How to regulate heterogeneous hospitals

Brigitte Dormont, Carine Milcent

How to regulate heterogenous hospitals?*

Brigitte Dormont†, and Carine Milcent§

January 14, 2005

Abstract

This paper presents two alternative payment systems to reduce hospital inefficiency. In both systems, one part of the payment is fixed ex ante and allows for observable patient and hospital heterogeneity. The first system is a mixed payment that retrospectively reimburses unobservable hospital heterogeneity specified by hospital fixed effects, but does not reimburse costs due to transitory moral hazard. The second system sets a prospective payment for all the non-observable characteristics, without reimbursing cost deviations due to either transitory moral hazard or hospital specific effects. The advantage of the first payment system is that it creates incentives to reduce transitory moral hazard while guaranteeing high quality of hospital services. Econometric estimates are performed on a sample of 7,314 stays for acute myocardial infarction observed in 36 French public hospitals

*We are grateful for helpful comments from Werner Antweiler of the Faculty of Commerce at the University of British Columbia, Alberto Holly of the University of Lausanne and Michel Mougeot of the University of Besançon. We also thank the participants of the NBER Summer Institute Workshop (Boston) as well as participants of the Crest-LEI and Delta seminars in Paris and participants of the twelfth European Workshop on Econometrics and Health Economics (Menorca) for useful comments. We are also wish to thank to two anonymous referees whose comments helped to improve the paper. This study was funded in part by grants from the DREES (Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques) of the French Ministry of Labor and Solidarity. Any errors are our responsibility.

†Corresponding author: Brigitte Dormont Thema UPX, Bâtiment G, 200 avenue de la république 92001 Nanterre cedex, France. Tel: 00 33 1 40 97 78 36. Fax: 00 33 1 40 97 59 73 . E-mail: dormont@u-paris10.fr.

‡Thema-CNRS, University of Paris10-Nanterre, France and the Institut d'Economie et de Management de la Santé (IEMS), Lausanne, Switzerland.

§Delta-CNRS, Ecole nationale Supérieure, Paris, France and the Institut d'Economie et de Management de la Santé (IEMS), Lausanne, Switzerland.

over the period 1994 to 1997. Transitory moral hazard is far from negligible: its standard error is about 30 % of the standard error we estimate for hospital fixed effects (permanent unobservable heterogeneity). Simulations show that a cost reduction of about 20 % can be expected from implementation of a payment system which allows for permanent unobserved heterogeneity and eliminates only transitory moral hazard.

JEL classification: C23, H51, I18

Keywords : Hospital costs, Prospective Payment System, moral hazard, unbalanced panel data.

1 Introduction

This paper proposes a payment system that creates incentives to increase hospital efficiency when hospitals are heterogeneous, without reducing quality of care.

In most Western European countries, a global budget system was introduced for cost containment purposes in the late 70's or early 80's. In these countries, most hospitals are public or publicly financed. The global budget system is still widely used in many European countries. It is also applied in Switzerland and in the USA for the hospitals managed by the Department of Veterans Affairs.

This method of payment consists of an annual budget fixed in advance which does not vary with the volume of services delivered. It has a number of drawbacks: underservice, risk selection or inefficiency. The type of drawbacks depends on whether the budget constraint is hard or soft. At present, there is growing pressure to reform hospital reimbursement systems through the introduction of a Prospective Payment System per DRG. In France, a gradual introduction of a PPS is planned for 2004-2005. This study explores different optimal reimbursement systems for heterogeneous hospitals as an alternative to the current global budget regime in France.

However, the implications of this paper are not restricted to the French case and go beyond the scope of hospital payment systems. Our approach could be applied to other areas of health care financing. We show how to identify a certain component of moral hazard, one that can have a sizeable impact on cost variability. Our identification method can be applied in situations where there is controversy about the sources of cost variability and the respective roles of inefficiency versus legitimate permanent heterogeneity.

An example in the USA is the controversy regarding the use of Adjusted Average Per Capita Cost (AAPCC) to calculate Medicare Managed Care Reimbursement. AAPCC is based on a blend of risk adjusted rates and of average expenditures computed at the

local level. Wennberg *et al.* (2002) observe that Medicare spending based on AAPCC varies widely between regions. For instance, the difference in lifetime Medicare spending between a typical sixty-five-year-old in Miami and one in Minneapolis is more than $50,000. The variations persist even after differences in health are corrected for. The controversy is about the reasons for the observed difference: is it due to moral hazard or to other differences that are not captured by risk adjusted rates ? Using our method, it would be possible to isolate one component of moral hazard and efficiency could be improved by an appropriate method of payment, while still reimbursing a part of unexplained cost heterogeneity between regions.

The theoretical foundations of a fully prospective payment system per stay have been defined by the yardstick competition model of Shleifer (1985). However, this model is based on rather unrealistic assumptions: homogeneity of hospitals, homogeneity of patients for the same pathology, fixed quality of care. Many studies have pointed to possible negative effects of careless implementation of a PPS, namely patient selection and lower care quality (Newhouse (1996)).

In order to avoid these drawbacks, many authors have advocated a mixed payment system, combining a lump sum and the actual cost. However, such a system is rather difficult to put into practice: its specification can depend on unobservable variables or functions. This leads to questions that we take up in the case of France. How can we identify the costs corresponding to efficient activity ? To what extent should patient and hospital heterogeneity be allowed for in a payment system?

Drawing on Shleifer's theory of yardstick competition, we develop an econometric model where hospital variability is explained by patient and hospital characteristics. From the regulator's perspective, some of these characteristics are observable and some are not. We propose two alternative payment systems in order to reduce hospital inefficiency.

We use a three dimensional nested database of 7,314 stays for acute myocardial in-

farction observed in 36 French public hospitals over the period 1994 to 1997. Information is recorded at three levels: stays are grouped within hospitals and hospitals are observed over several years. The structure of our panel data allows us to identify one component of unexplained cost variability: short term moral hazard.

This article is organized as follows. In section 2, we describe the data. In section 3, which is devoted to the theoretical background, we propose an extension of Shleifer's basic model and define an optimal payment rule. The specification of the cost function is given in section 4, which shows how we identify some components of unexplained cost variability and defines our two payment methods. Our results are presented in section 5, together with the methods and specification tests. In section 6, we simulate the implementation of our two payment methods and evaluate the potential budget savings. Section 7 concludes.

2 Description of the data

We have at our disposal a sample of 7,314 stays for acute myocardial infarction (AMI) observed in 36 French public hospitals from 1994 to 1997. In France, public[1] hospitals account the large majority of total admissions (2/3 of admissions for AMI). Our sample was extracted from the PMSI[2] cost database. Classification of stays by Diagnosis Related Group (DRG) is performed on the basis of diagnoses and procedures implemented during the stay. In order to obtain a high degree of homogeneity in pathologies, we selected patients who where at least 40 years old with acute myocardial infraction (AMI) as the main diagnosis and grouped in the same DRG: uncomplicated AMI (DRG 179).

For each stay, we have information about the cost of the stay, secondary diagnoses, procedures implemented, mode of entry into the hospital (coming from home or transferred

[1] In France and in this article, the term "public hospitals" means hospitals belonging to the public sector and most of private-not-for-profit hospitals.

[2] PMSI stands for *Programme de médicalisation des systèmes d'informations,* which collects information about hospital activity.

from another hospital), mode of discharge (return home or transfer), length of stay, age and gender of the inpatient.

The database gives access to rich, detailed information about stays. However, we cannot follow the same inpatient through successive hospital stays. There is no information about the patient's quality of life after the stay, about readmission just after the observed stay, about infections contracted during the stay. In addition, we have no information about the quality of services provided in terms of comfort or alleviation of pain. Participation in the cost database program is voluntary for hospitals and the number of participating hospitals is limited. They consent to give detailed information about their costs, which means that they must have accounting systems that enable them to provide such information.[3]

Our panel data exhibit a rather complex structure. Information is recorded at three levels. The panel is unbalanced in several dimensions: not only does the number of stays recorded vary across hospitals for a given year but also the length of the observation period varies across hospitals.

2.1 Patients and hospitals

Together with drug therapy (aspirin, beta blockers, etc.), uncomplicated AMI patients (DRG 179) can receive various treatments such as thrombolytic drugs, cardiac catheterization (hereafter denoted as CATH) and percutaneous transluminal coronary angioplasty (PTCA). Catheterization is a specialized procedure used to view the blood flow to the heart in order to improve the diagnosis. Angioplasty (PTCA) appeared more recently than bypass surgery. It is an alternative, less invasive procedure for improving blood flow in a blocked artery.

[3] Using an exhaustive database of AMI patients with no information about costs, we have carried out a comparative analysis of patient characteristics and procedures implemented. The results show that our data can be considered representative of AMI stays in French hospitals.

In France, the use of an innovative procedure such as catheterization or angioplasty does not lead to classification of a stay into a specific DRG.[4] These innovative procedures are most often performed within DRG 179: 76.1 % of CATHs and 82.8 % of PTCAs. Since they do not lead to classification in a specific DRG, these costly procedures would not lead to a specific payment under a prospective payment system. A payment system which does not take these procedures into account would therefore penalise the innovative hospitals which use them and give hospitals incentives to select patients.

Basic features of the data are presented in table 1. Most of the patients are men (73.8%). They are rather young. 89 % of patients come from home. 64 % of discharged patients return home and 36 % are transfered to another hospital.[5] Catheterization is performed for 38 % of the stays classified in DRG 179 and angioplasty in 12 %

Stays are recorded for 36 hospitals over the period 1994-1997 (table 2). A sizeable proportion of hospitals never perform catheterization or angioplasty. These procedures require specific skills and high-tech facilities. For a given year, a hospital is considered to be innovative (INNOV) if it has performed catheterization for at least 2 % of the stays or at least one angioplasty. A hospital can be non-innovative one year and perform high-tech procedures the year after. On average over the four years, 60% of hospitals are classified as innovative and these hospitals account for 71.5 % of the recorded stays.

To complete our database, we have also recorded information about hospital type from the SAE survey.[6] There are three types of hospitals: a CHR (*Centre Hospitalier Regional*) is a public teaching hospital with research activities; PRIV stands for a private not-for-profit hospital (these hospitals have only recently been subject the global budget system

[4] In the US classification, stays with angioplasty are grouped in a specific DRG (DRG 112).

[5] AMI with death are grouped in another DRG (GHM 180). The average death rate for all AMI patients is 9 %.

[6] The "Statistique Annuelle des Etablissements de santé" (SAE) is an annual survey which covers all French public hospitals.

and only partially so); PUB refers to other public hospitals.[7] All the CHR and most of the PRIVs are innovative hospitals.

Table 4 shows correlation coefficients between hospital type, innovative hospitals and averaged indicators computed at the hospital-year level (95 observations). CHRs are innovative and have a low rate of discharge through transfer to another hospital. Private not for profit hospitals (PRIVs) are characterized by a high rate of use of innovative procedures and a high rate of admissions through transfers. Other public hospitals are rather non innovative. Patient flows towards innovative hospitals appear clearly in (i) the positive correlation coefficients we find between admission rates through transfers and CATH or PTCA rates; (ii) the negative correlation coefficients we find between discharge rates through transfers and CATH rates.

2.2 Costs

Table 5 gives average costs. Average cost per stay is equal to 4,198 € with a standard error of 2,863 €. On average, a stay is more costly when an innovative procedure has been implemented. As concerns hospital characteristics, stays are more expensive in teaching and in private not-for-profit hospitals. Stays are also costlier in innovative hospitals.

2.3 Historical context

In France, public hospital budgets have been based on a global budget system for more than ten years, including the years 1994-1997 that we study. A complete information system which classifies inpatient stays by DRG has been set up, but a PPS has not been implemented. No reform of financing was undertaken from 1994 to 1997 (a gradual introduction of a PPS is planned for 2004-2005). Budgets have no direct link to the actual

[7] The SAE survey provides other indicators on hospitals, such as the number of beds, the occupation rate of beds, the diversification of activities within hospitals. However, the high number of missing observations makes a complete descriptive analysis impossible. On the basis of a restricted number of observations, we find that CHRs are large hospitals with highly diverse activities. On the other hand, private not- for- profit hospitals (PRIVs) concentrate on a small number of activities.

production of hospitals. Hospitals are managed by salaried administrators and do not keep the gains resulting from cost reduction efforts. In practice, the actual budget depends on the outcome of negotiations between the regulator and the hospital manager. In addition, hospitals are subject to a more or less soft budget constraint. This regulation leads to inequity and inefficiency in the allocation of ressources (Mougeot (1999)).

ici

3 Theoretical background

The models used to study hospital payment systems are devoted to the general problem of local monopoly regulation. They consider the theoretical framework of an agency relationship between the regulator and the hospital, where the regulator has poor information about the cost reduction effort provided by the hospital manager (moral hazard). For a particular disease, one assumes that the cost of one stay in a hospital h is given by: $C_h = a_h - e_h$, where a_h and e_h are private information of a hospital. a_h is a technology parameter which represents the hospital's cost characteristics. It is a decreasing function of hospital productivity. e_h represents the manager's effort to reduce cost. The higher the effort provided, the lower the moral hazard. A hospital exerting effort level e_h incurs a disutility denoted by $\varphi(e_h)$. $\varphi(.)$ is a continous function with $\varphi'(.) > 0$ and $\varphi''(.) < 0$. The services provided by hospital h generate a surplus $S_h > 0$. In return, the regulator compensates the hospital through a monetary transfer P_h. Hospitals are supposed to keep the rent earned through cost-reducing efforts and to face a hard budget constraint. Thus, each hospital h chooses its level of effort in order to maximise its utility given by :

$$U_h = P_h - \varphi(e_h) - C_h$$

Each hospital is supposed to be a local monopoly. One assumes that there is no collusion between hospitals. The regulator has to define the levels of transfers which maximise social welfare subject to the constraint that hospitals must not be in state of bankruptcy (λ takes distortions from taxation into account):

$$Max \sum_{h} (S_h + U_h - (1+\lambda)P_h), \text{ subject to : } U_h \geqslant 0 \; \forall \; h$$

3.1 The yardstick competition model

A prospective payment system (PPS) leads hospitals to exert the first-best level of effort and to have a balanced budget (with no rent and no deficit). A PPS is a fixed price contract. Since the payment is a lump-sum defined irrespective of actual cost, it gives the hospital a perfect incentive for cost reduction ($\varphi'(e^*) = 1$). At this stage, the problem is solved in part only. Indeed, a_h is a private information of the hospital: the level of the lump-sum fixed by the regulator can lead the hospital to bankrupty or generate rents. Thus, the problem of the regulator is to find the level of payments which is equal to the cost arising when the hospital is efficient.

The yardstick competition model (Shleifer (1985)) solves the problem of informational asymetries by assuming that the technology parameters are all identical between hospitals: $a_h = a \quad \forall h$. In this case, differences in costs are only caused by moral hazard:

$$C_h = a - e_h \tag{1}$$

The yardstick competition scheme consists in offering to each hospital a rule of payment defined on the basis of the average costs observed for all other hospitals than h at the end

of the year. The payment rule is:

$$P_h = \varphi(e^*) + \overline{C}_h, \quad \text{where } \overline{C}_h = \frac{\sum_{k \neq h} C_k}{H-1}. \tag{2}$$

H is the number of regulated hospitals.

Here, $\overline{C}_h$ is defined so as not to be influenced by C_h: the resulting payment is equivalent to a fixed price contract. Since the payment rule is announced at the beginning of the year, the average $\overline{C}_h$ is *ex post* equal to the cost corresponding to the first-best level of effort:

$$C_h = a - e^* = \overline{C}_h, \ \forall \ h$$

Transfers P_h are such that each hospital breaks even:

$$P_h = a - e^* + \varphi(e^*). \tag{3}$$

Expression (3) shows that P_h is a lump-sum equal to the level of cost corresponding to an efficient activity. In other words, P_h is equal to the level of costs of a hospital when there is no moral hazard. Given our notations, the additional costs induced by moral hazard is equal to $(e^* - e_h)$. The payment rule leads *ex post* to: $e_h = e^*$, $\forall h$. There is no longer moral hazard and the hospitals receive a payment (3) equal to the sum of the minimum level of costs $(a - e^*)$ and of the disutility of the optimal level of effort $\varphi(e^*)$.

This ideal representation sets up the theoretical foundations of a fully prospective payment system. This model is based on rather unrealistic assumptions: homogeneity of hospitals, homogeneity of patients for the same pathology, fixed quality of care.

Many studies have underscored the great diversity in the conditions of care delivery for hospitals (teaching status, share of low income patients, local wage level, etc.). For instance, Pope (1990) shows that input prices can differ according to location, and that a

hospital can be characterized by specific quality of services or severity of illness of admitted patients. These studies point out the risks of a fully prospective payment system: patient selection and lower care quality.

In order to avoid these drawbacks, many authors have tried to improve the basic model by removing hypotheses such as patient and hospital homogeneity. (Keeler (1990), Pope (1990), Ma (1994, 1998), Ellis (1998), Laffont and Tirole (1993)). It is also possible to consider extensions which introduce endogenous levels of number and quality of treatments (Ma (1994), Ellis (1998), Chalkley and Malcomson (2000)). Using various theoretical frameworks and hypotheses, some authors show that the social welfare can be improved by a mixed payment system combining a lump-sum and a reimbursement of the actual cost of treatment. However, the implementation of a mixed payment system is not straightforward: the proportions of the lump-sum and the actual cost are defined very differently, depending on the theoretical model used, its main hypotheses and its parameterisation. Moreover, its definition often relies on unobservable variables or functions (such as the effort disutility function, in Laffont and Tirole's model).

In this paper, we consider an extension of the basic Shleifer's model, where the regulator is supposed to use the information available about observable sources of hospital cost heterogeneity.

3.2 Extension of the basic model

Consider $C_{i,h,t}$ the cost of stay i in hospital h during year t. We now suppose that the sources of hospital cost variability are partially observable. The regulator is able to observe the share $\widetilde{C}_{iht}$ of the costs which is linked to observable patient and hospital characteristics. One has:

$$C_{i,h,t} = \underbrace{X'_{i,h,t}\gamma_t + W'_{h,t}\alpha + Q'_\lambda + c_t}_{\widetilde{C}_{iht}} + \underbrace{a - e_{h,t}}_{\Gamma_{ht}}. \tag{4}$$

$X'_{i,h,t}$ represents individual patient characteristics such as age-gender cross effects, admission and discharge modes, length of stay. $W'_{h,t}$ are observable hospital characteristics which can vary over time: the hospital's ability to perform innovative procedures, the implementation rates of high-tech procedures, the rates of admission or discharge through transfer. Q'_h are observable hospital characteristics which do not vary over time, such as the type: teaching, private not for profit or other public hospital.

In expression (4) the observed cost has two components: $C_{i,h,t} = \widetilde{C}_{iht} + \Gamma_{ht}$. The first one, $\widetilde{C}_{iht}$, is the observable hospital cost heterogeneity: $\widetilde{C}_{iht} = X'_{i,h,t}\gamma_t + W'_{h,t}\alpha + Q'_\lambda + c_t$. The second one, denoted Γ_{ht}, is equal to the cost heterogeneity considered by the basic Shleifer's model (see expression (1)), where e_{ht} is not observed by the regulator:

$$\Gamma_{ht} = a - e_{ht}. \tag{5}$$

Given these notations, the additional costs induced by moral hazard are, like in the basic model, equal to $(e^* - e_{ht})$. Consider :

$$\overline{\Gamma}_h = \frac{\sum_{t=1}^{T} \sum_{k \neq h} \Gamma_{ht}}{\sum_{t=1}^{T} (H_t - 1)}, \tag{6}$$

where H_t is the number of hospitals observed in year t.

The payment rule is now defined by:

$$P_{iht} = \widetilde{C}_{iht} + \varphi(e^*) + \overline{\Gamma}_h \tag{7}$$

Here, $\overline{\Gamma}_h$ is defined[8] so as not to be influenced by Γ_{ht}. Assuming that the explanatory variables of $\widetilde{C}_{iht}$ are exogenous, i.e. that the hospital cannot manipulate their level in reaction to the proposed payment, the result of the payment is a fixed price contract. As explained before, the average $\overline{\Gamma}_h$ corresponds *ex post* to a situation with no moral hazard:

$$\overline{\Gamma}_h = a - e^* \; \forall \; h \tag{8}$$

The payment rule leads *ex post* to: $e_{ht} = e^* \quad \forall \; h, t$. There is no longer moral hazard. On the basis of rule (7), each hospital receives a payment corresponding to the minimum level of costs, for a given activity.

Each hospital breaks even with transfers P_{iht} equal to:

$$P_{iht} = \underbrace{X'_{i,h,t}\gamma_t + W'_{h,t}\alpha + Q'_h\lambda + c_t}_{\widetilde{C}_{iht}} + \varphi(e^*) + \; a - e^* \tag{9}$$

4 Econometric specification of the cost function

When η_h are assumed to be random, the disturbance $\eta_h + \varepsilon_{h,t} + u_{i,h,t}$ has a ,

Our information is recorded at three levels (stays-hospitals-years), including the individual level of hospital stay.[9] The transition to the econometric specification makes it necessary to take into account disturbances which are linked to patients' and hospitals' unobserved heterogeneity, omited variables and measurement errors.

Theoretical model (4) thus becomes:

$$C_{i,h,t} = X'_{i,h,t}\gamma_t + W'_{h,t}\alpha + Q'_h\lambda + c_t + a + \eta_h + \varepsilon_{h,t} + u_{i,h,t} \tag{10}$$

[8] Here $\overline{\Gamma}_h$ is defined as an average computed over several years. This is to consider a definition in accordance with the computation done in our empirical approach, where $\overline{\Gamma}_h$ is defined over 4 years. Notice that this doesn't change any prediction of the effect of the payment rule, as soon as it is announced that the computation of $\overline{\Gamma}_h$ is updated every year.

[9] Therefore, our approach is different of papers which evaluate efficiency using data relative to average costs per hospital. A synthetic survey of this literature can be found in Linna (1998).

In contrast to theoretical model (4), we consider in econometric specification (10) a hospital specific effect η_h, a hospital-year specific effect ε_{ht} and a random error term at the patient level $u_{i,h,t}$. The structure of our data results in a nested structure of the disturbance, which means that each each successive component of the error term is imbedded within the preceding component. The random error term $u_{i,h,t}$ is assumed to be iid $(0, {\sigma_u}^2)$. It takes unobservable patient heterogeneity into account. $\varepsilon_{h,t}$ is a disturbance supposed to be iid $(0, \sigma_\varepsilon^2)$ and uncorrelated with $u_{i,h,t}$. We provide below detailed interpretations of η_h and ε_{ht}. Notice right here that the unobserved level of effort e_{ht}, which appears in theoretical model (5), is a component of the term $\eta_h + \varepsilon_{h,t}$.

As stated in the data section, the costs we observe result from an activity financed on the basis of a global budget system. Cost variability is therefore influenced by several factors: patient characteristics, hospital characteristics and inefficiency.

Why should inefficiency influence our "real" data ? Because of the way the global budget is implemented in France: as stated above in section 2.3, budgets have no direct link to the actual production of hospitals and the budget constraint is rather soft. In fact, inefficiency is more or less possible, depending on the generosity obtained by the hospital manager from the regulator when bargaining for the budget.

Given patient characteristics, cost variability can stem from hospital characteristics such as hospital type (CHR, PRIV, PUB) and size, diversification of activities, quality of services provided (performance of innovative procedures, comfort, alleviation of pain), skill level of nurses and doctors, quality of hospital management. Some of these factors are observable, some of them cannot be observed.

In this paper, we assume that the regulator has the same position as the econometrician. More exactly, we assume that the regulator has an access to our database (the PMSI database) to set the payments. Therefore, the sharing out of variables between observable and unobservable components is the same for the regulator and the econometrician. The

observable characteristics are the variables $X'_{i,h,t}$ for the patients and the variables $W'_{h,t}$ et Q'_h for the hospitals. In (10), c_t is a fixed temporal effect, which can be linked to technological progress, the pace of price growth and the general trend of hospital budgets. Given the observable characteristics, cost variability depends, in specification (10), on the term:

$$\eta_h + \varepsilon_{h,t} + u_{i,h,t}$$

4.1 Interpretation of hospital specific effects η_h

The hospital specific effects η_h can be assumed to be random or fixed. These effects allow us to specify the time-constant unobservable hospital heterogeneity. In our theoretical framework, we have considered that the regulator has poor information about the cost reduction effort provided by the hospital manager (moral hazard) but has information about *observable* sources of hospital cost heterogeneity $\widetilde{C}_{iht}$. However, the cost variability can also be influenced by *unobserved* hospital characteristics explaining its efficiency (adverse selection). Our theoretical framework considers only moral hazard. It does not address the issue of designing an optimal contract to deal with adverse selection.[10] Nevertheless, the costs we observe are influenced by unobserved hospital characteristics. Therefore we include adverse selection parameters into our econometric specification.

Then, η_h can be seen as the result of three components :

$$\eta_h = \eta_h^{as} + \eta_h^{mh} + \eta_h^{q}.$$

The components of η_h are interpreted in the following way: η_h^{as} is an adverse selection parameter. The hospital's activity is more or less costly, depending on its infrastructure or on the existence of economies of scale or of scope.[11] η_h^{mh} represents *long term* moral

[10] On this issue, see for instance Laffont and Tirole (1993).

[11] Finding evidence of links between hospital size (and diversification of activities) and the level of costs

hazard: the hospital management can be permanently inefficient. To provide an example of bad functioning, we can think about an obsolete elevator which is very slow and subject to frequent failures. Despite it should be replaced, this is not done during several years. The term η_h^q takes the time-invariant component of care quality into account. Transitory variations in quality are rather unlikely: the staff of public hospitals is binded to a unit. It cannot be moved from one unit to another one, according to the needs. The civil servant status involves rigidity in the hospital organization of work. Like for the staff, the public hospital facilities cannot be modified very quickly. Any decision depends on a central public administration. Thus, any modification of the quality of care in one public hospital requires a sizeable delay to be implemented. Since our data covers a period of four years, the possibility of a significant upgrading of quality is rather limited.

4.2 Interpretation of $\varepsilon_{h,t}$

The disturbance $\varepsilon_{h,t}$ is defined as the deviation, *ceteris paribus*, for a given year, of hospital h's cost in relation to its average cost level. It can be seen as the result of two components:

$$\varepsilon_{h,t} = \varepsilon_{h,t}^{mh} + \varepsilon_{h,t}^{tr}$$

$\varepsilon_{h,t}^{mh}$ is an indicator of the effect on costs of transitory cost-reducing effort, i.e an indicator of *transitory* moral hazard. For instance, the manager can be more or less rigorous when bargaining prices for supplies or for services delivered to the hospital by outside firms.

The remaining part of $\varepsilon_{h,t}$ should reflect the ordinary components of any disturbance, namely measurement errors and omitted transitory variables. Measurement errors are likely to be of slight importance: $\varepsilon_{h,t}^{tr}$ is replicated for each stay in the same hospital h during the same year t. Within this framework, a measurement error can only be a systematic

entails the use of the SAE survey. However, information about the accurate indicators is available only for a restricted number of observations in this survey. We did not find any significant results on these observations.

error in patient registration, or an error in hospital classification. These two possibilities are unlikely. The other possible components of $\varepsilon_{h,t}^{tr}$, are omitted transitory variables such as transitory variations in care quality or transitory shocks affecting hospital's costs. We have seen above that transitory significant changes in quality are rather unlikely within the organizational context of French public hospitals. On the other hand, any hospital can be affected by a shock in a given year. It may be, for instance, an electrical failure. Within the implementation of a Prospective Payment System, we think that the regulator would be well advised to classify *a priori* these incidents as moral hazard, in order to give hospitals incentives to declare them, when the extra costs they induce are justifiable and exceptional. But such a regulation is not applied to the hospitals of our sample. In our data, the variability of $\varepsilon_{h,t}^{tr}$ can therefore be affected by transitory unobserved shocks, the importance of which has to be evaluated empirically, in order to identify the share of $\varepsilon_{h,t}$ due to short term moral hazard.

Given the facts that (i) measurement errors are likely to be negligible; (ii) $\varepsilon_{h,t}^{tr}$ is mainly influenced by transitory shocks which are probably scarce, we think that the influence of $\varepsilon_{h,t}^{tr}$ on the variability of $\varepsilon_{h,t}$ is negligible. An econometric test based on the stochastic cost frontier (SCF) approach gives an empirical support to this conjecture (see section 5.3). Given this result, we can consider that the variability of $\varepsilon_{h,t}^{tr}$ is negligible and interpret the perturbation $\varepsilon_{h,t}$ as an indicator of transitory moral hazard.

4.3 Definition of two methods of payment

Econometric specification (10) can be written as follows:

$$C_{i,h,t} = \widetilde{C}_{iht} + a + \eta_h + \varepsilon_{h,t} + u_{i,h,t} \quad ,$$

where $\widetilde{C}_{iht}$ is the observable hospital heterogeneity. Consider the hospital-year means defined by $C_{.,h,t} = \frac{1}{N_{h,t}} \sum_{i=1}^{N_{h,t}} C_{i,h,t}$, where N_{ht} is the number of stays recorded in hospital h in year t. Computing means at the hospital-year level eliminates the perturbation $u_{i,h,t}$ linked to the sample distribution of stays ($u_{.,h,t} \xrightarrow{P} 0$ when N_{ht} is large[12]). Therefore, one has:

$$C_{.,h,t} \xrightarrow{P} \widetilde{C}_{.ht} + a + \eta_h + \varepsilon_{h,t}$$

In our theoretical model, we have $C_{i,h,t} = \widetilde{C}_{iht} + \Gamma_{ht}$ and the optimal payment is defined by (7): $P_{iht} = \widetilde{C}_{iht} + \varphi(e^*) + \overline{\Gamma}_h$. In order to put this payment into practice, the regulator has to establish a link between the theoretical concept Γ_{ht} and the perturbations of the econometric specification $\eta_h + \varepsilon_{h,t}$. In other words, he has to establish a link between the additional costs induced by moral hazard $\Gamma_{ht} - \overline{\Gamma}_h = e^* - e_{ht}$ and $\eta_h + \varepsilon_{h,t}$. The arguments presented above, together with our SCF analysis, allow us to consider that $\varepsilon_{h,t}$ can be interpreted as transitory moral hazard ($\varepsilon_{h,t}^{mh} \simeq \varepsilon_{h,t}$). The main difficulty concerns the hospital effect: is η_h a legitimate hospital heterogeneity (which would be part of $\widetilde{C}_{iht}$ if it were observable) ? Or is η_h long term moral hazard, which must be crushed by an appropriate method of payment ? We have seen above that η_h can be seen as the result of three components ($\eta_h = \eta_h^{as} + \eta_h^{mh} + \eta_h^{q}$) and that the moral hazard entails only one of these components ($-e_{ht} = \eta_h^{mh} + \varepsilon_{h,t}$). Given the fact that the components of η_h cannot be identified separately, the regulator is reduced to considering two extreme cases, whether η_h is supposed to be legitimate heterogeneity ($\eta_h^{mh} = 0$ and $-e_{ht} = \varepsilon_{h,t}$) or to be entirely due to moral hazard ($\eta_h = \eta_h^{mh}$ and $-e_{ht} = \eta_h + \varepsilon_{h,t}$).

4.3.1 Taking or not unobservable hospital heterogeneity into account

In our theoretical model, the unobserved cost heterogeneity is equal to (5): $\Gamma_{ht} = a - e_{ht}$. As stated above, the regulator can consider two cases as regards the components of moral

[12] On average, $N_{h,t}$ is equal to 77, with a minimum equal to 19 and a maximum equal to 250.

hazard: $-e_{ht} = \varepsilon_{ht}$ or $-e_{ht} = \eta_h + \varepsilon_{ht}$.

a) First method of payment

It relies on the assumption that hospital effects η_h are linked to a legitimate heterogeneity. Given our notations, this comes down to suppose $-e_{ht} = \varepsilon_{ht}$. Thus:

$$\Gamma^1_{ht} = a + \varepsilon_{ht}. \tag{11}$$

The rule of payment is given by:

$$P^1_{i,h,t} = X'_{i,h,t}\ \gamma_t + W'_{h,t}\ \alpha + Q'_h\ \lambda + c_t + \eta_h + \varphi(e^*) + \overline{\Gamma}^1_h\ , \tag{12}$$

with $\overline{\Gamma}^1_h$ defined by (6).

Assuming that hospitals keep the rent earned from more efficient operations, they will exert the optimal cost-reducing effort e^*. Ex post, the following equality will thus be verified : $\overline{\Gamma}^1_h = a - e^*$.

With payment rule (12), the regulator takes the observable characteristics $X'_{i,h,t}$, $W'_{h,t}$ and Q'_h into account. In addition, the payment P^1 allows for permanent unobserved hospital heterogeneity η_h, assuming that it is due to adverse selection or to the care quality. Nevertheless, this payment method still gives incentives to hospitals : cost deviations attributable to transitory moral hazard ε_{ht} are not reimbursed.

b) Second method of payment

The second method of payment is defined on the assumption that hospital effects η_h are entirely due to moral hazard. In this case, $-e_{ht} = \eta_h + \varepsilon_{ht}$ and:

$$\Gamma^2_{ht} = a + \eta_h + \varepsilon_{ht} \tag{13}$$

$$P^2_{i,h,t} = X'_{i,h,t}\ \gamma_t + W'_{h,t}\ \alpha + Q'_h\ \lambda + c_t + \varphi(e^*) + \overline{\Gamma}^2_h\ , \tag{14}$$

This second payment rule takes observable patient and hospital characteristics into account, but "crushes" unobserved heterogeneity $\eta_h + \varepsilon_{h,t}$. Implementing payment rule (14) comes down to interpreting all unobserved hospital heterogeneity as resulting from moral hazard.

4.3.2 Evaluating the *ex post* payments

We have seen in section 3.2 that payment rule (7) leads each hospital to provide the first best cost reduction effort e^*. Therefore, each hospital receives *ex post* a payment corresponding to the *minimum level of costs*, for a given activity.

To evaluate the payments which can arise from the implementation of such a payment rule and the corresponding potential budget savings, we must evaluate the level of costs linked to an efficient activity. Given our theoretical model the costs associated to an efficient activity are equal to the payments arising *ex post* when rule (7) is implemented.

The estimation of cost function (10) allows us to evaluate the costs associated to an efficient activity. However, the definition depends on the assumption relative to the components of moral hazard.

a) First method of payment

η_h being considered as a legitimate heterogeneity, one can estimate the *ex post* payments by:

$$\hat{P}{}^{1}_{i,h,t} = X'_{i,h,t}\overset{\wedge}{\gamma}_t + W'_{h,t}\overset{\wedge}{\alpha} + Q'_h\,\widehat{\lambda} + \hat{c}_t + \overset{\wedge}{\eta}_h + \hat{a} + \underset{h,t}{Min}\,(\overset{\wedge}{\varepsilon}_{h,t} + \hat{u}_{.,h,t})\,, \qquad (15)$$

where we are using consistent estimates of the parameters and disturbances of model (10).

Under assumption (11) and assuming that the most efficient hospital-year observation corresponds to an efficient activity (in other words, that the most efficient hospital provides the level of effort e^*), $\underset{h,t}{Min}\,(\overset{\wedge}{\varepsilon}_{h,t})$ is a consistent estimate of the cost reduction obtained[13] through the provision of effort e^*. Thus: $\underset{h,t}{Min}\,(\overset{\wedge}{\varepsilon}_{h,t}) \xrightarrow{P} \varphi(e^*) - e^*$. Then, $\hat{P}{}^{1}_{i,h,t}$ is a consistent

[13] On our data, the maximal cost reduction is equal to $\varphi(e^*) - e^*$ and not $-e^*$. Indeed the most efficient hospital wants to be reimbursed for the effort disutility $\varphi(e^*)$.

estimator of the cost corresponding to efficient activity. Indeed, the use of consistent estimates of the model implies $\hat{a} \xrightarrow{P} a$ and $\hat{u}_{.,h,t} \xrightarrow{P} 0$. Thus:

$$\left\{\hat{a} + \underset{h,t}{Min}\,(\hat{\varepsilon}_{h,t} + \hat{u}_{.,h,t})\right\} \xrightarrow{P} \varphi(e^*) + a - e^*.$$

b) Second method of payment

If the hospital effects η_h are entirely due to moral hazard, the *ex post* payments are estimated by:

$$\hat{P}^2_{i,h,t} = X'_{i,h,t}\hat{\gamma}_t + W'_{h,t}\hat{\alpha} + Q'_h\,\hat{\lambda} + \hat{c}_t + \hat{a} + \underset{h,t}{Min}\,(\hat{\eta}_h + \hat{\varepsilon}_{h,t} + \hat{u}_{.,h,t}), \qquad (16)$$

where we are using consistent estimates of the parameters and disturbances of model (10).

Here, we assume again that the most efficient hospital-year observation corresponds to an efficient activity (in other words, that the most efficient hospital-year provides e^*). But now we suppose that all the unobserved heterogeneity is resulting from moral hazard (13). Then we compute the payment by taking as a reference point the hospital for which the sum of unobservable characteristics η_h and transitory moral hazard $\varepsilon_{h,t}$ is minimal. More exactly, under (13), $\underset{h,t}{Min}\,(\hat{\eta}_h + \hat{\varepsilon}_{h,t})$ is a consistent estimate of the cost reduction obtained through the provision of effort e^*. Thus: $\underset{h,t}{Min}\,(\hat{\eta}_h + \hat{\varepsilon}_{h,t}) \xrightarrow{P} \varphi(e^*) - e^*$. Then, $\hat{P}^2_{i,h,t}$ is a consistent estimator of the cost corresponding to efficient activity. Indeed:

$$\left\{\hat{a} + \underset{h,t}{Min}\,(\hat{\eta}_h + \hat{\varepsilon}_{h,t} + \hat{u}_{.,h,t})\right\} \xrightarrow{P} \varphi(e^*) + a - e^*.$$

5 Estimation and results

We have chosen a linear specification for the cost function: the dependent variable is $C_{i,h,t}$ and not $Log(C_{i,h,t})$. It is well known that health care expenditures generally have a very

asymmetric distribution. In our case, however, the distribution is truncated on the right because of the selection of stays grouped in DRG 179 (uncomplicated AMI). More costly stays are grouped in other DRGs: complicated AMI or AMI treated by bypass surgery. The tests we have carried out on the distribution of $C_{i,h,t}$ have led us to the conclusion that it is closer to a normal than to a lognormal distribution. More exactly, normality tests led to reject the null hypothesis for both C and $Log(C)$. When we drop the 1% highest observed costs, the skewness is closer to the normal for C ($S = 0.509$) than for $Log(C)$ ($S = -1.117$). Taking Log displaces the distribution to the left, leading to a negative value of the skewness. In addition, one of the results presented above provides an *ex post* justification of our specification choice: we find that the estimates of ε_{ht} do not increase on the raw scale as average hospital costs increase.[14]

5.1 Estimation methods and specification tests

In model (10) the hospital specific effects η_h can be assumed to be random or fixed. They are linked to unobservable hospital characteristics: long term moral hazard, infrastructure, care quality. These characteristics can be correlated with the explanatory variables. For instance, care quality may be higher in a teaching hospital. Assuming that η_h is random comes down to assuming that unobserved heterogeneity is not correlated with the observed characteristics $X'_{i,h,t}, W'_{h,t}$ and Q'_h.

A specification test[15] led us to reject this hypothesis. Therefore, we specify η_h as a fixed

[14] We thank one referee for this remark.

[15] See Dormont and Milcent (2004). This test is not quite straightforward because our panel data is unbalanced in several dimensions: not only does the number of stays recorded vary across hospitals for a given year but also the length of the observation period varies across hospitals. Therefore our three-component error model (when η_h is random) is different from the unbalanced nested error component model considered by Baltagi, Song and Jung (2001) and we have to use the maximum likelihood estimator (MLE) defined by Antweiler (2001). To test for the independence of η_h , we have used an extension of the specification test proposed by Mundlak (1978) for the standard error component model. Writting the correlation between η_h and the explanatory variables as follows: $\eta_h = X'_{.,h,.}\pi_1 + W'_{h,.}\pi_2 + \beta_h$, where β_h is iid $(0, \sigma^2_\beta)$ and assumed to be uncorrelated with $\varepsilon_{h,t}$ nor with $u_{i,h,t}$, the independence test of η_h is equivalent to the restriction test for $H_0 : \pi_1 = \pi_2 = 0$ in the model (estimated by MLE): $C_{i,h,t} = X'_{i,h,t}\gamma_t + W'_{h,t}\alpha + X'_{.,h,.}\pi_1 + W'_{h,.}\pi_2 + a + c_t + \underbrace{\beta_h - \varepsilon_{h,t} + u_{i,h,t}}_{\zeta_{i,h,t}}$

effect. In this case, the model includes hospital dummies and it is not possible to identify parameters λ which reflect the influence of time-invariant variables Q'_h. Specification (10) becomes:

$$C_{i,h,t} = X'_{i,h,t}\gamma_t + W'_{h,t}\alpha + a + c_t + \eta_h + \underbrace{\varepsilon_{h,t} + u_{i,h,t}} \tag{17}$$

This model is a standard error component model, with a disturbance equal to $\varepsilon_{h,t}+u_{i,h,t}$. In this case, feasible generalized least squares (FGLS) lead to a consistent and asymptotically efficient estimate if $\varepsilon_{h,t}$ is not correlated with the explanatory variables.

Two specifications were estimated, related to different lists of explanatory variables W'_{ht}. Model (A) includes indicators close to verifiable characteristics such as the variable indicating whether or not the hospital is innovative and the average rates of admission and discharge through transfers. Model (B) includes additional variables such as the rates of use of innovative procedures, which can be more directly decided on by the hospital.

Tables 6 and 7 display the estimates of the models (A) and (B), and the associated specification tests.

Hausman's tests allowed us to validate the hypothesis that effects $\varepsilon_{h,t}$ are not correlated with the explanatory variables. Notice that the usual statistic of the Hausman test do not allow to consider variables W'_{ht}. So, there is no difference between the tests on models A and B. This test (denoted Hausman test 1) is equivalent to a test for no correlation between $X'_{i,h,t}$ and $\varepsilon_{h,t}$. It led not to reject the null hypothesis (table 7).

To test for the exogeneity of W'_{ht} we used intrumental variables to build another Hausman's specification test (denoted Hausman test 2). Here, we compared the estimator known to be consistent under the null and alternative hypotheses (the error component two-stage least square estimator, EC2SLS (Baltagi, 1981)) with an estimator which is consistent and efficient under the null hypothesis (the feasible generalized least squares estimator, FGLS). Instruments are the secondary diagnoses of the patient. A Sargan test has been imple-

mented in order to check the validity of the instruments used for this Hausman's test. In addition, we examined whether this test could be subject to the weak instrument problem (Staiger and Stock, 1997). For this purpose, we tested for the significance of the instruments in several equations, where each instrumented variable is explained by the instruments and the exogenous regressors. Hausman tests 2 led not to reject the null hypothesis for model (A) and (B) and the Sargan tests did validate the exogeneity of the intruments (table 7). In addition, we found a large significance of the partial correlation between instruments and endogenous explanatory variables, with high statistics[16] and levels of significance lower than 10^{-3}. All these tests validate the hypothesis that W'_{ht} are not correlated with ε_{ht} nor u_{iht}.

Given the fact that both effects have components related to moral hazard, we had to examine whether ε_{ht} could be correlated with the hospital effects η_h. For that purpose, we implemented a third Hausman test, comparing the FGLS applied to $C_{i,h,t} = X'_{i,h,t}\gamma_t + a + c_t + \eta_h + (\varepsilon_{h,t} + u_{i,h,t})$, where η_h is supposed to be fixed and $\varepsilon_{h,t}$ is supposed to be random and not correlated to η_h nor the $X'_{i,h,t}$, to the OLS applied to $C_{i,h,t} = X'_{i,h,t}\gamma_t + a + c_t + \eta_h + \varepsilon_{h,t} + (u_{i,h,t})$, where η_h and ε_{ht} are supposed to be fixed. The test relies on the fact that the OLS applied to the second model are consistent even when the $\varepsilon_{h,t}$ are correlated to the η_h. This test led us not to reject the null hypothesis, with a Wald statistic equal to 23.2 and a p-value close to 1 (the corresponding χ^2 has a degree of freedom equal to 64).

All these tests provide evidence that we cannot reject the hypotheses that the variables $X'_{i,h,t}$ and W'_{ht} are exogenous. Model (17) can be consistently estimated by the FGLS.

[16] For model (A), for instance, the Wald statistics to test the null hypothesis that the 8 instruments are not significant are equal to 559 for Innov, 116 for TI (rate of admission by transfer) and 751 for TX (rate of discharge by transfer) . The corresponding p-values are lower than 0.000.

5.2 Results

The estimated coefficients of the individual characteristics $X'_{i,h,t}$ are reported in table 6. The influence of individual stay characteristics are in accordance with the results generally obtained when studying costs of stays for acute myocardial infarction. The most costly stays are observed for men and cost is a decreasing function of age. One additional day induces, *ceteris paribus*, an average additional cost of about 330-400 Euros. In addition, the estimation of an incomplete specification using only individual patient characteristics $X'_{i,h,t}$ as explanatory variables reveals that 54.2 % of cost variability can be explained by observable patient heterogeneity. A payment system which would not take this heterogeneity into account would give hospitals incentives to select patients.

Once we have taken permanent differences in average costs into account through the fixed effects model, we do not find any significant effect of variable INNOV, nor of other variables W'_{ht}.

The fixed hospital effects specification allows us to obtain consistent estimates of η_h and $\varepsilon_{h,t}$ and of their standard errors σ_η and σ_ε. We have interpreted the disturbance $\varepsilon_{h,t}$ as an indicator of transitory moral hazard. Its influence on cost variability is far from negligible: its estimated standard error (373.4 or 391.6 - model A or B) is above 30 % of estimated σ_η (1213.2 or 1082.3).

To get an idea of the magnitude of the standard errors σ_η and σ_ε, one can compare them to the standard error of stay costs: 2,863 Euros (for an average cost equal to 4,198 Euros). In graphs 1 and 2, we relate estimated effects $\hat{\eta}_h$ and $\hat{\varepsilon}_{h,t}$ to the corresponding average cost per hospital $C_{.,h,.}$ and average cost per hospital-year $C_{.,h,t}$.[17] The observations have been sorted by increasing average cost. Hospital specific effects are linked to average costs per hospital but are far from explaining them entirely (graph 1). Graph 2 show that the magnitude of the transitory moral hazard is not connected to the size of average costs,

[17] These graphs are shown for model A.

giving an *ex post* justification to our linear specification.

5.3 SCF analysis

This analysis was implemented to give empirical support to the assumption that ε_{ht} is entirely due to transitory moral hasard. The basic SCF approach relies on the canonical "half normal" model (Greene, 2004), which uses a parametric specification in order to identify the inefficiency component. The disturbance is split into two components: a normal one, related to statistical noises and a half normal component, related to inefficiency. In our case, we have: $\varepsilon_{h,t} = \varepsilon^{mh}_{h,t} + \varepsilon^{tr}_{h,t}$, where $\varepsilon^{mh}_{h,t}$ is the transitory moral hazard and where $\varepsilon^{tr}_{h,t}$ is linked to measurement errors and transitory shocks. The SCF specification relies on the following assumptions:

$$\varepsilon^{tr}_{h,t} \sim N(0, \sigma^2_{\varepsilon^{tr}}) \text{ and } \varepsilon^{mh}_{h,t} = |\epsilon_{ht}|\,, \text{ with } \epsilon_{ht} \sim N(0, \sigma^2_{\epsilon}). \tag{18}$$

The asymmetry parameter, $\psi = \dfrac{\sigma_{\epsilon}}{\sigma_{\varepsilon^{tr}}}$ gives an evaluation of the magnitude of the inefficiency component. In section 4.2, we explained that $\sigma^2_{\varepsilon^{tr}}$ is likely to be negligible. If this conjecture is right, one should find $\sigma^2_{\varepsilon^{tr}} \to 0$ and $\psi \to \infty$. In this case, $\varepsilon_{h,t} \simeq \varepsilon^{mh}_{h,t}$.

Consider model (17). It can be written as follows:

$$C_{i,h,t} = X'_{i,h,t}\gamma_t + v_{h,t} + u_{i,h,t} \tag{19}$$

$$\text{with} \quad : \quad v_{h,t} = W'_{h,t}\alpha + c_t + \eta_h + \underbrace{\varepsilon^{mh}_{h,t} + \varepsilon^{tr}_{h,t}}_{\varepsilon_{h,t}} \tag{20}$$

In a first regression, we estimate (19), where the $v_{h,t}$ are specified as fixed effects and where $u_{i,h,t}$ is supposed to be iid $(0,\ \sigma^2_u)$. Given our assumptions and the fact that N_{ht} is large enough, the $v_{h,t}$ can be consistently estimated by OLS. This first step makes it

possible to eliminate the patient dimension from the data variability and to get observations at the hospital-year level.

In the second step, we use the first-step estimates $\hat{v}_{h,t}$ and consider the SCF specification, assuming (18) to estimate (20) by the maximum likelihood estimator. This allows us to identify the components of ε_{ht}.

Notice that the constant a has been deleted from the first step regression (19). In second step specification (20), the constant is taken by the hospital fixed effects into account (there is no reference hospital). To avoid multicolinearity, we deleted one year dummy. This treatment of the constant is adopted in order to avoid any *a priori* constraint upon the distribution of hospital effects η_h.

Our specification differs from the basic versions of panel data formulation of the SCF approach. In these versions, the inefficiency is supposed to be time-invariant and reflected by the individual effect (here, η_h). We think that this formulation is not appropriate in our case. As stated repeatedly above, η_h is not only affected by moral hazard, but also by heterogeneity and care quality. These two factors can be symmetrically distributed. Our three dimensional database allows us to consider a less constraining hypothesis.

The estimation of (20) by the maximum likelihood estimator (using the 95 observations of the first-step estimates $\hat{v}_{h,t}$) led to: $\widehat{\sigma}_{\varepsilon^{tr}} = 0.00025$ and $\widehat{\sigma}_{\epsilon} = 605.7816$. Thus $\widehat{\psi} \rightarrow \infty$ ($\psi = 2.394 \times 10^3$). This result gives an empirical support to our conjecture that the variability of ε_{ht} is entirely attributable to the transitory moral hazard. It was reasonnable to expect that measurements errors had a negligible importance at the hospital-year level. But we didn't know the share of the variability of ε_{ht} due to transitory shocks. This parametric SCF analysis show that it is likely to be negligible too.

Table 8 displays the distribution characteristics of various fixed effects and disturbances estimated in steps 1 and 2. It is interesting to notice that the hospital effects $\hat{\eta}_h$ obtained

in step 2 follow a normal distribution.[18] This result suggest that this unobserved hospital heterogeneity is likely to be more influenced by adverse selection and differences in care quality rather than long term moral hazard. However, this interpretation needs further empirical analysis to be confirmed.

6 Simulation of two methods of payment

Our econometric estimates encourage the implementation of a prospective payment system. Indeed, our results have revealed that the transitory moral hazard is far from negligible. As we have seen in section 4.3, the payment rule adopted by the regulator depends on whether η_h is supposed to be legitimate heterogeneity or to be entirely due to moral hazard. In the first case, it is defined by (12), in the second by (14). Under the assumptions of the theoretical model, these payment rules should give to each hospital an incentive to provide the first best cost reduction effort, leading *ex post* to payments

$$(15) \quad : \quad \hat{P}{}^{1}_{i,h,t} = X'_{i,h,t}\overset{\wedge}{\gamma}_t + W'_{h,t}\overset{\wedge}{\alpha} + \hat{c}_t + \overset{\wedge}{\eta}_h + \hat{a} + \underset{h,t}{Min}\,(\overset{\wedge}{\varepsilon}_{h,t} + \hat{u}_{.,h,t})$$

$$(16) \quad : \quad \hat{P}{}^{2}_{i,h,t} = X'_{i,h,t}\overset{\wedge}{\gamma}_t + W'_{h,t}\overset{\wedge}{\alpha} + \hat{c}_t + \hat{a} + \underset{h,t}{Min}\,(\overset{\wedge}{\eta}_h + \overset{\wedge}{\varepsilon}_{h,t} + \hat{u}_{.,h,t})$$

We can simulate the implementation of the two payment rules on our data. The first method of payment exerts a softer constraint on hospitals than the second method of payment. Indeed, payment P^2 ignores all unobserved heterogeneity $(\eta_h + \varepsilon_{h,t})$. With payment P^1 the regulator takes the time-invariant unobservable heterogeneity (η_h) into account, whether it is due to inefficient management or to particularly good care quality.

[18] As we have seen above, $\overset{\wedge}{\varepsilon}_{h,t}$ appears to be entirely half-normal.

6.1 Potential budget savings

Table 9 gives the potential budget savings which can be expected from the implementation of such payment rules.[19] They are computed by measuring the difference between total costs C_{iht} and total *ex post* payments $\hat{P}{}^{1}_{i,h,t}$ or $\hat{P}{}^{2}_{i,h,t}$. We can observe that the bracket defined by P^1 and P^2 is quite wide: the payment rule P^1 leads to potential savings of about 20 %; the payment rule P^2 leads to potential savings of between 51 % and 56 %, depending on the model considered (B or A).

P^1 is indeed the least constraining payment system. Yet, it still leads to substantial potential savings (20 %) because (i) it provides incentives to reduce the costs due to transitory moral hazard $\varepsilon_{h,t}$, (ii) the variability of costs due to transitory moral hazard is sizeable. We thus recommend this method of payment. It avoids using the hospital with the poorest care quality as a benchmark for cost. It takes permanent unobservable differences of quality between hospitals into account. This strategy is advisable, given that quality is a variable that cannot be verified by the regulator.

The next step is to determine which model should be used to establish payments.

- In our estimations and simulations, we have taken the length of stays into account. Nevertheless, the type of payment system that we suggest implementing should not be retrospective in the sense that it should be calculated by stay and not by day. Therefore, we propose reimbursing on the basis of the estimated coefficient of the length of stay in the cost function multiplied by a suitable indicator of the length of stay (an average indicator taking differences in patient and hospital characteristics into account).

- The main difference between the models A and B is that model B integrates charac-

[19] Our simulation are carried out under the assumption that the estimated coefficients remain unchanged despite the payment reform.

teristics such as the frequency of innovative procedures. The reason for integrating procedure rates into the payment system is to avoid patient selection and skimping on treatment. On the other hand, there is a risk of creating incentives for excessive use of procedures (McClellan, 1997). We notice that all variables W'_{ht} are not significant when estimating fixed effects model (table 6) and that potential budget savings do not differ when implemented on the basis of model A or B. It is interesting to notice that taking into account heterogeneity through the hospital fixed effects lead to a non significant influence of variables such as TI and TX (rate of admissions or discharges through tranfers) as well as the frequency of innovative procedures. Indeed, these variables can be manipulated by the hospitals in the short run. It should be more difficult for the hospitals to manipulate their own value of η_h , which derive from the estimation process.

Table 10 records correlation coefficients between costs and payments. A high correlation rate means that the incentives for selecting patients are limited. We observe that substantial budget savings displayed in table 9 are compatible with high correlation coefficients, especially in the between dimension, which is based on the yearly mean by hospital.

6.2 Share of retrospective payment in the first method of payment

Payment method P^2 can be seen as a prospective payment, relaxed by the kind of risk adjustment resulting from the fact that we take observable patient heterogeneity into account. On the other hand, the first method of payment is partly retrospective because it reimburses costs differences due to the hospital effects η_h. More exactly, one can distinguish the following prospective and retrospective components of the first method of payment:

$$\hat{P}^1_{i,h,t} = \underbrace{X'_{i,h,t}\hat{\gamma}_t + W'_{h,t}\hat{\alpha} + \hat{c}_t + \hat{a} + \underset{h,t}{Min}\,(\hat{\varepsilon}_{h,t} + \hat{u}_{.,h,t})}_{\text{Prospective}=\hat{F}_{i,h,t}} + \underbrace{\hat{\eta}_h}_{\text{Retrospective}} \quad . \tag{21}$$

Let us consider the classical expression of a mixed payment as a weighted average of a lump-sum F and the actual cost of treatment C : $P = \mu F + (1-\mu)C$. Using the expression (21), one can compute $\mu_{i,h,t} = \frac{\hat{P}^{1}_{i,h,t} - C_{i,h,t}}{\hat{F}_{i,h,t} - C_{i,h,t}}$.

We have obtained (for model A): $\overset{\wedge}{\mu} = 46,2\,\%$, with a standard error equal to 12,9 %. We have to underline that this sample mean ($\overset{\wedge}{\mu}$) provides an evaluation which is not a rule of payment. It results from an *ex post* computation, which allows us to know the weight of retrospective payment induced by the implementation of payment rule P^1.

7 Conclusion

Hospital heterogeneity is a major issue in defining an optimal reimbursement system.

In this paper, we have considered an extension of the basic yardstick competition model, allowing for the existence of observable sources of heterogeneity. We have applied an econometric approach to the identification and evaluation of observable and unobservable sources of cost heterogeneity. The use of a three dimensional nested database makes it possible to identify transitory moral hazard, and to estimate its effect on hospital cost variability.

In our specification, observable hospital characteristics and hospital specific effects enable us to take hospital heterogeneity into account. We obtain two alternative payment systems. The first takes all unobservable hospital heterogeneity into account, provided that it is time invariant, whereas the second ignores unobservable heterogeneity. Simulations show that substantial budget savings - at least 20 % - can be expected from the implementation of such payment rules.

The first method of payment seems advisable to us: it has the great advantage of reimbursing high quality care. It leads to substantial potential savings, because it provides incentives to reduce costs linked to transitory moral hazard, whose influence on cost variability is far from negligible. Thus, our study shows that: (i) one component of moral

hazard can be easily identified with three-level panel data: transitory moral hazard (ii) this component of moral hazard is sizeable. Therefore, substantial budget savings can be obtained from the implementation of a payment rule which eliminates only this component.

Moreover, this payment system is easy to implement, provided the regulator has information about costs of hospital stays. One drawback is that it would give higher reimbursements to hospitals which are costlier because of permanently inefficient management. The choice between the two methods of payment depends on the weights assigned to efficiency and care quality in the social objective function used by the regulator.

Our payment rules could be extended to other areas of health care financing. Considering again the example of AAPCC to calculate Medicare Managed Care reimbursements, our method would make it possible - not to identify the sources of geographical cost heterogeneity - but to identify the transitory moral hazard in the local-year dimension. If this component of moral hazard has a sizeable influence on cost variability, the savings derived from eliminating it can be substantial.

In order to induce effective budget savings, the implementation of our payment rules requires the following: hospitals would have to earn the rents arising from improved efficiency and they would have to face a hard budget constraint. What can be infered from our simulations is limited by the fact that they are carried out under the hypothesis that behaviors are supposed to remain unchanged except as concerns moral hazard. In other words, hospitals are supposed not to adopt strategic behaviors in reaction to a reform of the payment system. Moreover, our evaluations of the budget savings assume a constant level of activity. Our finding of a potential saving of 20 % means that greater efficiency could have induced a saving of 20 % to finance the hospital activity observed during the period 1994-1997. One important difference between the PPS and the global budget is that the level of activity is in principle not capped within the PPS. An increase in activity could make hospital expenditures rise, even if hospitals progressed in efficiency.

8 References

Apparitio, S., A.-M. Brocas and J.-C. Moisdon, 1999. La place du PSI dans l'allocation des ressources en île-de-France, Agence Régional d'Hospitalisation d'île-de-France - Rapport technique.

Antweiler, W., 2001. Nested random effects estimation in unbalanced panel data, Journal of Econometrics vol 101: pp 295-313

Baltagi, B. H., 1981. Simultaneous equations with error components, Journal of Econometric vol 17 : pp 189-200

Baltagi, B. H., S. H Song. and B. C. Jung, 2001. The unbalanced nested error component regression model, Journal of Econometrics, vol 101: pp 357-381

Chalkley, M. and J. M Malcomson, 2000. Government purchasing of health services, in: Culyer A.J. and Newhouse J.P. editors, Handbook of Health Economics, Vol. 1A (North Holland, Amsterdam), Chapter 15, 847-890.

Direction des Hôpitaux de Paris, mission PMSI, 1996. Le PMSI, analyse médico-économique de l'activité hospitalière, La lettre d'information hospitalières, special issue

Dormont, B. and C. Milcent, 2004. The sources of hospital cost variability. Health Economics, Vol. 13: 927-939

Ellis, R. P., 1998. Creaming, Dumping, skimping: Provider competition on the intensive and extensive margins, Journal of Health Economics, vol 17: pp 537-555

Greene, W. 2004. Distinguishing between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the WHO's Panel Data on National Health Care Systems, Health Economics, Vol 13, Issue 10: pp 959-980

Keeler E. B., 1990. What proportion of hospital cost differences is justifiable ?, Journal of Health Economics 9(3), 359-365.

Laffont J. J. and J. Tirole, 1993. A theory of incentives in procurement and regulation,

MIT Press

Linna M., 1998. Measuring hospital cost efficiency with panel data models, Health Economics, 7: pp 415-427.

Lopez-Casasnovas, G. and M. Saez, 1999. The impact of teaching Status on Average Costs in Spanish Hospitals, Health Economics, vol 8, n°7: pp 641-651

Ma, A. C. T., 1994. Health care payment systems: cost and quality incentives, Journal of Economics and Management Strategy, vol 3, n°1: pp 93-112

Ma, A. C. T., 1998. Health care payment systems: cost and quality incentives- Reply, Journal of Economics and Management Strategy, vol 7, n°1: pp 139-142

McClellan M., 1997. Hospital reimbursement incentives : an empirical analysis , Journal of Economics and Management Strategy, 6(1) : 91-128.

Mougeot M., 1999. Régulation du système de santé, CAE, La Documentation Française, Paris.

Newhouse J. P., 1996. Reimbursing health plans and health providers : efficiency in production versus selection , Journal of Economic Literature, Vol. XXXIV : 1236-1263.

Pope, G., 1990. Using hospital-specific costs to improve the fairness of prospective reimbursement, Journal of Health Economics, vol 9, n°3: pp 237-251

Rosko, M. 2001. Cost efficiency of US hospital: A Stochastic Frontier Approach, Health Economics, Vol 10: pp 539-551

Shleifer, A., 1985. A theory of Yardstick Competition, Rand Journal of Economics, vol 16: pp 319-327

Staiger D. and Stock JH., 1997. Instrumental variables regression with weak instruments. Econometrica, 65, pp 557-586

Wennberg, R., E Fisher, J. Skinner, 2002. Geography and The Debate Over Medicare Reform, Health Affairs. Available: http://content.healthaffairs.org/cgi/content/full/hlthaff.w2.96v1/D