we perform 20 Halton draws specific to each individual. For a given draw, the likelihood is evaluated by computing the utility derived from each of the J^C alternatives. We then calculate the simulated probability of selecting each alternative, conditional on the draw. The likelihood of selecting each alternative is then the average of these simulated probabilities. Each iteration of the likelihood requires computing utility $N \times J^C \times 20$ times. We estimated the model using parallel programming on a cluster of 20 processors.

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[38]

$$X_{j,t}^{MR} = \mathcal{W}_j \widehat{\mathcal{N}_j} \mathcal{D}_t + p_t A_j^{NBS} + \hat{\theta}_j \alpha p_t (A_j^{BS}) + (1 - \hat{\theta}_j) p_t A_j^{BS}, \tag{26}$$

where $\mathcal{W}_j$ is the number of weeks worked in alternative j, $\widehat{\mathcal{N}_j}$ 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 per (14). 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.[39] 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 (23) and (26).[40] 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.[41]

[38] Note that the fact that we only observe a lower bound to A^{NBS} 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 A^{NBS} is provided within the *per diem* period and does not affect income.

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

[40] 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.

[41] 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.

5.3.3 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.[42] 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.

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

$\mathcal{R}_1$ 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.[43]

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

$$P^{\mathcal{B}}\left(Z_j, A_m^{NBS}\right) \tag{27}$$

denote the probability of observing a given alternative $\left(Z_j, A_m^{NBS}\right)$ for an MR physician, from (18).

[42] 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.

[43] 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.

The individual contribution to the likelihood function is given by

$$
\begin{aligned}
l_i(\widetilde{\gamma}_i, \beta_i) = & \left[\frac{e^{u_{ij}^{FFS}}}{\sum_{k \in J} e^{u_{ik}^{FFS}}} \right]^{D_{i1}} \\
& \times \left[(1-\psi)\, P^{EFF}\left(Z_j, A_m^{NBS} \right) \right]^{D_{i2}} \\
& \times \left[\psi\, \frac{e^{u_{ij}^{FFS}}}{\sum_{k \in J} e^{u_{ik}^{FFS}}} \right]^{D_{i3}} \\
& \times \left[\psi\, \frac{e^{u_{ij}^{FFS}}}{\sum_{k \in J} e^{u_{ik}^{FFS}}} + (1-\psi)\, \frac{e^{u_{ij}^{Eff}}}{\sum_{k \in J} e^{u_{ik}^{Eff}}} \right]^{(1-D_{1i}-D_{2i}-D_{3i})} .
\end{aligned}
\tag{28}
$$

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.

6 Results

6.1 Parameter Estimates

We estimated three versions of the quadratic utility function (16): first, in the absence of heterogeneity; second, with observed heterogeneity; and third, with observed and unobserved heterogeneity. Each case incorporates constrained choice of the compensation system – the contribution to the likelihood of observation i, conditional on γ_o, is given by (28).[44] The results are presented in Table 4. The first column presents results without heterogeneity. The second column presents results when observed heterogeneity is introduced into the linear and quadratic terms for non-clinical hours worked (h_o), weeks of leisure (L), hours of leisure per day (l), non-billable services (NBS), billable services (BS) and income (X). These coefficients are permitted to vary with age and gender. Finally, the third 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.

[44] We also estimated the model without taking account of constrained choice. The results for these specifications were generally less satisfactory than those presented.

Table 4: Mixed Logit Parameters: Quadratic Utility

	Multinomial Logit		Observed heterogeneity		Full specification	
	Parameter	(Stud.)	Parameter	(Stud.)	Parameter	(Stud.)
γ^o	2.7e-01***	(3.58)	5.2e-01***	(5.48)	5.3e-01***	(5.11)
σ^o	–	–	–	–	5.7e-02*	(1.66)
$\gamma^o \times Male$	–	–	1.8e-01***	(4.55)	1.5e-01***	(3.12)
$\gamma^o \times Age$	–	–	-6.7e-03***	(4.49)	-6.9e-03***	(4.29)
γ^L	4.5e-09	(0.00)	2.1e-14	(0.00)	1.8e-08	(0.39)
$\gamma^L \times Male$	–	–	4.5e-14	(0.00)	-7.5e-09	(0.17)
$\gamma^L \times Age$	–	–	4.6e-14	(0.00)	-3.6e-10	(0.01)
γ^l	4.8e-01***	(6.30)	2.2e-01	(1.16)	2.7e-01*	(1.31)
σ^l	–	–	–	–	2.1e-02*	(1.51)
$\gamma^l \times Male$	.	–	-9.7e-02	(0.83)	-8.9e-02	(0.81)
$\gamma^l \times Age$	–	–	7.0e-03*	(1.51)	7.1e-03*	(1.55)
γ^{NBS}	-2.9e-02	(1.25)	-1.7e-02	(0.60)	-8.0e-03	(0.25)
$\gamma^{NBS} \times Male$	–	–	1.3e-02*	(1.43)	1.4e-02	(1.27)
$\gamma^{NBS} \times Age$	–	–	-4.0e-05	(0.10)	-1.1e-04	(0.24)
γ^{BS}	1.1e-02	(0.58)	7.6e-03	(0.31)	8.9e-02**	(2.04)
σ^{BS}	–	–	–	–	8.3e-02***	(7.04)
$\gamma^{BS} \times Male$	–	–	1.9e-03	(0.18)	-4.1e-03	(0.17)
$\gamma^{BS} \times Age$	–	–	4.6e-04	(1.07)	5.7e-04	(0.75)
γ^x	8.4e-02***	(4.10)	9.5e-02***	(3.24)	9.7e-02***	(2.36)
σ^x	–	–	–	–	3.2e-02***	(6.17)
$\gamma^x \times Male$	–	–	-8.1e-04	(0.06)	-8.2e-03	(0.45)
$\gamma^x \times Age$	–	–	-4.4e-04	(0.87)	4.2e-04	(0.56)
β^o_L	3.1e-04	(0.14)	9.6e-04	(0.44)	9.5e-04	(0.45)
β^o_l	-2.2e-03***	(4.26)	-2.4e-03***	(4.65)	-2.4e-03***	(4.28)
β^o_{NBS}	-1.3e-04	(0.70)	-5.2e-05	(0.26)	-2.4e-05	(0.11)
β^o_{BS}	-9.2e-04***	(3.78)	-8.2e-04***	(3.65)	-6.6e-04***	(2.83)
β^o_x	2.1e-04	(1.02)	9.8e-06	(0.04)	1.2e-05	(0.05)
β^L_l	-5.8e-04	(0.38)	-6.2e-04	(0.28)	-4.4e-04	(0.35)
β^L_{NBS}	-2.4e-03***	(2.62)	-1.8e-03**	(2.22)	-1.9e-03**	(2.29)
β^L_{BS}	-2.6e-03***	(3.36)	-2.0e-03***	(2.77)	-2.3e-03***	(2.92)
β^L_x	2.2e-03***	(3.36)	1.7e-03***	(2.58)	1.8e-03***	(2.71)
β^l_{NBS}	1.7e-04	(1.05)	8.5e-05	(0.60)	7.4e-05	(0.53)
β^l_{BS}	-1.0e-05	(0.07)	-1.1e-04	(0.82)	-1.1e-04	(0.84)
β^l_x	-4.0e-04***	(2.61)	-3.3e-04***	(2.73)	-3.5e-04***	(2.91)
β^{NBS}_{BS}	-3.1e-05	(0.37)	2.3e-05	(0.28)	1.6e-04**	(1.73)
β^{NBS}_x	1.4e-04*	(1.41)	3.5e-05	(0.32)	2.5e-05	(0.20)
β^{BS}_x	2.1e-04**	(1.86)	1.0e-04	(0.83)	-4.3e-05	(0.29)

Table 4: Mixed Logit Parameters: Quadratic Utility (Continued)

	Multinomial Logit		Observed heterogeneity		Full specification	
	Parameter	(Stud.)	Parameter	(Stud.)	Parameter	(Stud.)
β^o	-4.3e-04	(0.71)	-7.3e-03***	(4.05)	-9.3e-03***	(4.02)
$\beta^o \times Male$	–	–	-4.2e-03***	(3.31)	-3.8e-03***	(2.72)
$\beta^o \times Age$	–	–	2.0e-04***	(4.72)	2.1e-04***	(4.51)
β^L	1.1e-03	(0.13)	5.3e-03	(0.39)	4.3e-03	(0.52)
$\beta^L \times Male$	–	–	-1.8e-03**	(1.81)	-1.8e-03**	(1.84)
$\beta^L \times Age$	–	–	-8.6e-05*	(1.63)	-7.6e-05*	(1.56)
β^l	-1.7e-03***	(6.43)	-7.1e-04	(0.91)	-8.9e-04	(1.08)
$\beta^l \times Male$	–	–	4.4e-04	(0.90)	4.0e-04	(0.86)
$\beta^l \times Age$	–	–	-2.8e-05*	(1.41)	-2.8e-05*	(1.47)
β^{NBS}	-7.1e-05*	(1.59)	-6.6e-05	(0.65)	-2.7e-04**	(2.20)
$\beta^{NBS} \times Male$	–	–	-6.2e-06	(0.12)	-3.5e-05	(0.61)
$\beta^{NBS} \times Age$	–	–	5.4e-07	(0.30)	3.3e-06*	(1.57)
β^{BS}	-1.8e-04***	(3.41)	-9.3e-05	(0.89)	-3.3e-04	(1.12)
$\beta^{BS} \times Male$	–	–	5.5e-05	(1.13)	1.4e-04	(0.98)
$\beta^{BS} \times Age$	–	–	-1.9e-06	(1.07)	-6.9e-06	(1.06)
β^x	-2.1e-04***	(2.88)	-1.9e-04**	(1.91)	-1.5e-04	(1.16)
$\beta^x \times Male$	–	–	2.9e-06	(0.07)	7.2e-05	(1.26)
$\beta^x \times Age$	–	–	1.2e-06	(0.77)	-2.7e-06	(1.10)
ψ	4.6e-01***	(6.41)	4.6e-01***	(6.42)	4.6e-01***	(6.28)
Log-Likelihood	-4057.0		-3969.9		-3562.6	

Legend. Significance levels: *** 10%, ** 5%, *** 1%.
Note. ML estimation of the model on the sample of pediatricians (N=819 observations). The left-hand side provides point estimates and standard errors of the Multinomial Logit model that includes only parameters from the quadratic function defined over practice variables. Observable heterogeneity is added in the model displayed in the middle of the table. The right-side includes unobserved heterogeneity through random coefficients, assumed normally distributed.

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). This requirement is satisfied for 86% of observations in the model with observed heterogeneity. It is satisfied for 85% of the observations in the multinomial-logit model (with no heterogeneity) and for 60% of the observations in the model with both observed and unobserved heterogeneity. In the interests of selecting a model that best fits the data, while respecting theoretical restrictions, we concentrate on the version of the model with observed heterogeneity. Note, we choose this specification in spite of the fact that the likelihood function increased substantially (from -3969 to -3562) upon the introduction of unobserved heterogeneity. This reflects the tradeoff between fitting the sample data

Table 5: Model Fit

	Observed Total	Predicted 2002	Observed 2002
Weekly Total Hours (h^t)	44.62	45.45	44.18
____clinical (h^c)	39.37	39.41	38.86
____non clinical (h^o)	5.25	6.04	5.32
Weeks ($\mathcal{W}$)	49.32	49.51	50.00
Total Services$(A)^a$	149.04	141.02	140.71
___Non-billable (A^{NBS})	64.87	60.24	61.98
____Billable (A^{BS})	84.16	80.77	78.73
Effort $\left(e = \frac{A^{NBS}+A^{BS}}{h^c * \mathcal{W}}\right)$	76.76	72.27	72.42
Annual income^a (X)	144.08	150.45	149.58

[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.

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. This ensures that predictions are based on economic, rather than a statistical, criteria. 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 this issue in section 6.2.3.

Table 5 compares the predictions with observed choices for the year 2002 (the only post-reform year in our sample). The middle column of the table gives the average predicted values 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 weeks worked, average number of billable services, average effort (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.[45]

[45]Caution should be exercised in interpreting the statistics over non-billable services. Since 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.

Table 6: Elasticity of practice variables

	Bench-mark	Income	Hourly wage rate			Fee per unit of service		
		$\varepsilon_{k/y}$	$\varepsilon_{k/w}$	$\widetilde{\varepsilon}_{k/w}$	$\frac{wh^t\mathcal{W}}{y}\varepsilon_{k/y}$	$\varepsilon_{k/p}$	$\widetilde{\varepsilon}_{k/p}$	$\frac{pA}{y}\varepsilon_{k/y}$
Weekly Total Hours (h^t)	45.04	0.004	0.018	0.010	0.008	0.088	0.039	0.049
——— clinical (h^c)	39.37	-0.002	0.000	0.005	-0.004	0.080	0.106	-0.026
——— non clinical (h^o)	5.67	0.041	0.143	0.052	0.092	0.142	-0.429	0.571
Weeks ($\mathcal{W}$)	49.28	-0.005	-0.006	0.005	-0.010	-0.023	0.041	-0.064
Total Services$(A)^a$	138.67	-0.037	-0.082	-0.001	-0.081	-0.092	0.415	-0.507
—— Non-billable (A^{NBS})	58.85	-0.042	-0.097	-0.003	-0.094	-0.280	0.306	-0.586
——— Billable (A^{BS})	79.83	-0.032	-0.071	0.001	-0.072	0.046	0.495	-0.449
Effort $\left(e = \frac{A^{NBS}+A^{BS}}{h^c * \mathcal{W}}\right)$	71.47	-0.030	-0.077	-0.010	-0.067	-0.154	0.263	-0.418

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

6.2 Policy simulations

Estimation of a structural model allows us to simulate the impact of different compensation policies on physician labour supply 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.

6.2.1 Elasticities through policy simulations

Table 6 provides results on elasticities of practice variables with respect to non-labour income, hourly wage rate, and fee per service. 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-labor income equal to \$10,000.[46] 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 important changes in fee per service unit 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

[46] 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.

Slutsky decompositions of uncompensated elasticities into compensated and total income elasticities: $\varepsilon_{k/w} = \widetilde{\varepsilon}_{k/w} + wh^t \frac{W}{y}\varepsilon_{k/y}$ and $\varepsilon_{k/p} = \widetilde{\varepsilon}_{k/p} + \frac{pA}{y}\varepsilon_{k/y}$, to compute the wage and fee per service compensated elasticities of each practice variable.[47]

Results from the second panel of Table 6 indicate that, as expected, physicians' average clinical weekly hours of work, weeks of work, and the volume of (billable and non-billable) services are negatively affected by an increase in non-labour income. However, non-clinical hours of work increase with non-labour income. 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 -0.005 for weeks of work and -0.037 for services. Moreover, physicians' effort, 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 -0.030.

Results from the third panel indicate that uncompensated own wage elasticities of total weekly hours and weeks of work are 0.018 and -0.006, respectively. This suggests that physicians' labour supply curves for weekly hours and weeks are upward sloping and backward bending respectively, but with a modest response of these variables to a change in the wage rate. Moreover, the compensated own wage elasticities are positive, although quite small, being estimated at 0.010 and 0.005 respectively. The cross compensated wage elasticity on effort is negative but also small (=-0.010), indicating that an increase in the wage rate induces slightly a physician to provide less services per unit of time.

The last panel provides results regarding fee per service elasticities. The own uncompensated elasticity on services is negative and close to -0.1. 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 on non-billable services (=-0.280) than on billable services (= -0.046). The compensated own elasticity of services is positive as expected [see eq. (6)], and quite large (= 0.415). Furthermore, a compensated increase in the fee per service positively affects the physician's effort, with an elasticity of 0.263. These results indicate that physicians respond much more to incentives in terms of services and work effort than in terms of hours and weeks. Notice finally that the compensated elasticity of non-clinical hours with respect to fee per service is negative and quite high in absolute value (=0.429). Our results indicate that a compensated increase in the fee per service induces the physician to spend less time in teaching and administrative activities and more time to perform clinical services.

In short, our results on elasticities suggest that physicians (pediatricians) react to incentives in the directions predicted by the theory. However, the own compensated and uncompensated elasticities of hours and weeks are very small. This result is similar to those reported in studies focusing

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

Table 7: Treatment effects of MR

	FFS	Group Free MR		Individual Free MR		Mandatory MR	
		Practice	Variation	Practice	Variation	Practice	Variation
Weekly Total Hours (h^t)	43.72	45.45	3.94 %	46.68	6.77 %	46.65	6.68 %
____clinical (h^c)	39.09	39.41	0.82 %	39.66	1.45 %	39.67	1.48 %
_____non clinical (h^o)	4.63	6.04	30.32 %	7.03	51.67 %	6.98	50.61 %
Weeks ($\mathcal{W}$)	49.62	49.51	-0.21 %	49.44	-0.36 %	49.46	-0.31 %
Total Services$(A)^a$	155.43	141.02	-9.27 %	130.14	-16.27 %	131.31	-15.52 %
___Non-billable (A^{NBS})	67.58	60.24	-10.85 %	54.64	-19.15 %	55.35	-18.09 %
_____Billable (A^{BS})	87.86	80.77	-8.06 %	75.50	-14.06 %	75.96	-13.54 %
Effort $\left(e = \frac{A^{NBS}+A^{BS}}{h^c * \mathcal{W}}\right)$	80.14	72.27	-9.82 %	66.38	-17.17 %	66.93	-16.49 %
Annual income^a (X)	137.18	150.45	9.68 %	160.90	17.29 %	159.98	16.63 %

[a] Thousands of (1996) Can. Dollars.
Note. Average practice behavior (in terms of practice variables) 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.

on physicians who are not self-employed: for example, Sloan (1975); Noether (1986); Saether (2005) found that the wage elasticities are modest or insignifiant in this context. On the other hand, the (compensated) elasticity of services and effort are positive and much larger. This suggests that physicians adjust their practice behaviour much more in terms of their volume of services and their work effort than in terms of labour supply. Another important result is that non-clinical hours seem to be quite sensitive to a compensated change in the fee per service.

6.2.2 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 7) 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 7, labeled "Group Free MR". The results are instructive in many ways. First, notice the reform increased the number of weekly hours worked, by nearly 4 %. Yet this is almost entirely due to increases in non-clinical hours; clinical hours increased only very slightly, less than 1%. This suggests that the *per diem* incorporated into the MR payment system did induce physicians to spend more 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 9%. This reflects physicians responding to (large) monetary incentives.

The MR compensation system reduced the marginal payment for services received by physicians 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 6.2.4. Our results show that effort decreased with the reform suggests that physicians spent more time with their patients under MR (recall effort is defined as clinical hours per service completed). This may have implications for the quality of health care provided. Finally, the number of weeks worked is insensitive to the mode of payment. This is consistent with empirical results obtained elsewhere; see, for example, Sloan (1975).

6.2.3 Mandatory Reforms and Selection

Given the voluntary nature of the observed reform, a natural question to address is how a mandatory 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 7. They suggest that a mandatory reform would have had considerable effects on services provided (a decrease of 15.5% relative to FFS) and non-clinical hours (an increase of 50.6% 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 16.5% relative to FFS. The cost of the program would also be significantly affected (an increase of 16.6% relative to FFS).

The mandatory reform changes two things vis-à-vis the observed reform: first, it removes choice (and hence selection) and second it removes constraints on an individual's choice of a MR compensation system. To decompose the overall effect into its component parts (and isolate the importance of selection) 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 7, labelled "Individual Free MR." They are very close to the results from the mandatory reform. This suggests that constraints on choice are the important factor in explaining the difference between the actual and mandatory reforms. Even though workers who switched to MR were low-productive workers, many high-productive workers – 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. Selection (on observables) is not responsible for a large part of the observed effect of the reform.[48] Physicians who are currently observed under FFS

[48]Similar results were obtained when we allowed for unobservable heterogeneity, suggesting that selection on unob-

Table 8: Practice under a cost-preserving wage under free MR

	FFS	MR contract			
		Constant cost		Actual	
		Practice	Variation	Practice	Variation
Weekly Total Hours	43.72	45.04	3.01%	46.68	6.77 %
clinical (h^c)	39.09	38.39	-1.79%	39.66	1.45 %
non clinical (h^o)	4.63	6.65	43.55%	6.98	50.61 %
Weeks ($\mathcal{W}$)	49.62	49.65	0.07%	49.44	-0.36 %
Total Services$(A)^a$	155.43	145.37	-6.47%	130.14	-16.27 %
Non-billable (A^{NBS})	67.58	62.90	-6.93%	54.64	-19.15 %
Billable (A^{BS})	87.86	82.47	-6.13%	75.50	-14.06 %
Effort $\left(e = \frac{A^{NBS}+A^{BS}}{h^c * \mathcal{W}}\right)$	80.14	76.27	-4.83%	66.38	-17.17 %
Annual income^a (X)	137.18	137.18	0.00%	160.90	17.29 %
Per Diem (3.5 hours)	–	$123.55		$300.00	

[a] Thousands of (1996) Can. Dollars.
Note. Average practice behavior (in terms of practice variables) predicted by model II (accounting for observed heterogeneity) in 2002 depending on whether physicians are paid according to: a mandatory Fee-for-Service (*first column*) ; the Mixed Remuneration scheme chosen freely at the individual level, associated to a *per diem* that maintain health care costs at a constant level. The *last column* provides the percentage variation in practice induced by the change.

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.

6.2.4 The Welfare Effects of (Constant-Cost) Contracts

One striking feature that is highlighted by the simulations in Sections 6.2.2 and 6.2.3 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 7. 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, contracts that keep annual payments to physicians equal to those observed pre-reform (under FFS). We restrict attention to reforms under which physicians freely choose to adopt MR. This allows some measure of welfare comparison across contracts since physicians are necessarily (weakly) better off under the new contract and costs are constant. The actual realized benefits will then depend on how physician behaviour changes under the new contract as this affects patient welfare.

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 deter-

servables does not play a substantial role either.

mined endogenously at a level that holds expected costs constant.[49] The results are given in Table 8 (we replicate the results of the individual-free reform under the observed contract from Table 7 for ease of comparison). The effects on physician behaviour would have been less pronounced under a constant-cost contract. The *per diem* paid to physicians in this case would be $35.29 per hour, or $123.55 per 3.5 hour period (compared to $300 in the observed contract), a reduction of 59%. Yet, physician behaviour would also change substantially. The volume of services provided would not decrease by as much as under the observed contract. Total services would decrease by 6.5% relative to FFS (rather than by 16% under the observed contract). The time spend at work would not increase by as much as under the observed contract. Weekly hours would increase by 3% relative to FFS (rather than an increase of 6.7% under the observed contract). The time spent per service would therefore decrease (by 13 %) relative to the observed contract. This raises concerns over the quality of services. If we interpret the time spent per service as a measure of quality, then the welfare benefits of the constant-cost contract relative to the observed contract are not clear – services increase but the quality of services goes down.

7 Conclusion

We have developed and estimated a structural labour supply model that incorporates work effort into the standard consumption/leisure trade-off. This generates endogenous prices since effort affects the opportunity cost of leisure and hours worked affect the marginal return to effort. 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 strongly affect physicians' work effort and the volume of services provided. The voluntary MR reform led to a 9% reduction in the volume of services provided. The effect on hours and weeks worked was less pronounced: hours spent at work increased by 4% and the change in weeks worked was negligible. A mandatory reform would have a substantially larger effect on behaviour: hours worked would increase by 6.7%, services would decrease by 15% and time spent per service would increase by 16%. The cost per physician would increase by over 16%, 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 $123.55

[49]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.

dollars per 3.5 hours would generate substantially smaller effects on physician behaviour: services would decrease by 6.5%, hours worked would increase by 3% and time spent per service would increase by 4.8%.

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 work effort, 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 effort 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 work effort. Extending data sets to include information on health outcomes would also be helpful. We have concentrated on the time spent per service as a measure of the quality of health care. 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 only recently 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

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. HOLMÅS (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.

BLANK, R. M. (1988): "Simultaneously Modeling the Supply of Weeks and Hours of Work among Female Household Heads," *Journal of Labor Economics*, 6(2), 177–204.

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

BLUNDELL, R., AND T. MACURDY (1999): "Labor Supply: a Review of Alternative Approaches," in *Handbook of Labor Economics, Vol. 3A*, ed. by O. Ashenfelter, and D. Card, pp. 1559–1696. Elsevier.

BOLDUC, D., B. FORTIN, AND M.-A. FOURNIER (1996): "The Effect of Incentive Policies on the Practice Location of Doctors: A Multinomial Probit Analysis," *Journal of Labor Economics*, 14(4), 703–732.

CHIAPPORI, P.-A., AND B. SALANIÉ (2003): "Testing Contract Theory: A Survey of Some Recent Work," in *Advances in Economics and Econometrics, Eight World Congress*, ed. by M. Dewatripont, L. Hansen, and S. Turnovsky, pp. 115–149. Cambridge University Press, Cambridge (MA).

COPELAND, A., AND C. MONNET (2009): "The Welfare Effects of Incentive Schemes," *The Review of Economic Studies*, 76(1), 93–113.

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.

DICKINSON, D. L. (1999): "An Experimental Examination of Labor Supply and Work Intensities," *Journal of Labor Economics*, 17(4), 638–670.

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.

EDLEFSEN, L. E. (1981): "The Comparative Statics of Hedonic Price Functions and Other Nonlinear Constraints," *Econometrica*, 49(6), 1501–1520.

——— (1983): "The signs of compensated price effects in quantity/quality models," *Economics Letters*, 12(1), 1–6.

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 Statistics*, 52(2), 121–133.

FERRALL, C., AND B. SHEARER (1999): "Incentives and Transactions Costs Within the Firm: Estimating an Agency Model using Payroll Records," *Review of Economic Studies*, 66, 309–338.

FORTIN, B., N. JACQUEMET, AND B. SHEARER (2008): "Policy Analysis in the health-services market: accounting for quality and quantity," *Annales d'Economie et de Statistique*, 91-92, 293–319.

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.

GRUBER, J., AND M. OWINGS (1996): "Physician Financial Incentives and Cesarean Section Delivery," *Rand Journal of Economics*, 27(1), 99–123.

HANOCH, G. (1980): "Hours and Weeks in a Theory of Labor Supply," in *Female Labor Supply : Theory and Estimation*, ed. by J. P. Smith, pp. 119–165. Princeton University Press, Princeton (NJ).

HAUSMAN, J. A. (1980): "The effect of wages, taxes, and fixed costs on women's labor force participation," *Journal of Public Economics*, 14(2), 161–194.

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

HECKMAN, J. J. (1974): "Effects of Child-Care Programs on Women's Work Effort," *Journal of Political Economy*, 82(2), S136–S163.

HECKMAN, J. J., AND E. VYTLACIL (2001): "Policy-Relevant Treatment Effects," *American Economic Review*, 91(2), 107–111.

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

KEANE, M., AND R. MOFFITT (1998): "A Structural Model of Multiple Welfare Program Participation and Labor Supply," *International Economic Review*, 39(3), 553–589.

LAZEAR, E. P. (2000): "The Power of Incentives," *American Economic Review*, 90(2), 410–414.

LIN, C.-C. (2003): "A Backward Bending Labor Supply Curve without an Income Effect," *Oxford Economic Papers*, 55, 336–343.

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

MA, C.-T. A., AND T. G. MCGUIRE (1997): "Optimal Health Insurance and Provider Payment," *American Economic Review*, 87(4), 685–704.

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.

MARGIOTTA, M., AND R. MILLER (2000): "Managerial Compensation and the Cost of Moral Hazard," *International Economic Review*, 41(3), 309–338.

MARSCHAK, J. (1953): *Studies in Econometric Methods*chap. Econometric Measurements for Policy and Prediction, pp. 1–26. Wiley, New York (NJ).

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.

MCFADDEN, D. L. (1978): "Modelling the Choice of Residential Location," in *Spatial Interaction Theory and Residential Location*, ed. by A. Karlkvist, pp. 75–96. North Holland, Amsterdam.

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.

PAARSCH, H., AND B. SHEARER (2000): "Piece Rates, Fixed Wages and Incentive Effects: Statistical Evidence from Payroll Records," *International Economic Review*, 41(1), 59–92.

——— (2009): "The Response to Incentives and Contractual Efficiency: Evidence from a Field Experiment," *European Economic Review*, 53(5), 481–494.

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

SHEARER, B. (2004): "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment," *Review of Economic Studies*, 71(2), 513–534.

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.

TODD, P. E., AND K. I. WOLPIN (2008): "Ex Ante Evaluation of Social Programs," *Annales d'Economie et de Statistique*, 91-92(06-022).

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

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 Appendix

A.1 Indexes

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

$$\begin{cases} A_i^t = \sum_a A_{a,i}^t \; p_a^{1996} & \text{if } 1996 \leq t < 2000, \\ A_i^t = \sum_a (A_{a,i}^t \; p_a^{2000}) \; \dfrac{\sum_a A_{a,i}^{2000} \; p_a^{1996}}{\sum_a A_{a,i}^{2000} \; p_a^{2000}} & \text{if } 2000 \leq t \leq 2002. \end{cases} \tag{29}$$

The same price are used for weighting billable and non-billable services. The variable A_i^t in (29) then stands for either non-billable services, $A_i^t = NBS_i^t$, or billable ones, $A_i^t = BS_i^t$, 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 $\overline{A}_a^t$ 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:

$$\begin{cases} p^t = \dfrac{\sum_a \overline{A}_a^{1996} \; p_a^t}{\sum_a \overline{A}_a^{1996} \; p_a^{1996}} & \text{if } 1996 \leq t < 2000, \\ p^t = \dfrac{\sum_a \overline{A}_a^{2000} \; p_a^t}{\sum_a \overline{A}_a^{2000} \; p_a^{2000}} \dfrac{\sum_a \overline{A}_a^{1996} \; p_a^{2000}}{\sum_a \overline{A}_a^{2000} \; p_a^{2000}} & \text{if } 2000 \leq t \leq 2002. \end{cases} \tag{30}$$

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 Caps

The ceiling does not apply to emergency from 1996 to 2001 whereas the hospital activities are excluded from 2001. Moreover, the income earned at private practice is decreased by a fixed proportion, denoted by a ($a = 35\%$ for all specialities except diagnostic radiology, for which $a = 75\%$) before cap application. Since we do not distinguish practice based on the health care places where they take place, we cannot generate consumption after the true ceiling. We instead adjust the caps so as to include the allocation of practice between health care places. Let R_i^p denote the income earned by physician i at private practice and R_i^e the income apart for private practice included in the cap (except for emergency for 1996-2001, hospital since 2001). For each observation, we define the share of practice in each practice place as: $s_i^e = R_i / R_i^e$ and $s_i^p = R_i / R_i^p$ where R_i is the observed total income for physician i. Denoting by $\widetilde{C}$ the level of the cap for the specialty the physician i belongs to, the physician's income is subject to the cap if:

$$s_i^e.R_i + s_i^p.R_i(1-a) \geq \widetilde{C} \Leftrightarrow R_i \geq \frac{\widetilde{C}}{s_i^e + s_i^p.(1-a)} \equiv C_i.$$

This transformation provides us a cap on the whole income that incorporates practice allocation.

A.3 Non-Convex Budget Set

Let the budget set be given by

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

Let (X^o, e^o, h^o) and (X^1, e^1, h^1) be on the frontier of XM with $h_1 > h_o, e_1 > e_o$ and

$$X^o - pe^o h^o - wh^o = X^1 - pe^1 h^1 - wh^1 = y. \tag{31}$$

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

$$(X^3, e^3, h^3) = \lambda(X^1, e^1, h^1) + (1-\lambda)(X^o, e^o, h^o). \tag{32}$$

Convexity requires

$$X^3 - pe^3 h^3 - wh^3 \leq y \tag{33}$$

or,

$$\lambda X^1 + (1-\lambda)X^o - p\left[\lambda e^1 + (1-\lambda)e^o\right]\left[\lambda h^1 + (1-\lambda)h^o\right] - w\left[\lambda h^1 + (1-\lambda)h^o\right] \leq y. \tag{34}$$

But

$$y = \lambda X^1 + (1-\lambda)X^o - \left[\lambda pe^1 h^1 + (1-\lambda)pe^o h^o\right] - \left[\lambda wh^1 + (1-\lambda)wh^o\right]. \tag{35}$$

So (34) can be written

$$\begin{aligned}
-p\left[\lambda e^1 + (1-\lambda)e^o\right]\left[\lambda h^1 + (1-\lambda)h^o\right] &\leq -p\left[\lambda e^1 h^1 + (1-\lambda)e^o h^o\right] \\
\left[\lambda e^1 + (1-\lambda)e^o\right]\left[\lambda h^1 + (1-\lambda)h^o\right] &\geq \lambda e^1 h^1 + (1-\lambda)e^o h^o \\
\lambda(\lambda-1)e^1 h^1 + \lambda(\lambda-1)e^o h^o &\geq \lambda(\lambda-1)e^1 h^o + \lambda(\lambda-1)e^o h^1 \\
e^1 h^1 + e^o h^o &\leq e^1 h^o + e^o h^1 \\
\left(e^1 - e^o\right)\left(h^1 - h^o\right) &\leq 0.
\end{aligned} \tag{36}$$

But this contradicts $h_1 > h_o, e_1 > e_o$, so the budget set is not convex.

A.4 Equivalence

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

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

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

$$\begin{aligned}
U(X^\star, h^\star, e^\star) > U(X', h', e') \quad & \forall (X', h', e') \in XM, (X', h', e') \neq (X^\star, h^\star, e^\star) \iff \\
U(X^\star, h^\star, A^\star/h^\star) > U(X', h', A'/h') \quad & \forall (X', h', e') \in XM, (X', h', e') \neq (X^\star, h^\star, e^\star) \iff \\
u(X^\star, h^\star, A^\star) > u(X', h', A') \quad & \forall (X', h', e') \in XM, (X', h', e') \neq (X^\star, h^\star, e^\star).
\end{aligned} \tag{37}$$

A.5 Comparative Statics

To perform the comparative statics, we use the transformed utility function $u = u(X, h, A)$. Assuming that the second-order conditions are satisfied, let the expenditure function $m(w, p, \bar{u})$ be the solution to the standard dual program $\min_{\{X,h,A\}} \; X - wh - pA$ subject to $\bar{u} - u(X, h, A) \leq 0$. 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,

$$\begin{aligned} \frac{\partial m(w,p,\bar{u})}{\partial w} &= \quad -\tilde{h}(w,p,\bar{u}) \\ \frac{\partial m(w,p,\bar{u})}{\partial p} &= \quad -\tilde{A}(w,p,\bar{u}). \end{aligned} \tag{38}$$

Also, from the concavity of the expenditure function,

$$\begin{aligned} \frac{\partial m(w,p,\bar{u})^2}{\partial w^2} &= \quad -\frac{\partial \tilde{h}(w,p,\bar{u})}{\partial \hat{w}} \leq 0 \\ \frac{\partial m(w,p,\bar{u})^2}{\partial p^2} &= \quad -\frac{\partial \tilde{A}(w,p,\bar{u})}{\partial \hat{p}} \leq 0, \end{aligned} \tag{39}$$

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

$$\frac{\partial m(w,p,\bar{u})^2}{\partial w \partial p} = \quad -\frac{\partial \tilde{h}(w,p,\bar{u})}{\partial p} \underset{<}{\overset{>}{}} 0, \tag{40}$$

which demonstrates (7). Finally,

$$\tilde{e}(w,p,\bar{u}) = \quad \frac{\tilde{A}(w,p,\bar{u})}{\tilde{h}(w,p,\bar{u})}, \tag{41}$$

from which

$$\frac{\partial \tilde{e}(w,p,\bar{u})}{\partial p} = \quad \frac{\tilde{h}\frac{\partial \tilde{A}}{\partial p} - \tilde{A}\frac{\partial \tilde{h}}{\partial p}}{\tilde{h}^2} \underset{<}{\overset{>}{}} 0, \tag{42}$$

which demonstrates (8).