Unofficial payments for acute state hospital care in Kazakhstan. A model of physician behaviour with price discrimination and vertical service differentiation.

Robin Thompson
Centre for Health Economics,
University of York, UK

Ana Xavier*
LICOS Centre for Transition Economics,
Katholieke Universiteit Leuven, Belgium

Autumn 2002

*Correponding author: *LICOS Centre for Transition Economics, Katholieke Universiteit Leuven, Deberiotstraat 34, 3000 Leuven, Belgium. Tel: +32 16 32 65 92. Fax: +32 16 32 65 99. Email: ana.xavier@econ.kuleuven.ac.be*

Abstract

In most of the countries in transition from a planned to a market economy (Former Soviet Union (FSU) and Central and Eastern Europe (CEE)) patients are routinely asked to pay unofficially for the medicines and other supplies that ought to be free. They are often described as "payments to individuals or institutions in cash or in kind made outside official payment channels for services that are meant to be covered by the public health care system". Despite their illegality, surveys undertaken in Bulgaria, Poland, Turkmenistan, and Tajikistan found that 43%, 46%, 50%, and 70% of the patients paid for officially free services.

We consider a simple model of discriminatory pricing and service differentiation in which state salaried physicians employed in a monopoly state acute hospital adjust the quality of care to the level of unofficial payment paid by the patient. On one hand, low motivated and poorly paid physicians exploit their monopoly position by choosing the payment / quality combination to be provided with the knowledge that corruption is largely ignored. There is a cost involved in the provision of each type of treatment (e.g. the potential fine imposed if found). On the other hand, the general quality of the health care provided by the sate is perceived to be poor and some patients are willing to pay unofficially in an attempt to improve the quality of care in some way. Patients have heterogeneous preferences for care quality. Physicians exploit their position as providers and the demand for quality, and offer differing levels of care quality to paying and non-paying patients.

This behavioural model is then tested using a unique dataset obtained from a survey of 1508 discharged hospital surgical and trauma patients treated in three hospitals in Almaty City, Kazakhstan in 1999. Data include information on patients' experience in hospital including where and how much was paid unofficially and patients' socio-economic characteristics. Each patient is identified by an ICD10 code and most surgical and trauma conditions for which entitlement is free are represented. We use waiting time for admissions and hospital length of stay (LOS) as observable measures of quality. Process indicators, such as "time" have traditionally been used to monitor hospital performance. Waiting is used as a quality proxy in studies of health service demand and surveys suggest that patients pay to reduce the wait. LOS is taken as a measure of quality in that it is determined by physicians and may proxy not only more attention paid to the patient by the doctor but also fewer post-surgical complications in a context where post-hospital follow up is very limited. We then compare these results with those using a categorically (ordered) variable reflecting the subjective view of the patients towards the health care quality received.

We use both OLS analysis and, in the case of the categorical variable, ordered probit models to test whether, as suggested by the theory, surveys and anedoctal reports, patients are paying for increased treatment quality. We account for heteroskedasticity, potential endogeneity, and test the general model specification. We look at both pooled and unpooled hospital data. We find that: patients are paying to decrease the time for surgical admissions; patients are paying to stay longer in hospital (or more realistically not to discharged early); there is some patient heterogeneity. These results conform to the theoretical modelling and the anedoctal reports suggesting that patients are paying for quality and physicians are exploiting their monopoly power to charge informal payments, which result in an increase in their incomes.

This paper contributes to the literature in various ways: a) as one of the first attempts to use economic theory to model unofficial payments and related physician behaviour; b) we use original data from Kazakhstan to conduct econometric analysis so as to explore whether prior payment influences the quality of care received (measured using process indicators) while previous studies were limited to answer the who, how much, when and to whom; c) the unofficial pricing behaviour of state salaried physicians in a public sector hospital may offer insights into the general behaviour of physicians.

Keywords: transition economies, unofficial or informal payments for health care, length of stay, ordered probit and marginal effects

JEL: I1, P3

1. Introduction

In many emerging market economies of the former Soviet Union (FSU) and Central and Eastern Europe (CEE) patients are routinely asked and expected to *unofficially* pay for medicines and medical supplies required for their medical treatment (World Bank, 2000a). These payments have been described as "payments to individuals or institutions in cash or in kind made outside official payment channels for services that are meant to be covered (*without direct charge*) by the public health care system"(Lewis, 2000). Their legality is either not clear or they are illegal. Recent surveys of patients undertaken in Bulgaria, Poland, and Turkmenistan found that 43%, 46%, and 50% respectively paid for services that were officially free (Delcheva *et al.*, 1997; Ladbury, 1997; Chawla *et al.*, 1998). In Tajikistan, 70% of survey respondents stated they expected to have to pay for health care (Mirzoev, 1999). Thompson and Witter (2000) and Ensor (2000) present typologies of these payments.

Unofficial payments are rooted in systems of bargaining and connections inherited from the socialist system (Smith, 1973). The planned and rigid nature of health care provision led patients to search for mechanisms to obtain faster and better services (*e.g.* more doctor's attention) than those they would obtain as the basic state services, (Gaal, 1999a,b; Kornai, 2000). These payments increased the financial reward to the medical activity, highly demanding but little recognised.

The widespread existence of unofficial payments for health care is also closely related to the impact of economic restructuring including the closure of state and private enterprises and increased unemployment, which resulted in a decline in tax revenue and subsequent reductions in government health sector funding. Government failure in addressing the scope and scale of service provision (downsizing services and reducing staff), as a result of resource constraints, has led to a gap in state resources necessary to fund the existing level of provision. Chronic shortages, coupled with inadequate equipment resulted in patients or their relatives being routinely asked to cover the shortfall in health care funding by paying, through unofficial channels, for medicines and other supplies required for their medical treatment, which are scarce due to tighter budget constraints.

Unofficial payments feature in countries where health workers salaries are low relative to other state and private sector professions, delays are commonplace - in Lithuania and Ukraine, workers were waiting up to three months to be paid, with reports of longer delays in Russia (Healy and McKee, 1998), and the private sector - which could provide extra income - is practically non-existent. The majority of unofficial payments go to physicians: *e.g.* in Poland 81% of payments were paid to physicians, with the rest being paid to other health workers (Chawla *et al.*, 1998). In Estonia, 60% of physicians reported receiving at least one non-cash gift each week and some

received a monetary tip (Barr, 1996). This monetary tip constituted an average amount of around 18.5% of their monthly salary. Unofficial payments double the average gross salary of physicians in Poland (Chawla *et al.*, 1998) and are five times the salary of Albanian specialist doctors (Healy and McKee, 1998), while in Hungary they constitute 62% of the net income of physicians (Kornai, 2000). In the Czech Republic over 27% of patients gave gifts to obtain better treatment and 7% paid out of fear of receiving no treatment (Masopust, 1989). The presence of widespread corruption, weak monitoring, and minimal sanctions, and their enforcement, for those who are caught taking such payments, fuels unofficial payments. Lack of information and non-reporting by patients and physicians' lack of accountability to a higher authority help to maintain the system.

More generally, unofficial payments can be viewed as an attempt to improve service quality receive in chronically underfunded state facilities (Thompson and Witter, 2000; Lewis *et al.*, 2001). These improvements are wide-ranging and may include, for example, more effective medicines than those offered without charge by the state, minimally invasive surgical technologies rather than conventional surgery, or simply more "effort" undertaken by the physician. Anecdotal reports suggest that motivating physician "effort" is one of the key reasons for payment (Thompson and Witter, 2000). Field (1998) suggests that unofficial payments "constitute a countervailing power at the disposal of the patient to exert some kind of control over the physician". Lewis *et al.* (2001) find that patients paid to save time and Gaal (1999b) argues that payments are made so as to change providers' attitude towards the patient and adapt treatment to patient's convenience. For Kornai (2000) these payments are a "bribe" paid by patients to doctors to ensure extra-attention, "jump" the queue thus obtaining a shorter period of waiting, obtain a better bed or a chosen doctor.

Unofficial payments are likely to be inequitable because patients' access to services or quality of care depends on their ability to pay. In Bulgaria payments made to health workers represented 3 to 14% of a patient's average monthly income with the cost of a surgical procedure around 83% of the average income (Delcheva *et al.*, 1997). Bognar *et al.* (2000) found that payments were income independent. Thus, it is perhaps not surprising that, in Kazakhstan, relatives of patients often advertise in newspapers for monetary support and that the Russia Longitudinal Monitoring Survey finds "lack of money" as the main reason for the inability to obtain medicines cited twice as often in 1996 as in 1994 (Liu *et al.*, 1998). Some people may be delaying care and avoiding the health sector all together (Lewis, 2001). Unofficial payments may also undermine investment in equipment and facilities because they are channelled to individuals not to the system.

However, these payments play an important role in sustaining health care systems in many countries where, despite government efforts, public revenues generated officially have been limited (World Bank, 2000b). They represent a significant slice of the total health care expenditure and a

financial supplement for health workers whose wages were kept low before and after transition started (Kornai, 2000). In Kazakhstan, in 1996, they constituted 25-30% of the state budget considering medicines alone and patients may pay around US$50 for inpatient medicines (Ensor and Savelyeva, 1998; Thompson and Witter, 2000). The purchasing of drugs is indeed a common source of unofficial expenditure although inpatient care is the most costly item (Lewis, 2001).

Much of the unofficial payments literature has focused on differing types of unofficial payment and the contribution these payments make to total health care spending (*e.g.* Thompson and Witter, 2000; Lewis, 2001). According to Lewis (2001), "a greater understanding is important if abuse of the system is to be addressed and resolved".

We therefore develop and test an economic model of physician behaviour in which state salaried physicians adjust the quality of care to the level of unofficial payment paid by the patient in a monopoly state acute hospital setting, offering differing levels of service quality to paying and non-paying patients. The model is motivated by the perception that the general quality of state health care provision is poor and some patients are willing to pay unofficially in an attempt to improve care quality. Demand is the result of patients' perceptions of quality and associated preferences for the low and basic quality services, and the prices charged. On the supply side poorly paid and demotivated physicians exploit their monopoly position and better information concerning care by engaging in discriminatory pricing and service differentiation, doing so with the knowledge that corruption is largely ignored by the state. The physician chooses the payments to charge for the low and the high quality services given the patients' demands (*i.e.* the physician chooses the quality - payment combination). The theory is then tested using data from a survey of 1508 discharged acute hospital (surgery and trauma) patients in Kazakhstan.

This paper contributes to the existing literature in various ways. Firstly, it is one of the first attempts to use both economic theory and econometric tools to analyse the issue of unofficial payments and explore whether prior payment influences the quality of care received (measured using process and subjective indicators of quality). While previous studies were limited to answer the "whom, how much, when and to whom" of the matter and tend to provide anecdotal reports of physician behaviour with estimates of spending collected through primary surveys, we use a discriminatory price differentiated service model to formalise the unofficial market for quality of care and both ordinary least squares and ordered probit analysis to test it. There are few, if any published English language studies, which formally attempt to model and test physician behaviour within an unofficial payments context despite the number of anecdotal reports suggesting that physicians may indeed be exploiting their monopoly position engaging in discriminatory unofficial pricing and service differentiation. Secondly, we use detailed data on discharged acute hospital

patients that in general are very difficult to obtain. The data are gathered for patients who had an intervention that was officially free of charge, fact that allow us to identify clearly the amount paid unofficially. Some of the previous studies did/could not distinguish between official and unofficial payments for care thus providing only a rough idea of what the latter might be. Finally, the unofficial pricing behaviour of state salaried physicians working in the hospital sector might offer some insights into the behaviour of physicians working in formal health care systems elsewhere.

The paper is organised as follows: Section 2 describes the Kazak health care system. Section 3 develops a model of physician and patients' behaviour in a context of an unofficial market for health care quality. Section 4 describes the data and methods used to test the model. Section 5 presents the results of OLS regressions. Section 5 discusses the results and concludes.

2. The health care system in Kazakhstan

Before independence, the Ministry of Health in Kazakhstan administered policy made in Moscow through a centrally organised hierarchical structure, from the republic level to the *oblast* or city administrations, then to the subordinate rayon level. The Kazak health care system featured most of the usual characteristics of a Soviet health care system (see Ryan, 1978 for a detailed description of the organisation of Soviet health care): services were, in principle, accessible and mostly free to everyone; funding was based on capacity rather than activity; over emphasis was given to specialist training and there was a dependence on hospitalisation, with long lengths of stay; and incentives focused on penalties for failure rather than incentives for success (Ensor and Rittmann, 1997). The weaknesses of the Soviet health care system have been well documented (European Observatory, 1999). Since independence they have been exacerbated by declining health sector spending, a product of deep economic recession. National income halved between 1991 and 1995, while government revenue fell by more than 70% (World Bank, 1997). The acute funding crisis and over-emphasis on inpatient care resulted in resources being extremely thinly spread.

Kazakhstan began the 1990s with a government funded, tax-based, health care system. A mandatory health insurance system was established in 1996 and dissolved in 1998, largely due to enterprises being unable to pay contributions to the fund, a large informal workforce, inability of the regional administrations to cover the socially protected population, particularly the growing unemployed, and a collapse in the confidence in the fund with allegations of corruption and misappropriation of reserve funds. Health care now comes from two main sources (similar position to pre-insurance funding): the government budget and out-of-pocket payments (official and unofficial). A 1994 survey of 5000 households in South Kazakhstan found that informal payments were common for both outpatient and inpatient care. On an inpatient basis, the subject of this

paper, payment was made to providers 11% of the time and 12% to surgeons. In addition, 25-42% of those hospitalised provided their own bedding, clean laundry and food, and 57% provided their own medicines (Sari *et al.*, 2000). A decree formalising user charges was introduced in 1999 (European Observatory, 1999). The ability of a significant proportion of the population to pay for health care is limited; a living standards survey undertaken in 1996 found that over a third of the population lived below a "subsistence minimum" living standard (World Bank, 1998).

Whilst entitlement to comprehensive health care was a feature of the pre-independence system, in recent years entitlement benefits have become confusing. This has partly been the result of the insurance experiment where services were separated into two "packages": basic (provided by insurance) and guaranteed (paid for by the state). Confusion is enhanced by shortages relating to chronic underfunding and health sector corruption. In principle primary health care consultations are free, although medicines are not free for the non-exempt. Yet, even the exemption system does not function well and many individuals have to pay for medicines that should be free. Hospital benefit entitlement is particularly confusing and whether a patient pays depends on whether an illness is acute/not acute, resource availability, and health worker corruption. For example, individuals requiring elective surgery are increasingly required to pay whereas those who are admitted as acute/emergency patients are, again in principle, exempt from payment. However, as the empirical results in this paper show, in reality the vast majority of patients pay for hospital care.

The health care system is dominated by hospital care and the number of days a patient spends in hospital appears to be quite important. In countries like the UK post-hospital follow-up care is increasingly important as the length of hospital stay is reduced. In Kazakhstan, however, post-hospital follow-up care is weak and anecdotal reports suggest that patients are willing to pay to stay in hospital for reassurance, as once they leave hospital follow-up care is non-existent.

3. Physician and patient's behaviour: an unofficial health care market

In this section we model patients' and physicians' behaviour looking at the parallel and unofficial market for health care within a monopoly state provider. Given the apathetic attitude of government towards corruption in some of the countries of the FSU and CEE, state salaried physicians might well adopt patterns of market behaviour within state hospitals and explore an element of monopoly power thus creating an unofficial market for health care.

On one side of this market we have the patients for whom the general quality of state health care provision is perceived to be poor. Consequently, some patients are willing to pay unofficially for services (*e.g.* medicines, surgeon's time) that are free so as to improve the quality of care received. Patients have different preferences for quality, resulting in the demand for quality of care

being a function of payment. On the other side of this unofficial market we have the state salaried physicians employed in a monopoly state acute hospital setting.[1] Often unmotivated and poorly paid, they adjust the quality of care to the level of unofficial payment paid for by the patients, given the preferences of the latter. As said, health workers have a strong monopoly power over medical knowledge (diagnostic and treatment) and patients' discharge, which they can exploit to obtain unofficial payments without a significant cost to them (*e.g.* no extra working time and sanctions are weak). They can allocate scarce state medicines and medical supplies to patients who pay unofficially, keep patients in hospital or discharge them early (a significant power due to the lack of follow-up service provision outside hospital). Hence, physicians are seen as profit/income maximisers choosing the payment -quality combination given patients' demand for their services. They exploit their monopoly position by engaging in discriminatory pricing and service differentiation doing so with the knowledge that corruption is largely ignored by the state. There are many anecdotal reports of FSU and CEE state physicians adopting differential unofficial pricing strategies and considerable evidence suggesting that patients are willing to pay unofficially for an improvement in the quality of health care (Thompson and Witter 2000).

An important issue is whether unofficial payments are made for entitled services or some enhanced level of care. Patients may be asked to supply medicines and supplies required for their treatment because the hospitals do not have these. Or patients may be asked to purchase medicines and supplies that are available and paid for through the state budget but often with a delay, which patients may not wish/be able to bear. Or a corrupt health worker may simply ask a patient for a payment to ensure access to a basic level of service and/or imply that payment is linked to higher quality care. The patient accessing acute care is unlikely to know, or be in a position to question, whether the care is in fact some enhanced level or the entitled level. However, whatever the reason, if she perceives that no payment leads to a sub-desirable care, she may be willing to pay.

Information asymmetry coupled with endemic unofficial payments places acute hospital patients in an extremely vulnerable position. While patients do not know what services should be provided as part of their entitlement to state health care, physicians are fully aware of this entitlement: medical standards define the scope of services to be provided for each diagnostic category, include the scope and scale of diagnostic tests, medicines, and medical supplies and define how many days that a patient should stay in hospital. The health worker can exploit his knowledge to obtain unofficial payments, allocating state resources to those who pay unofficially and discharging accordingly.

[1] We only examine the decision making process of physicians not of the hospital as a whole or its management team.

3.1. A short review of the related literature

The physician agency literature can provide some useful insights into the behaviour and motivation of state salaried physicians employed in the transition world. Whilst the literature is predominantly written within the North American context, a number of parallels can be established with salaried state physicians working within endemic unofficial payment systems of the FSU and CEE when they adopt patterns of market behaviour within state hospitals and explore some monopoly power.

Reviewing the literature McGuire (2000) argues that there are not many alternatives to a profit maximising model subject to a demand. Many papers present no formal conception or model of the behaviour of the physician firm, while others looked at physicians as profit maximisers setting prices for their services. An element of monopoly power, which is explored by the physician in the context of complete information, is present in most models. Several authors (*e.g.* Gaynor and Gertler, 1995; Ma and McGuire, 1997; Phelps, 1997; Dranove and Satterthwaite, 2000) analyse location, specialty, and care quality as elements that turn physicians into imperfect substitutes and, as such, there is an element of monopoly power with the demand curve sloping downwards. McGuire (2000) presents a model of monopolistic competition where the price and quantity of physician services are found by maximising the physician's profit, subject to the constraint on patient net benefit. Patients face an all or nothing offer and all available consumer surplus is extracted. With market power and the non-retradability of healthcare, the physician possesses the prerequisites for the exercise of first-degree price discrimination. Hence, we believe the profit maximising assumption is useful to analyse a context where patients are willing to pay unofficially for extra quality and physicians are willing to exploit their power to provide it.[2]

The literature on general discriminatory pricing (outside the health care sector) is large (Tirole, 1988; Varian 1987) and it is well understood that non-retradability is behind models of this nature. Gaynor (1994) and Folland, Goodman, and Stano (1997) recognise that physician services are heterogeneous and non-retradable thus supporting price discrimination. Focusing on the physician's self interest, Kessel (1958), suggests that differences in physician fees could be explained by differences in demand. Ruffin (1973) describes a "charity-competition" model in which price discrimination emerges as result of doctor utility maximisation. Feldstein (1979) uses a monopoly model to analyse physician's pricing behaviour. The insights of these models are our departing point. We believe this literature fits well with the context under analysis.[3]

[2] One may wish to add others arguments to the physician objective function but we wish to concentrate on this particular aspect of physician behaviour.

[3] The literature on corruption may also provide useful insights concerning unofficial payments (see Bardhan, 1997 for a review). Lui (1985) presented an equilibrium queuing model of bribery where customers pay bribes in order to obtain a better position in the queue. The size of the bribe was linked to the opportunity costs of time for the individual. Myrdal

3.2. Setting the quality - unofficial payment combination

Consider now, in the context of the parallel and unofficial market for health care described before, the demand for two competing health care interventions or processes used to treat the same condition but differing in some quality characteristic. The good being traded is treatment and each consumer consumes one unit of the good, that is, a patient consumes care only once at a time (*e.g.* one operation only). According to some measure the treatment can have two different quality levels or indeed be considered as two treatments - the low quality and the high quality treatment. For example, a patient may be given two choices of surgery: low quality (*e.g.* basic/conventional surgery) or high quality surgery (*e.g.* cholycystectomy). The hospital physician might not officially be permitted to use this technology for the treatment but has unofficial access to it. Alternatively, quality may be measured by some physician input such as time (or "effort") devoted to the patient so that the low quality treatment corresponds to basic consultation time and high quality treatment means additional doctor's time. Another possible definition of treatment quality in the transition context might be that where an acute surgical patient is given a choice of post-operative care, implied by two differing lengths of stay proposed by the operating surgeon. The patient may not know what specific interventions will be administered post-operatively, however longer length of stay may associated with increased patient's utility because of the reassurance of knowing that if any problems occur the physician will be on hand to address them. Shorter lengths of stay for the acute surgical patient would in this context create disutility because of perceived inadequate follow-up on discharge.[4] Shorter lengths of stay in this context therefore might be recognised as some basic or low level of health care quality with longer length of stay perceived as an enhanced or high level of quality. Also time spent waiting before admission may be perceived as a quality measure: the longer the wait the lower the quality of care according to patients (*e.g.* Propper, 2000).

The indirect utility (measured in monetary terms) each patient derives from treatment depends on the price she pays and on the quality obtained given her taste parameter (Tirole, 1988):

$$U^P = \begin{cases} \theta\varphi - p & \text{if consumer pays } p \text{ and consumes quality } \varphi \\ 0 \end{cases} \qquad (1)$$

(1968) argued that corrupt officials might, instead of speeding up queues, actually cause administrative delays in order to attract more bribes, although Bardhan (1997) suggests that, in the context of pervasive and cumbersome regulations, corruption may actually improve efficiency. Galasi and Kertesi (1989) modelled bribes for quality in socialist countries and showed that all consumers were worse off when some of them paid bribes to obtain higher than official quality care. With fixed inputs, bribery reduces the quality available to those paying fixed or no price and induces more corruption, resulting in everybody paying bribes yet obtaining quality no higher than the official level (see also Kornai, 2000).

[4] There are a number of reports to suggest this is the case, particularly in rural areas (Ensor and Thompson, 1999)

where φ is a positive parameter describing quality, with $\varphi=(\varphi_L,\varphi_H)$ and $\varphi_L<\varphi_H$, and L and H referring respectively to the low quality and to high quality treatments as perceived by patients[5]. θ is a positive real number that describes the taste for quality[6] and p is the price of treatment with $p=(p_L,p_H)$ and $p_L < p_H$, that is, the low quality treatment is charged a lower price (if it were more expensive then no one would buy it). The utility obtained with treatment is separable in price and quality, the rationale behind it being that all consumers prefer a higher quality for a given price but a consumer attributing a high value to quality is willing to pay more to obtain higher quality of care. The parameter θ is distributed according to some density function, $f(\theta)$, reflecting the variation in tastes among patients, and a cumulative distribution function $F(\theta)$ defined between zero and a maximum value of $\theta=\theta^M$, $[0, \theta^M]$), with $F(0)=0$ and $F(\theta^M)=1$. The utility of no treatment is zero.

A patient chooses the high quality treatment rather than the low quality one if the utility obtained with the former is higher than that obtaining with the latter and higher than no treatment:

$$U_H^P \geq U_L^P > 0 \;\Rightarrow\; \theta\varphi_H - p_H \geq \theta\varphi_L - p_L > 0 \;\Rightarrow\; \theta^{Hc} \geq \frac{p_H - p_L}{\varphi_H - \varphi_L} > 0 \qquad (2)$$

Patients choose the low quality treatment whenever the utility associated with it is higher than no treatment

$$U_L^P > 0 \;\Rightarrow\; \theta\varphi_L - p_L > 0 \;\Rightarrow\; \theta^{Lc} \geq \frac{p_L}{\varphi_L} > 0 \qquad (3)$$

Thus,

- All those, who have preferences for quality higher than threshold θ^{Hc}, buy the high quality care *e.g.* longer length of stay (basic) or shorter admission time.
- All the patients, who have preferences for quality higher than threshold θ^{Lc}, buy the low quality treatment *e.g.* shorter length of stay (early discharge) or longer admission time.
- All the other patients for whom the threshold $\theta < \theta^{Lc}$ are excluded from care.

Given N potential patients whose preferences for quality vary according to the density function above, a proportion of these will buy the high quality treatment, another will buy the lower quality treatment and some will buy no care. Integrating the density function using the boundaries defined by the above critical levels of the taste parameter, θ, we obtain the demand for high quality and low quality treatment and the demand for no treatment as functions of the unofficial payment:

[5] For analytical simplicity and illustrative purposes we use only two discrete levels of quality. This context can be easily extended to one of a continuous range of quality levels.

[6] It can also be seen as the inverse of the marginal rate of substitution between income and quality. In that case $f(\theta)$ may be related to the distribution of income among the potential consumers of the good treatment. Assuming the case of income, this means that all consumers derive the same surplus from the treatment (in our case they get cured) but some consumers, the wealthier, have a lower marginal utility of income and thus a higher θ.

$$D_H = D(N, p_L, \varphi_L, p_H, \varphi_H) = N \int_{\theta^{Hc}}^{\theta^M} f(s)ds = N\left[1 - F\left(\frac{p_H - p_L}{\varphi_H - \varphi_L}\right)\right] \quad (4)$$

$$D_L = D(N, p_L, \varphi_L, p_H, \varphi_H) = N \int_{\theta^{Lc}}^{\theta^{Hc}} f(s)ds = N\left[F\left(\frac{p_H - p_L}{\varphi_H - \varphi_L}\right) - F\left(\frac{p_L}{\varphi_L}\right)\right] \quad (5)$$

$$D_{NC} = D(N, p_L, \varphi_L) = N \int_{0}^{\theta^{Lc}} f(s)ds = N\left[F\left(\frac{p_L}{\varphi_L}\right)\right] \quad (6)$$

Assume now that the physician knows that the demand for his services is composed of heterogeneous consumers some of those with stronger preferences for the higher quality good.[7] The physician chooses the unofficial payments p_L and p_H to maximise his utility knowing the above demand functions. We assume there may be a possible cost involved in the unofficial provision of treatment for example the potential sanction imposed on the physician if found to be charging unofficial payments, which is separable and linear. The doctor's maximisation problem is

$$\max_{p_H, p_L} U^D = (p_L - c_L)D_L(p_L, p_H, ..) + (p_H - c_H)D_H(p_L, p_H, ..) \quad (7)$$

and the first order conditions are, with $i,j=L,H$ and $i \neq j$:

$$\frac{\partial U^D}{\partial p_i} = D_i(p_i, p_j, ..) + (p_i - c_i)\frac{\partial D_i(p_i, p_j, ..)}{\partial p_i} + (p_j - c_j)\frac{\partial D_j(p_i, p_j, ..)}{\partial p_j} = 0$$

Rearranging the terms we have that, with $i,j=L,H$ and $i \neq j$:

$$D_i(p_i, p_j, ..) + p_i \frac{\partial D_i(p_i, p_j, ..)}{\partial p_i} + p_j \frac{\partial D_j(p_i, p_j, ..)}{\partial p_i} = c_i \frac{\partial D_i(p_i, p_j, ..)}{\partial p_i} - c_j \frac{\partial D_j(p_i, p_j, ..)}{\partial p_i}$$

This shows that the physician as a monopolist chooses the unofficial payments so that marginal revenue equals marginal cost. Further rearranging (see Appendix 1) gives us, with $i,j=L,H$ and $i \neq j$:

$$\frac{(p_i - c_i)}{p_i} = \frac{1}{\varepsilon_{ii}^D} - \frac{(p_j - c_j)D_j \varepsilon_{ji}^D}{\varepsilon_{ii}^D R_i} \quad (8)$$

with ε^D_{ii} and ε^D_{ji} the direct and cross demand elasticities of treatment which are, when goods are substitutes:

$$\varepsilon_{ii}^D = -\frac{p_i}{D_i(p_i, p_j, ..)}\frac{\partial D_i(p_i, p_j, ..)}{\partial p_i} \quad and \quad \frac{\partial D_i(p_i, p_j, ..)}{\partial p_i} < 0$$

$$\varepsilon_{ji}^D = -\frac{p_i}{D_j(p_i, p_j, ..)}\frac{\partial D_j(p_i, p_j, ..)}{\partial p_i} \quad and \quad \frac{\partial D_j(p_i, p_j, ..)}{\partial p_i} > 0$$

Relationship (8) shows that the mark-up price of treatment i the term on the left-hand side of the equation is a function of: 1) the inverse of the own elasticity of demand $1/\varepsilon^D_{ii}$, which is positive

by our definition; and 2) the cross-elasticity and the mark-up for the other good. As the treatments are substitutes (patients can have only one type of treatment), the cross-elasticity, ε^{D}_{ji}, is negative according to the definition. Thus, *the mark-up price for good i is greater than just the inverse of the own elasticity of demand. It so appears that quality discrimination makes all patients pay a higher price for care*. Both Galasi and Kertesi (1989) and Kornai (2000) reached a similar conclusion.

As our aim is to test the relationship between quality of care and payment, we assume that patients' preferences over quality are distributed uniformly, that is, θ follows a uniform distribution. Thus we have $F(\theta^M)=1;\ F(0)=0;\ F(\theta^{Lc})=\frac{\theta^{Lc}-0}{\theta^M-0}=\frac{\theta^{Lc}}{\theta^M};\ F(\theta^{Hc})=\frac{\theta^{Ec}-0}{\theta^M-0}=\frac{\theta^{Hc}}{\theta^M}$ and the demands simplify to

$$D_H = D(N, p_L, \varphi_L, p_H, \varphi_H) = N\left[F(\theta^M) - F(\theta^{Hc})\right] = N\left[1-\frac{\theta^{Hc}}{\theta^M}\right] = N\left[\frac{\theta^M - \left(\frac{p_H - p_L}{\varphi_H - \varphi_L}\right)}{\theta^M}\right]$$

$$D_L = D(N, p_L, \varphi_L, p_H, \varphi_H) = N\left[F(\theta^{Hc}) - F(\theta^{Lc})\right] = N\left[\frac{\theta^{Hc}-\theta^{Lc}}{\theta^M}\right] = \frac{N}{\theta^M}\left(\frac{p_H - p_L}{\varphi_H - \varphi_L} - \frac{p_L}{\varphi_L}\right)$$

while the first order conditions are:

$$\frac{\partial U^D}{\partial p_i} = \frac{N}{\theta^M}\left(\frac{p_j - p_i}{\varphi_j - \varphi_i} - \frac{p_i}{\varphi_i}\right) + (p_i - c_i)\left(-\frac{N}{\theta^M(\varphi_j - \varphi_i)} - \frac{N}{\theta^M \varphi_i}\right) + (p_j - c_j)\left[\frac{N}{\theta^M(\varphi_j - \varphi_i)}\right] = 0 \quad \text{(A)}$$

Solving the pair of equations defined by (A) with respect to the unofficial payments p_H and p_L we obtain the optimal values of p_L and p_H:

$$\begin{cases} p_L = \dfrac{\theta^M \varphi_L + c_L}{2} \\ \\ p_H = \dfrac{\theta^M \varphi_H + c_H}{2} \end{cases} \quad \text{(9) which rearranging is equivalent to} \quad \begin{cases} \varphi_L = \dfrac{2p_L - c_L}{\theta^M} \\ \\ \varphi_H = \dfrac{2p_H - c_H}{\theta^M} \end{cases} \quad \text{(9a)}$$

This relationship is the basis for the empirical analysis that follows.[8] Equation (9a) shows that the quality dimension is positively related to the unofficial payment and negatively related to the cost of providing care quality unofficially. To a higher care quality corresponds a higher price and cost. If the results that follow show a positive association between payment and quality then the above

[7] Alternatively, some consumers have different marginal rates of substitution between income and quality of treatment.
[8] We chose to have quality on the left hand side and price on the right hand side because our data suggests that patients pay before entering hospital and receiving treatment so that relationship (9a) reflects better the reality of paying unofficially for care in transition countries. Moreover, to establish whether a higher payment leads to higher quality (which implies a direction of causality) we must estimate relationship (9a).

model may be a good representation of patients and physicians' behaviour in what concerns unofficial payments.

4. Data

The data used in this analysis come from a randomly selected survey of 1508 discharged surgical and trauma inpatient patients treated in three hospitals in Almaty City, Kazakhstan, in 1999. The survey was conducted in January 2002 with a maximum of nine months elapsing between discharge and interview. Given the sensitivity of the survey (unofficial payments are part of an unofficial market for care) patients were surveyed in their homes. The questionnaire was conducted in the Russian language with the help of the staff from the School of Public Health, Kazak State Medical Univeristy. Each of the 1508 patients included in the analysis is identified by an ICD10 code. Thirty-seven codes were included in the survey representing the most common surgical and trauma conditions treated in each of the departments. They were also chosen because individuals suffering from one of these codes were entitled to free care and thus all that was paid in hospital constituted an unofficial payment. The ICD10 codes were aggregated into four crude resource groups (RG1-4) based on information on resource use provided by the Almaty City Health Administration. Patients were surveyed about their experience in hospital and related expenditure. They were asked if they, or their relatives/friends on their behalf, had paid (monetarily or non-monetarily) and the amount paid in the admission department (AD), in the surgical/trauma ward, for medicines, and diagnostic tests.[9] Patients stated how many minutes they had spent in the AD, the number of nights they had spent in hospital and on the general quality of treatment. These three variables were chosen as indicators of quality of care so as to explore the relationship between the quality of care and unofficial payments. Information was obtained on patients' socio-economic status: age, gender, education, occupation, exemption status, and household expenditure (on food, utilities, clothes, cigarettes and alcohol, cars, education, health care and pharmaceuticals, family celebrations and support to relatives) as a proxy for household income. We also gathered information on the referral type (self-referral, polyclinic doctor or specialist, hospital specialist) and on whether the patient had surgery.

4.1. Dependent variable: quality of care

Measuring quality in health care is generally a complex business. Arrow (1963) recognised nearly thirty years ago that "uncertainty as to the quality of the product is perhaps more intense than for any other important commodity". One indicator of quality is the health worker's "effort" and it is

argued that unofficial payments are given to "motivate" physicians to provide more "effort". McGuire (2000) indeed suggests that "the care or effort that a doctor puts into a decision or treatment matters to the patient but it is difficult to incorporate into a payment system" as it is not directly observable. Thus it becomes one of the most difficult quality indicators, leading to contractability issues. The discussion on how to measure quality of care is ongoing (for example Campbell *et al.* (2000) suggest access to care and effectiveness as two main dimensions of quality of care) but the measurement of quality of care is however becoming more common through an increasingly available array of indicators. These are often divided into three groups: outcome measures (*e.g.* mortality rates), process or volume measures (*e.g.* day surgery, waiting time, length of stay), and patient satisfaction, in an attempt to capture some aspect of quality due to the absence of a correct, complete, and tangible measure of quality. The problems with measuring quality and contracting on outcomes (for these depend on patient characteristics *e.g.* genetics and the technological process of care) have meant that process indicators such as time, more tangible and more easily measured, have traditionally been used to measure quality and monitor and pay providers. For example McCall (1996), states that "the amount of time a doctor spends interviewing and examining you and explaining things reflects how genuinely concerned a doctor is". Also, lengths of stay and waiting times for inpatient and outpatient appointments are typically used to monitor hospital performance. Thus, in order to explore the relationship between unofficial payments and the process of care, we make use of two process measures of health care quality - 1) the waiting time (number of minutes) spent by the patient in the AD and 2) a patient's length of stay - LOS - or number of days spent in hospital, and complement these with a variable that measures patient's subjective measure of care quality. They are explained below.

4.1.1 Time spent in the admission department

As said, time measures have often been use as process indicators of quality because they are more easily measured. A number of authors model the demand for inpatient care (*e.g.* Goddard *et al.*, 1995; Martin and Smith, 1999; Gravelle *et al.*, 2003) examining the impact of waiting time. Waiting time is frequently used as a quality proxy in studies of health service demand (*e.g.* Propper, 2000 where individuals vary in their valuation of this quality parameter) and long waits have been seen by the general population as an unsatisfactory characteristic of the NHS (Bosanquet, 1988). Such models provide useful insights to the analysis of unofficial payments for quality in a transition country such as Kazakhstan, where state health care workers engage in quality enhancing activities within the state hospital structure and patients entering hospital chose differing quality services.

[9] We find hardly any variation in these latter two categories *i.e.* everybody appears to be paying for drugs and tests.

Moreover, previous studies of unofficial payments (*e.g.* Gaal, 1999a,b; Kornai, 2000; Lewis *et al.*, 2001) suggest that patients use such payments to obtain faster services than they would otherwise and jump the queue and face a shorter wait or to save time. In other words patients are paying an extra fee for immediate referral. Waiting could thus be seen as a measure of health care quality, and the shorter the wait the higher the quality of care.

Table 1 (appendix 2) shows, by hospital and resource group, the first indicator of quality: the number of minutes that patients spend in the admission department (AD) prior to a clinical intervention. The average period of time a patient spends in the department is approximately 55 minutes with small variations between hospitals and within resource groups. A further glance at Table 1 shows that a large number of patients fall into resource groups 2 and 3. There are few patients coded as group 1 in hospital 3 (trauma). As the time for admission is positively skewed a log transformation was performed on the variable (**lnadmwait**). Furthermore, the variable was standardised by ICD10 code. A negative coefficient for the unofficial payment is to be expected if a longer waiting time is seen as the inverse of a higher care quality.

4.1.2. Length of hospital stay (LOS)

The empirical analysis discussed in this paper also focuses on length of stay (LOS) as an indicator of health care quality. Variations in LOS may point to differences in the quality of health care provision although we may need to distinguish between the developed world and that of transition. In OECD countries, Barnum and Kutzin (1993) argue, longer stays do not necessarily contribute to higher-quality care (although patients may not necessarily perceive it to be so): LOS for most conditions has decline during the last thirty years in most OECD countries and the health of the population has not declined. Improvements in the technical quality of hospital care and most importantly a much wider availability of community care and local facilities to provide follow up care have made this possible[10] although concerns are sometimes raised about early discharge, post-surgical complications, and hospital readmission. In the transition world the situation is quite different. Health facilities are limited in number and often located in cities far from an important part of the population. Post-hospital follow-up is poor or non-existent and transport to hospital is limited and costly, especially from remote areas. Quality of care has regressed with the transition process and the consequent economic crisis. In this context, a longer stay in hospital increases patients' reassurance and decreases the probability of post-treatment complications and

[10] In several OECD countries and for reimbursement purposes, LOS has been used as a proxy for resource use and technical inefficiency.

readmission, as doctors monitor the patient condition for longer. Hence, a longer LOS may be perceived as better quality of care by patients in the transition world.

Table 2 (in appendix 2) shows length of hospital stay by hospital and resource group, the mean of which is approximately 14 days. There are large differences between hospitals 1 and 2 (surgical) and hospital 3 (trauma). In hospitals 1 and 2 length of stay is under 10 days where as in hospital 3 length of stay is over 20 days. As one might expect there are also differences in length of stay between resource groups. Once again the data is positively skewed so that a log transformation was performed (**lnlos**). Furthermore, the variable was standardised by ICD10 code. A positive sign is thus expected for the coefficient estimate associated with payment if a longer LOS proxies a higher (perceived) quality of care.

4.1.3. Subjective measure of the quality of care

Patients were also asked to characterise the quality of care received using an ordered categorical variable ranging from "Very poor" and "Poor" to "Satisfactory", "Good", and "Very good" quality of care. Patients had to choose one of the five categories. We believe this subjective measure of the quality of care received can complement the analysis using the process variables just discussed. It reflects patients' perception of the level of quality received, perhaps proxying patients' satisfaction, and allows us to use an extra non-process quality indicator. Moreover, if the empirical results are consistent across models using process or subjective/categorical variables then the analysis can be deemed more robust. Table 3 (appendix 2) shows the distribution of patients' responses to the questionnaire. We also sum the two last categories, "Poor" and "Very poor", into one category ("Poor plus very poor" - in italic) which results in a smoother distribution of responses across categories, from "Poor plus very poor" to "Satisfactory", "Good" and "Very good".

Overall, more than 55% of the patients considered the quality of care received to be "Good" with around 13% considered it to be "Very good". About 22.3% of the patients thought the quality of care was "Satisfactory" while 6.2% thought of it as "Poor" and "1.5% "Very Poor" (with almost 8% perceiving quality of care to be poor or very poor). The same sort of distribution can be seen across hospitals with the "Good" category registering the highest percentage of answers (always more than 50%) followed by "Satisfactory" (between 19% and 25%) and "Very good" (between 11,3% and 14%). "Poor" and "Very poor" vary between 4% and 8.8% and 1% to 2% respectively. Hospital 3 (Trauma) registers the best scores in that only 5% considered care to be of a poor or very poor quality (and more than 60% considered it to be good) while for hospital 2 this percentage is around 11% (and for hospital 1 is 6.6%). In the regression analysis and given the smallest percentage of responses associated with category "Very poor" we use the pooled "poor plus very

poor" category. We expect that payment is positively associated with a better quality of care while negatively associated with a worst category of care.

4.2. Independent Variables

As we focus on the relationship between the quality of care obtained and the unofficial payment made we gathered information on whether, how much, and where patients paid. The general idea developed through the interviewing process was that payment negotiation takes place as soon as the patient arrives to the hospital in the AD and before treatment takes place (*e.g.* patients seek to reduce admission time by paying) with patients agreeing to a certain amount for a certain quality level. However, although negotiation and agreement take place in the AD and before treatment, some patients do not pay all at once in the AD (*e.g.* because they cannot afford) and some pay after admission takes place while in the ward. As a result, given the information gathered with the questionnaire, we consider two unofficial payments variables: 1) Payment1 the amount of unofficial payment made by the individual before treatment takes place and in the AD; and 2) Payment2, the amount of unofficial payment made after admission takes place when already in the ward. Both are in their logs due to their skewed distribution. We also consider payment-hospital interactions.

In order to isolate the association between the quality of care and payment it is important to understand and control for other factors. Martin and Smith (1996) conclude that LOS is related to patient characteristics and hospital characteristics. Studies typically find that patient age and severity or DRG status are important determinants of LOS (*e.g.* Godfarb *et al.*, 1983; Cairns and Munroe, 1992) and patients of lower socio-economic status have longer LOS (*e.g.* Epstein *et al.*, 1990). Hence, as regressors we use age, gender, and resource groups to proxy for severity, and occupation, income and exemption to account for socio-economic differences.

The importance of hospital characteristics and organisational factors in determining length of stay has been established in some studies (*e.g.* Cannoodt and Knickman, 1984; Burns and Wholey, 1991; Xiao *et al.* 1997; Westert *et al.* 1993). These suggest that the organisation of discharge and unplanned admissions and physician workload can be relevant in determining LOS, as can the way by which inpatient services are financed (constant *per diem* fee or prospective payments may respectively increase or decrease it). Ensor and Thompson (1999) highlight that, whereas FSU and CEE countries used the criteria of beds and bed-days to fund hospitals, encouraging long LOSs, the system has now been replaced by case-based and prospective payment using DRGs introducing new incentives to promptly discharge patients. Medical standards specify the number of days a patient should stay in hospital. We do not have such detailed information on providers (although medical standards and the reimbursement method apply equally to all the 3

hospitals studied thus reducing the need to control for such variables) and therefore control for differences across hospitals (*e.g.* number of beds or doctors) using dummy variables.

Table 4 (appendix 2) provides a list of the dependent and independent variables used in the models. Table 5 presents some summary statistics. As before on average patients spent around 55 minutes in the AD waiting for a clinical intervention while spending around 14 days in hospital. About 33% of the patients paid in the AD while 22% paid when already in the ward with 6% of the patients paying in both places. On average individuals paid 2,950 KZT before receiving any clinical intervention and 1,797 KZT while already in the ward. Average montly income is around 20,683 KZT with the smallest income equal to 1,131 KZT. Around 25% and 50% of the respondents had respectively a monthly income of less then 10,000 and 17,340 KZT. Payment and income do vary considerably across individuals as suggested by the standard deviation. Half of the respondents are male and average age is 43 years. 13% are students, 18% are unemployed, 12% work in private companies, 4% are self-employed, 25% are retired and 10% are housewives. About 28% of the respondents are considered exempted from any payment for health care. Around 40% of all patients go to hospitals 2 and 3 while 19% go to hospital 1. More than 50% of patients are coded as RG3, with 33% coded as RG2 and around 7% in each of the other two diagnostic groups.

5. Regression analysis

The econometric models described and tested below explore the relationship between unofficial payment and (1) number of minutes spent in the AD; (2) LOS or number of days spent in hospital and (3) the patient's subjective measure of quality. We start by exploring the two first relationships by undertaking simple linear least squares' analysis (OLS regressions). Four specifications are estimated in total using the two different process indicators - admission wait and LOS - as dependent variables. We estimate the relationship between each of the two quality indicators and the unofficial payment as a binary variable, thus exploring whether the act of paying (**Pay1** and **Pay2**) is associated with the admission time or the LOS. We then use the amount paid (**lnpayment1** and **lnpayment2**) as the regressor and check in which way the continuous variable relates to waiting time or LOS. We consider the two unofficial payments separately but simultaneously (rather then their sum), the reason being that some patients although agreeing to pay when in the AD and before treatment, pay later, when already in the ward. Patients may not have the required amount ready at the moment of admission. Keeping payments separate allows us to distinguish each payment's effect and control for potential endogeneity reasons (see below).

When considering quality of care as perceived by patients, an ordered probit model (Maddala, 1983; Greene, 1993; STATA, 2002) is the estimation method used to account for the

ordinal nature of the dependent variable. Perceived quality is indeed a categorical variable ordering care using four categories from "Poor plus very poor" to "Very Good". This model is built around a latent regression $y^*=xb+e$ where y^* is unobserved and what is observed is for example $y=0$ if $y^*<\mu_1$; y=1 if $\mu_1<y^*<\mu_2$; $y=2$ if $\mu_2<y^*<\mu_3$; … $y=J$ if $\mu_J<y^*$. The μs are parameters to be estimated, together with the β, and constitute cut off points. The idea is that each respondent could potentially define y^* (given the regressors and unobservables) but given the categories proposed chooses the one closest to his answer. Assuming the residuals follow a normal distribution we can define the probabilities of the different possible outcomes as: $Prob(y=0)=Pr(xb+e<\mu_1)=Pr(e<\mu_1-xb)=F(\mu_1-xb)$;$Prob(y=1)=Pr(\mu_1<xb+e<\mu_2)=Pr(\mu_1-xb<e<\mu_2-xb)=F(\mu_2-xb)-F(\mu_1-xb)…Pr(y=J)=Pr(\mu_J<xb+e)=1-F(\mu_J-xb)$. Note that the marginal effects of the regressors, x, on the probabilities ($\partial F(x,b)/\partial x$) are not equal to the coefficients only, but rather: $\partial \Pr(y=0)/\partial x=-f(\mu_1-xb).b$, $\partial \Pr(y=1)/\partial x=[f(\mu_1-xb)-f(\mu_2-xb)].b$, … $\partial \Pr(y=J)/\partial x=f(\mu_j-xb).b$. Thus, to understand the relationship between payment and quality of care we need a fair amount of calculation. We need to see how the probability for each category changes with payment.

In terms of regression diagnostics used to assess the specification of the models these include firstly, in the case of the OLS the calculation of variance inflation factors (VIF) to assess for multicollinearity of the regressors. Second, we address potential heteroskedasticity by specifying the Huber/White/sandwich estimator of variance. Finally, we compute the Ramsey reset test for each of the OLS models estimated[11] and a Wald (χ_2) test for each of the ordered probit models estimated. These test for the general model specification namely in terms of omitted variables. Finally, we test for the potential endogeneity of payment in the ward (**lnpayment2**) using the Durbin-Wu-Hausman test (Davidson and Mackinnon, 1993). The endogeneity problem may arise from the fact that patients once experiencing the wait in the AD and LOS change their preferences concerning waiting and therefore the payment they make in the ward. In other words, **lnpayment2** may be endogenous.[12] Thus, we regress the potential endogenous variable on all the other explanatory variables together with any other variables that may help explaining payment. We compute the predicted residuals from this regression and introduced them into the original regression (OLS or ordered probit) so as to establish the significance of the corresponding coefficient estimate. An estimate that is statistically significantly different from zero suggests that

[11] This Ramsey Reset test introduces the predicted values of the dependent variable in their second, third, and fourth power into the regression and tests the joint significance of the respective coefficient estimates. It amounts to estimate $y=xb+zt+u$ (where z stands for the three powers of the predicted values of y) and test $t=0$. The Wald test consists of introducing the square of the predicted values of the dependent variable in the regression and testing the significance of the associated coefficient estimate.

[12] The definition of Pay1 makes it exogenous.

endogeneity may be in place. In that case the variable has to be instrumented for and a two stage least squares (2SLS) regression run in the case of OLS or the predicted values of the variable used instead of the observed values in the case of the ordered probit.

5.1. Time spent in the Admission Department (minutes)

As said, we examine the relationship between the time spent in the AD, defined as the total time a patient spends in the AD, from the time of admission to hospital to transfer to theatre or the ward, and a) the act of paying and b) the amount of unofficial payments made. The models can formally and generically specified as:

$$\begin{aligned} \ln admwait = {} & \beta_0 + \beta_1 RG + \beta_2 age + \beta_3 gender + \beta_4 pay + \beta_5 \ln income \\ & + \beta_6 occupation + \beta_7 Exemption + \beta_8 hospital + e \end{aligned} \tag{10a}$$

$$\begin{aligned} \ln admwait = {} & \beta_0 + \beta_1 RG + \beta_2 age + \beta_3 gender + \beta_4 \ln payment + \beta_5 \ln income \\ & + \beta_6 occupation + \beta_7 Exemption + \beta_8 hospital + \beta_9 hospital * \ln payment + e \end{aligned} \tag{10b}$$

with hospital 1, a surgical provider, and RG2 the reference hospital and the reference resource group. The results are presented in Tables 6 (act of paying) and 7 (amount of payment) in appendix 2. We specify the model using pooled data and data for each of the hospitals to assess differences across hospitals. When using the continuous payment variable we also specify a hospital-payment interaction model with the pooled data.

Looking at table 6 and at the models that pass the Ramsey Reset test – the regression models for hospital 1 and 2, we find that paying the AD - **Pay_1** -is negatively and significantly associated with a shorter time for admission in hospital 1, a surgical hospital, as indicated by the robust coefficient estimates. The act of paying is not associated with the time waiting for admission in hospital 2. The remaining models in Table 6 are deemed misspecified so that we cannot say much about the act of paying and its relation to the admission time when in those cases. Hence, surgical patients in hospital 1 waiting for surgery may perceive it to be worth making a payment in an attempt to decrease admission time. In hospital 1, being a **student** decreases the time in the AD while being retired increases it. In hospital 2, workers in the private sector- **privwork** - face a lower waiting time in the AD. Paying in the hospital ward – **Pay_2** - appears to be positively associated with admission wait. However, further inspection shows that it may be the case in hospital 3 (trauma). Moreover, both models do not pass the Reset test so that the positive association must be seen with caution.[13]

[13] Patients in hospital 3 may be agreeing to pay but pay after admission and "loose" time in the AD in the process of bargaining. Also, this variable may include payments made for reasons other than admission time, so that it is not necessarily the case that a negative relationship should be observed.

Examination of the amount paid in the AD (**lnpayment1**) and in the ward (**lnpayment2**) and their relation with admission time (Table 7), in the models considered well specified as suggested by the Ramsey Reset test, indicates that paying unofficially in the AD or in the ward is significantly associated with a lower waiting time for admission in the case of hospital 1, the surgical provider. These payments are not related to admission time in the case of hospital 2. Looking at the remaining models namely that with the interactions we can see that indeed the two payments are negatively related to admission time in hospital 1 only. The other models are again misspecified. When looking at socio-economic factors, we find that **retired** individuals (as compared to state workers) wait longer for admission to surgery in hospital 1 while **students** have shorter admission waits. Private sector workers - **privwork** - face a lower wait in the AD in hospital 2. Income – **lnincome** - is positively associated with waiting time for surgery in hospital 1, the rationale being that income is often a proxy for health status so that those richer and thus healthier may wait longer.

Finally, testing for potential endogeneity of the payment in the ward (**lnpayment2**) suggests that payment in the ward is not endogenous when admission time (lnadmwait) is analysed.[14] Therefore, we can conclude that, in hospital 1, paying in the AD is associated with a shorter wait, and the higher the unofficial payment, the shorter the admission wait is. In other words, patients in hospital 1 paid both in the AD and in the ward so as to reduce the time for admission. If time spent in the AD is indeed a proxy for quality of care so that the lower the time spent in the AD, the higher the quality, then the results support the theoretical model developed previously: patients pay unofficially to obtain better quality of care and physicians provide a differentiated service. These results also support the anedoctal reports of surgical patients interviewed during the survey process.

5.2. Length of hospital stay (days)

The second set of models examines the relationship between the dependent variable defined as number of days spent in hospital and the unofficial payments. The model is generally specified as:

$$\ln LOS = \beta_0 + \beta_1 RG + \beta_2 age + \beta_3 gender + \beta_4 pay + \beta_5 \ln income + \beta_6 occupation + \beta_7 Exemption + \beta_8 hospital + e \quad (11a)$$

$$\ln LOS = \beta_0 + \beta_1 RG + \beta_2 age + \beta_3 gender + \beta_4 \ln payment + \beta_5 \ln income + \beta_6 occupation + \beta_7 Exemption + \beta_8 hospital + \beta_9 hospital * \ln payment + e \quad (11b)$$

Again we make use of both unofficial payments - in the AD and in the ward - as explanatory variables the rationale being that patients may agree to pay for the care quality (*e.g.* LOS) they are

[14] Note that we first of all wish to look for evidence of an association between payments and quality after controlling for other variables. As such endogeneity is not an issue. For diagnostic rigour and for those wishing to learn further on the direction of causality we test for potential endogeneity.

to receive when they first encounter the hospital staff, that is, whilst in the AD and before admission, but some may however need to pay in instalments.

Table 8 shows the robust results of the regression of hospital LOS and the act of paying unofficially (binary variable) in the AD and in the ward. In the models that pass the Ramsey Reset test –pooled model and that of hospital 1 - it can be seen that patients admitted to **hospital 1** stayed in hospital less time than those going to the other hospitals with **hospital 3** registering the longer LOS. Those coded in **RG3** and **RG4** stay longer in hospital. Patients paying unofficially (in the AD or in the ward) have a longer stay in hospital, that is, **pay_1** and **pay_2** (especially in hospital 1) are positively and significantly associated with a longer LOS. Finally, being a man or a housewife increases hospital LOS while income decreases it. When looking at hospital 1 it is mainly the ward payment that appears to be related to LOS.

Examination of the amount paid in the AD or in the ward (Table 9) shows that these are positively and significantly associated with a longer stay in hospital. **Lnpayment1** and **lnpayment2** are positively related to LOS in all hospitals, with the association between lnpayment1 and LOS and that between lnpayment2 and LOS stronger in hospital 3 and 2 respectively. **Age** is positively related with LOS. Both **housewives** and **men** spend a longer time in hospital while **students** stay less long. Patients in **RG3** and **RG4** spend the longest in hospital. Income, perhaps reflecting health status, is negatively related to LOS: the less healthy stay longer in hospital.

Note that the R^2 values are quite high and all models, but that for hospital 2, pass the Ramsey Reset test. Checking for the potential endogeneity of the unofficial payment made by patients when in the ward - **lnpayment2** we find that it may take place in the contexts of the pooled model and that for hospital 3. Running a 2SLS estimation, where we instrument lnpayment2 using the entire set of the explanatory variables above plus education, referral type and surgery variables, we obtain a stronger positive and significant relationship between the amount paid in the ward and the LOS in (Table 9A in appendix 2).[15]

We can conclude that in acute surgery and trauma hospitals in urban Kazakhstan, paying unofficially in the AD and in the ward are related to a longer LOS in hospital. Moreover, the bigger the payment made, the longer is the stay, especially in the trauma context. Therefore, if LOS is a proxy for quality, which is potentially the case in Kazakhstan where post-hospital treatment is virtually non-existent, transport to hospital quite limited and expensive, and thus increased stay in hospital reassuring, then it can be said that patients are paying to improve the quality of care they

[15] Note that instruments in a cross-section context may be limited namely when using a survey. Thus, it is argued that the initial OLS analysis may still be "first-best". We present both models. If the focus is the sign not the magnitude of the association the results are in line with each other.

receive. Or, possibly, patients pay not to be discharged too early but have the required LOS for their condition.

5.3. *Perceived quality (categorical ordered variable)*

The second model examines the relationship between the perceived quality of care - a categorical variable taking four values (1=Poor plus very poor; 2=Satisfactory; 3=Good; 4=Very good) and ranking patients' perceptions in relation to the care received - and the unofficial payments. As mentioned, given the ordering characteristic of the dependent variable we use an ordered probit estimation. Again we make use of both types of unofficial payments, in the AD and in the ward. Due to the increased number of computations we have to undergo to analyse the relationship between payment and quality of care in this context we concentrate on the amount of payment made (continuous payment variables) only. The model can be generally specified as:

$$\begin{aligned} Perceived\ Quality = {} & \beta_0 + \beta_1 RG + \beta_2 age + \beta_3 gender + \beta_4\ ln\ payment + \beta_5\ ln\ income \\ & + \beta_6 occupation + \beta_7 Exemption + \beta_8 hospital + \beta_9 hospital * ln\ payment + e \end{aligned} \quad (12)$$

We test for endogeneity of the payment in the ward (**Lnpayment2**) and find that it may be endogenous. We therefore estimate **Lnpayment2** using the entire set of the explanatory variables above plus education, referral type and surgery variables and introduce the predicted values of **Lnpayment2** from this regression (**lnpayment2_hat**) into the ordered probit analysis (table 10B).

The results on table 10B and for the models passing the Wald test for general specification (non-pooled models and pooled with interactions) suggest that indeed a higher payment in the AD and a higher payment in the ward are associated with a higher perceived quality of care received as indicated by the positive and significant coefficient estimates of the payment variables. However, to fully understand the relationship between payment and quality of care we need to calculate the marginal effect of the regressors on each of the probabilities, that is, see how the probability for each category changes with payment.

Table 11 presents three possible cases (among the many potential combinations of values for the regressors). First, we use the mean value of all the explanatory variables. Second, we consider all dummies equal to zero (thus marginal effects are computed for **females**, in **Hospital 1**, in **RG2**, that are **state workers** and **non-exempted** patients, considering **mean age** and **income**). Finally, we attribute the value of one or zero to each dummy depending on whether its mean is closer to one or zero and whether it is one of the most representative groups (thus the marginal effects are computed for **males**, in **Hospital 3**, in **RG3**, that are **state workers, not exempted** from formal hospital charges, given **average age** and **income**). Overall, that is, for the above combinations and for all hospitals payment in the AD (**Lnpayment1**) and payment in the ward (**Lnpayment2**) decrease the

probability of receiving "Poor and very poor" as well as "Satisfactory" quality of care, and increase the probability of receiving a "Good" and "Very good" quality of care. The association between **Lnpayment1** and quality of care is the strongest for hospital 2 followed by hospital 3, while that between **Lnpayment2** and quality of care is the strongest for the strongest for hospital 2 followed by hospital 1, as suggested by the interaction terms. Patients going to **hospital 3** as compared to hospital 1 are more likely to consider care as "Good" and "Very good" and less likely to consider it "Satisfactory" or "Poor and very poor". The opposite holds for those going to **hospital 2** as compared to 1 (and 3). Those in **RG1** are more likely to classify care as of a "Good" or "Very good" category**.** The same is observed for those in **RG3,** while the opposite holds for those in **RG4** (although these two relationships do not appear to be significant). **Age** has no effect on perceived quality. **Men** are more likely to consider the quality of care to be "Poor and very poor" or "Satisfactory" rather than "Good" or "Very good". The **richer** the patients are the more likely they are to define care quality as "Poor and very poor" and "Satisfactory" (although relationship is not significant). In terms of occupation being **retired** increases the likelihood of seeing care as "Good" and "Very good" while it decreases the other two categories (with similar results for student, unemployed, and private workers when in the case of hospital 2). **Exempt** has the opposite effect.

Concentrating on the payment variables and their marginal effect, we have also looked at various other possible combinations different from those above (another type of diagnostic group or occupation considering each of the hospitals and each gender type) as shown in table 12. As before, payment decreases the probabilities of receiving "Poor and very poor care" and "Satisfactory care" while increasing the probabilities of obtaining "Good" and "Very good" care for most combinations. The exception can be observed when considering retired patients in hospitals 1 and 3. When these are taken into account payment increases the probability of receiving "Very good care" while decreasing the probability of receiving care of a lower level. Similarly if we consider patients coded as RG1 in hospital 3. In more detail one can seen that for example that in absolute terms the marginal effects of payment on "Poor and very poor" quality and on "Good" are higher for men while women have stronger marginal effects for "Satisfactory" and "Very good" care quality. Patients coded in RG4 have a higher marginal effect of payment (in absolute terms) than those in RG3, while these have a higher marginal effect than those in RG1 when looking at "Poor and very poor" quality and "Good" quality. The opposite holds when looking at "Very good" quality. Again retired individuals have a higher marginal effect of payment for the two top classes.

6. Preliminary Discussion and conclusions

In this paper we use theoretical and empirical analysis to investigate unofficial payments for health care in the transition world using the example of Kazakhstan. Following claims that patients perceive state health care provision to be poor and thus are willing to pay unofficially in an attempt to improve the quality of care received (*e.g.* reduce the time spent in the admission department), we presented a theoretical model that formalised this informal market for health care quality that takes place within state facilities. Patients' utility was thus assumed to depend on health care quality and monetary payment with patients having heterogeneous preferences for care quality. The resulting demand was therefore a function of payment and quality level. Physicians on the other side, having more information on diagnostic and information and enough ability to manipulate queues, decide upon resource use and treatment, exploited their monopoly position and maximised their unofficial income by engaging in discriminatory pricing and service differentiation (*i.e.* offering differing levels of service quality to paying and non-paying patients), doing so with the knowledge that corruption is largely ignored by the state. In equilibrium the level of unofficial payment made was thus positively associated with the level of quality of care.

We then conducted an empirical exploration of whether, other things being equal, patients in Kazakhstan are indeed paying unofficially to see the quality of care they receive improved. We constructed a unique data set from a survey of 1508 discharged patients treated in three hospitals in Almaty City, Kazakhstan for conditions whose treatment was completely and officially free of any charge. Given the sensitivity of the issue they were interviewed in their homes and shortly after discharge. We gathered information on their social and economic status and their experience in hospital. Bearing in mind the problems associated with the measurement of health care quality (McGuire, 2000) and in the absence of a correct, complete and tangible indicator, we computed two process measures of quality of care and patients' subjective evaluation using a categorical variable to investigate whether the above positive relationship between unofficial payment and care quality does hold, that is, whether payment is related to patients wanting to spend less time in the admission department, whether they would want to spend more time in hospital, and whether in their opinion they received better care. It is likely that the acutely ill patient relies on the physician to address his health care needs promptly and effectively. Yet, we think it realistic to assume that patients would want to be processed quicker in the admission department (*e.g.* Propper 2000; Bishai *et al.* 2000.) or that patients may wish to stay in hospital for as long as it takes to be reassured that their health status has indeed improved as a result of the treatment when follow-up care is poor or inexistent and

transport to hospital expensive. We find significant variation in terms of amounts paid, length of stay and perceived quality of care.

We conducted both OLS and ordered probit analysis, controlling for potential heteroskedasticity and endogeneity, and conducting the necessary and respective diagnostic tests. The (robust and well specified) results obtained showed significant associations between unofficial payments and quality of care received as measured by two process indicators - time a patient spends in the admission department and the number of days spent in hospital as well as by patients' subjective (categorical/ordered) measure of care quality. Indeed, the empirical analysis suggested that: 1) *paying* in the admission department before treatment takes place *and the amount paid* both in the admission department and in the ward is associated with a *reduction in the admission time* in hospital 1 (surgery); 2) *paying and the unofficial payment made* in the admission department and in the ward are *associated with longer length of stay in all hos*pitals; and 3) the amount *paid unofficially increases the likelihood* of considering the *quality of care to be "Good" or "Very good"*. Note that the positive association between payment and LOS may reflect, in reality, not extra days than necessary but simply the fact that patients are paying to remain in hospital for the number of days specified by medical standards, while those not paying stay in hospital are discharged too early (as specified by medical standards). By discharging non-paying patients earlier physicians are freeing up beds for other patients who might pay. It is also likely that patients may also be paying for reassurance and increased physician "effort". In that case the two process indicators are proxying only some dimensions of quality (although anedoctal reports suggest that patients may be paying to reduce the wait for admission). However we do find evidence of a strong association between payment and process as well as payment and the subjective evaluation of care received, that is, the results are robust across the three indicators which suggests that patients that make an unofficial payment do indeed receive different health care than that of those that do not pay anything. Moreover, the results and their consistency support the theoretical formalisation developed for the unofficial market for health care quality whereby patients with heterogeneous valuations of quality pay to receive better care and physicians exploit their monopoly power and provide a differentiated service to those paying and not paying.

This paper is a first attempt to use both economic theory and econometric tools to analyse the issue of unofficial payments and explore whether prior payment influences the quality of care received (measured using process and subjective indicators of quality) as compared to previous studies that were restricted to answering the "whom, how much, when and to whom" of the matter. We use a discriminatory price differentiated service model to formalise the unofficial market for quality of care and both ordinary least squares and ordered probit analysis to test it. We also use a

unique and detailed data set on discharged acute hospital patients that accurately identifies the amount paid unofficially while most previous studies did/could not distinguish between official and unofficial payments for care. Future research may extend the analysis to other transition countries. It would be also interesting to look at length of stay and its relation to unofficial payment using survival analysis.

Reading

Arrow, K. (1963). Uncertainty and the welfare economics of medical care, *American Economic review* 53(5): 941-973.

Bardhan, P. (1997). Corruption and development: A review of the issues, *Journal of Economic Literature* XXXV: 1320-1346.

Barnham, H. and Kutzin, J. (1993). *Public Hospitals in Developing Countries*, John Hopkins University Press, Baltimore, Maryland.

Barr, N. (1996) The ethics of Soviet medical practice: behaviour and attitudes of physicians in Soviet Estonia, *Journal of Medical Ethics* 22: 33-40.

Bishai, D. and Lang, H. (2000). The willingness to pay for wait reduction: the disutility of queues for cataract surgery in Canada, Denmark, and Spain, *Journal of Health Economics* 19: 219-230.

Bognar, G., Robert, I. and Kornai, J. (2000). Gratitude payments in the Hungarian health sector, *Kozgazdasagi Szemle* 47: 293-320.

Burns, L. and Wholey, D, (1991). The effects of patient, hospital and physician characteristics on length of stay and mortality, *Medical Care* 293: 251-271.

Cairns, J. and Monroe, J. (1992). Why does length of stay vary for orthopaedic surgery?, *Health Policy* 223: 297-306.

Campbell, S., Roland, M. and Buetow, S (2000). Defining quality of care, *Social Science and Medicine* 51: 1611-1625.

Cannoodt, L. and Knickman,J. (1984). The effect of hospital characteristics and organisational factors on pre- and postoperative lengths of hospital stay, *Health Services Research* 195: 561-585.

Chawla, M., Berman, P. and Kawiorska, D. (1998). Financing health services in Poland: New evidence on private expenditures, *Health Economics* 7: 337-346.

Delcheva, E. Balabanova, D. and McKee, M. (1997). Under-the-counter payments for health care: Evidence from Bulgaria, *Health Policy* 42: 89-100.

Donaldson, C. and Shackley, P. (1997). Does "process utility" exist? A case study of WTP for laproscopic cholycystectomy, *Social Science and Medicine* 44(5): 699-707.

Dranove, D. and Satterthwaite, M. (1992). Monopolistic competition when price and quality are imperfectly observable, *Rand Journal of Economics* 23(4): 518-534.

Eisenberg, J. (1986). *Doctors' Decisions and the Cost of Medical Care*, Health Administration Press, Ann Arbor, MI.

Ensor, T and Savelyeva, L. (1998). Informal payments for health care in the Former Soviet Union: some evidence from Kazakstan, *Health Policy and Planning* 13(1): 41-49.

Ensor, T. (2000). The unofficial business of health care in transitional Europe, *Eurohealth* 6(2) Spring Issue.

Ensor, T. and Rittmann, J. (1997) Reforming health care in the Republic of Kazakhstan, *International Journal of Health planning and Management* 12: 219-234.

Ensor, T., Thompson, R. (1999). Rationalising rural hospital services in Kazakstan, *International Journal of Health Planning and Management* 14: 155-167.

Epstein, A., Stern, A. and Weissman, J. (1990). Do the poor cost more? A multi-hospital study of patients' socio-economic status and use of hospital resources, *New England Journal of Medicine* 32216: 1122-1128.

European Observatory (1999). *Health Systems in Transition: Kazakhstan*, European Observatory on Health Care Systems.

Feldstein, P. (1979). *Health Care Economics*, John Wiley and Sons, New York.

Field, M. (1989). The position of the Soviet physician, *Milbank Quarterly* 66(2): 182-201.

Folland, S., Goodman, A. and Stano, M. (1993). *The Economics of Health and Health Care*, Prentice-Hall International, London, UK.

Forster, M. and Jones, A. (2001) Starting and quitting smoking, *Journal of the Royal Statistical Society*, 164(3): 517-547.

Gaal, P. (1999a). Informal payments in the Hungarian health services, mimeo.

Gaal, P. (1999b). Under-the-table payment and health care reform in Hungary, mimeo.

Galasi, P. and Kertesi, G. (1989). Rat race and equilibria in markets with side payments under socialism, *Acta Oeconomica* 41: 267-292.

Gaynor, M. (1994). Issues in the industrial organisation of the market for physician services, *The Journal of Economics and Management Strategy*, 3(1): 211-255.

Gaynor, M. and Gertler, P. (1995). Moral hazard and risk spreading in partnerships, *Rand Journal of Economics* 26: 591-614.

Goddard, J., Malek, M. and Tavakoli, M. (1995). An economic model of the market for hospital treatment for non-urgent conditions, *Health Economics* 4: 41-55.

Godfarb, M., Hornbrook, M. and Higgins,C. (1983). Determinants of hospital use: A cross-diagnostic analysis, *Medical Care* 211: 48-66.

Hamilton, B., Hamilton, V. (1997). Estimating surgical volume-outcome relationships applying survival models: accounting for frailty and hospital fixed effects, *Health Economics* 6: 383-395.

Healy,J and McKee,M . (1997). Health sector reform in Central and Eastern Europe, *Health Policy and Planning* 12 (4) 286-295.

Jones, A (2000) Health econometrics in Culyer, A.J. and Newhouse, J.P. (eds) *Handbook of Health Economics*, Volume 1, Chapter 6, Elsevier Science, Amsterdam.

Kessal, R. (1958). Price discrimination in medicine, *Journal of Law and Econometrics* 1: 20-53.

Kornai, J. (2000). Hidden in an envelope: gratitude payments to medical doctors in Hungary. Mimeo for the Festschrift in honour of George Soros.

Ladbury, S. (1997). *Social Assessment Study: Turkmenistan.* World Bank report mimeo.

Lewis, M. (2000), *Who Is Paying for Health Care in Eastern Europe and Central Asia?* World Bank

Liu, Y., Rao, K. and Fei, J. (1998). Economic transition and health transition: comparing China and Russia, *Health Policy* 44: 103-122.

Lui, F. (1985). An equilibrium queuing model of bribery, *Journal of Political Economy*, 93(4): 760-81.

Ma, C. and McGuire, T. (1997). Optimal health insurance and provider payment, *American Economic Review* 87(4): 685-704.

Martin, S. and Smith P. (1999). Rationing by waiting lists: an empirical investigation, *Journal of Public Economics* 71: 141-164

Martin, S. and Smith, P. (1996). Explaining variations in inpatient length of stay in the National Health Service, *Journal of Health Economics* 15: 279-304.

Masopust, V. (1989). Bribes in health care and patients opinions, Medline abstract, source Cesk Zdrav 37 (6-7): 299-307.

McCall, T. (1996). *Examining Your Doctor.* Citadel Press, Seacaucus, NJ

McGuire (2000). Physician Agency in Culyer, A.J. and Newhouse, J.P. (eds) *Handbook of Health Economics*, Volume 1, Chapter 9, Elsevier Science, Amsterdam.

Mirzoev, T. (1999). Corruption in Tajikistan as seen by the private sector, mimeo.

Myrdal, G. (1968). *Asian Drama.* Vol 2, Random house, New York.

Phelps, C. (1997). *Health economics*, 2nd edition, Harper Collins, New York.

Propper, C. (2000). The demand for private health care in the UK. *Journal of Health Economics* 19: 855-876.

Ruffin, R. and Leigh,D. (1973). Charity, competition, and the pricing of doctors' services. *The Journal of Human Resources* 8(2): 212-22

Ryan, M. (1978). The organisation of Soviet medical care. Professional Seminar Consultants, Inc.

Sari, N., Langenbrunner, J. and Lewis, M. (2000). Affording out-of-pocket payments for health services. *Eurohealth*, 6:2: Spring Issue.

Smith, H. (1973). *The Russians.* Sphere books.

Stata Press (2001). *Stata 7: Reference manuals.* Stata Corporation

Thompson, R. and Witter, S. (2000). Informal payments in transition economies: implications for health sector reform, *International Journal of Health Planning and Management* 15: 169-187.

Tirole, J. (1988). *The Theory of Industrial Organisation*, MIT Press, London.

Varian, H. (1987). *Intermediate Economics.* Norton, New York.

Westert, G., Niebor,A. and Groenewegan, P. (1993). Variation in duration of hospital stay between hospitals and between doctors within hospitals. *Social Science and Medicine* 376: 833-839.

World Bank (1998). *Kazakhstan: Living Standards During the Transition.* Report No: 17520-KZ.

World Bank (2000a). Armenia institutional and governance review, mimeo

World Bank (2000b). "Health" chapter in Czech Republic: Public expenditure review, mimeo.

Xiao, J., Douglas, D., Lee, A. and Vemuri, S. (1997). A Delphi evaluation of the factors influencing length of stay in Australian hospitals, *International Journal of Health Planning and Management* 12: 207-218.

Appendix 1

The doctor's maximisation problem is

$$\max_{p_H, p_L} U^D = (p_L - c_L)D_L(p_L, p_H, ..) + (p_H - c_H)D_H(p_L, p_H, ..)$$

And the first order conditions are, with $i,j=L,H$ and $i \neq j$:

$$\frac{\partial U^D}{\partial p_i} = D_i(p_i, p_j, .) + (p_i - c_i)\frac{\partial D_i(p_i, p_j, .)}{\partial p_i} + (p_j - c_j)\frac{\partial D_j(p_i, p_j, .)}{\partial p_i} = 0$$

Rearranging the terms we have that

$$D_i(p_i, p_j, .) + p_i \frac{\partial D_i(p_i, p_j, .)}{\partial p_i} + p_j \frac{\partial D_j(p_i, p_j, .)}{\partial p_i} = c_i \frac{\partial D_i(p_i, p_j, .)}{\partial p_i} - c_j \frac{\partial D_j(p_i, p_j, .)}{\partial p_i}$$

which shows that the monopolistic doctor chooses prices so that marginal revenue equals marginal costs. Dividing both sides by p_i we obtain

$$\frac{(p_i - c_i)}{p_i}\frac{\partial D_i(p_i, p_j, .)}{\partial p_i} = -\frac{D_i(p_i, p_j, .)}{p_i} - \frac{p_j - c_j}{p_i}\frac{\partial D_j(p_i, p_j, .)}{\partial p_i}$$

Further rearrangement yields

$$\frac{(p_i - c_i)}{p_i} = -\frac{D_i(p_i, p_j, .)}{p_i\left[\frac{\partial D_i(p_i, p_j, .)}{\partial p_i}\right]} - \frac{p_j - c_j}{p_i\left[\frac{\partial D_i(p_i, p_j, .)}{\partial p_i}\right]}\frac{\partial D_j(p_i, p_j, .)}{\partial p_i}$$

and $$\frac{(p_i - c_i)}{p_i} = -\frac{1}{\frac{p_i}{D_i(p_i, p_j, .)}\left[\frac{\partial D_i(p_i, p_j, .)}{\partial p_i}\right]} - \frac{p_j - c_j}{p_i\left[\frac{\partial D_i(p_i, p_j, .)}{\partial p_i}\right]}\frac{\partial D_j(p_i, p_j, .)}{\partial p_i}$$

which is equivalent to

$$\frac{(p_i - c_i)}{p_i} = \frac{1}{\varepsilon_{ii}^D} - \frac{p_j - c_j}{p_i\left[\frac{\partial D_i(p_i, p_j, .)}{\partial p_i}\right]}\frac{\partial D_j(p_i, p_j, .)}{\partial p_i}$$

$$\frac{(p_i - c_i)}{p_i} = \frac{1}{\varepsilon_{ii}^D} + \frac{(p_j - c_j)p_i D_j}{\varepsilon_{ii}^D D_i(p_i, p_j, .)p_i D_j}\frac{\partial D_j(p_i, p_j, .)}{\partial p_i}$$

for the first term on the right hand side (rhs) is the inverse of the own elasticity of demand, *i.e.*

$$\varepsilon_{ii}^D = -\frac{p_i}{D_i(p_i, p_j, .)}\frac{\partial D_i(p_i, p_j, .)}{\partial p_i}$$

Diving and multiplying the denominator of the second term on the rhs by D_i we obtain

$$\frac{(p_i - c_i)}{p_i} = \frac{1}{\varepsilon_{ii}^D} + \frac{(p_j - c_j)}{-\frac{p_i}{D_i(p_i, p_j, .)}\left[\frac{\partial D_i(p_i, p_j, .)}{\partial p_i}\right]D_i(p_i, p_j, .)}\frac{\partial D_j(p_i, p_j, .)}{\partial p_i}$$

$$\frac{(p_i - c_i)}{p_i} = \frac{1}{\varepsilon_{ii}^D} + \frac{(p_j - c_j)}{\varepsilon_{ii}^D D_i(p_i, p_j, .)}\frac{\partial D_j(p_i, p_j, .)}{\partial p_i}$$

As again

$$\varepsilon_{ii}^{D} = -\frac{p_i}{D_i(p_i, p_j, .)} \frac{\partial D_i(p_i, p_j, .)}{\partial p_i}$$

Finally multiplying and diving the second term by p_i and D_j

$$\frac{(p_i - c_i)}{p_i} = \frac{1}{\varepsilon_{ii}^{D}} + \frac{(p_j - c_j) p_i D_j}{\varepsilon_{ii}^{D} D_i(p_i, p_j, .) p_i D_j} \frac{\partial D_j(p_i, p_j, .)}{\partial p_i}$$

$$\frac{(p_i - c_i)}{p_i} = \frac{1}{\varepsilon_{ii}^{D}} - \frac{(p_j - c_j) D_j \varepsilon_{ji}^{D}}{\varepsilon_{ii}^{D} R_i}$$

as

$$\varepsilon_{ii}^{D} = -\frac{p_i}{D_i(p_i, p_j, .)} \frac{\partial D_i(p_i, p_j, .)}{\partial p_i} \text{ and } \frac{\partial D_i(p_i, p_j, .)}{\partial p_i} < 0$$

$$\varepsilon_{ji}^{D} = -\frac{p_i}{D_j(p_i, p_j, .)} \frac{\partial D_j(p_i, p_j, .)}{\partial p_i} \text{ and } \frac{\partial D_j(p_i, p_j, .)}{\partial p_i} > 0 \text{ when substitutegoods}$$

Appendix 2: Tables

Descriptive analysis

Table 1: Minutes spent in admission department by hospital and resource group (mean, s. d., no)

Hospital	RG 1	RG 2	RG 3	RG 4	Total
1	60.0 (56.8) 18	61.3 (60.0) 90	57.2 (51.8) 110	57.9 (51.9) 41	58.6 (54.9) 259
2	51.2 (42.6) 80	54.9 (48.1) 257	54.4 (48.8) 193	57.7 (44.8) 38	54.4 (47.2) 568
3	50.0 (0.0) 1	46.9 (39.1) 118	56.9 (58.2) 402	66.5 (104.1) 20	55.0 (57.0) 541
Total	52.8 (45.1) 99	54.1 (48.8) 465	56.2 (54.7) 705	59.6 (63.0) 99	**55.5** (52.8) 1368

Table 2: Days spent in hospital by hospital and resource group (mean, s.d ,no)

Hospital	RG 1	RG 2	RG 3	RG 4	Total
1	6.0 (3.8) 24	5.9 (4.1) 107	7.8 (5.4) 126	8.6 (4.0) 42	7.0 (4.8) 299
2	6.3 (2.5) 84	7.9 (6.4) 266	10.4 (5.3) 211	14.3 (6.2) 41	9.0 (6.0) 602
3	3.0 (0.0) 1	23.0 (25.8) 127	22.8 (19.2) 447	19.1 (18.6) 27	22.7 (20.7) 602
Total	6.2 (2.8) 109	11.3 (15.5) 500	17.1 (16.3) 784	13.3 (11.0) 110	**14.1** (15.5) 1503

Table 3: Distribution of responses (no and percentage) characterising the quality of care (categorical variable) by hospital

	Hospital 1		**Hospital 2**		**Hospital 3**		**Total**	
Perceived quality	**Obs.**	**Percent**	**Obs.**	**Percent**	**Obs.**	**Percent**	**Obs.**	**Percent**
Very good	34	11.33%	78	12.94%	85	14.05%	197	13.06%
Good	171	57%	313	51.91%	374	61.82%	858	56.9%
Satisfactory	75	25%	146	24.21%	115	19.01%	336	22.28%
Poor	17	5.67%	53	8.79%	24	3.97%	94	6.23%
Very poor	3	1%	13	2.16%	7	1.16%	23	1.53%
Poor plus very poor	*20*	*6.6%*	*6*	*10.95%*	*31*	*5.12%*	*117*	*7.76%*
Total	300		603		605		1508	

Table 4: Variables used in the empirical analysis

Variable code	***Description***
Dependent	
Lnadmwait	The log of the number of minutes an individual spends in the Admission Department
Lnlos	The log of the number of days an individual spends in hospital
Perceived quality	Categorical ordered variable**:** 1=not good, 2=satisfactory, 3=good and 4=very good
Independent	
Socio-economic variables	
Age	Age, in years
Male	Binary gender, male = 1
Student, unemploy, statwork, privwork, selfwork, retired, Houswife	Student, unemployed, state employee, private company employee, self employed, retired, housewife, (Dummy variables)
Exempt	Registered exempt = 1(Dummy variable)
Lnincome	The log of the household adjusted monthly consumption expenditure (income proxy) in local currency (KZT)
Payment variables:	
Pay_1 Pay_2 Lnpayment1 Lnpayment2	Pay_1 and Pay_2 (binary variables: 1=patient paid and 0=patient did not pay) The log of the amount of KZT paid in the Admission Dept The log of the amount of KZT paid in the ward
Hospital specific variables	
Hospital 1, Hospital 2, Hospital 3 (Trauma)	Hospitals 1, 2 and 3 (Dummy variables)
Hosp1*lnpayment, Hosp2*lnpayment, Hosp3*lnpayment	Hospital payment interactions

Table 5: Descriptive statistics

Variables	Obs	Mean	Std. Dev.	Min	Max
Dependent variables					
Admwait	1368	55.512	52.765	3	720
Lnadmwait	1368	3.782	0.749	1.792	6.583
LOS	1496	13.556	13.081	1	90
LnLOS	1496	2.592	0.622	1.386	4.533
Independent variables					
Pay_1	1508	0.338	0.473	0	1
Payment1	1452	2949.345	6145.867	0	52000
Lnpayment1	1452	3.397	3.528	1.099	10.859
Pay_2	1508	0.202	0.402	0	1
Payment2	1483	1796.129	4743.295	0	35000
Lnpayment2	1483	2.538	3.025	1.099	10.463
Age	1508	42.989	18.004	5	89
Male	1508	0.505	0.500	0	1
Income	1494	20683.010	14745.390	1130.348	144000
Lnincome	1494	9.726	0.660	7.030	11.878
Student	1508	0.133	0.340	0	1
Unemploy	1508	0.184	0.387	0	1
Privwork	1508	0.117	0.322	0	1
Selfwork	1508	0.036	0.188	0	1
Retired	1508	0.253	0.435	0	1
Houswife	1508	0.102	0.303	0	1
Exempt	1508	0.284	0.451	0	1
RG1	1508	0.072	0.259	0	1
RG2	1508	0.334	0.472	0	1
RG3	1508	0.521	0.500	0	1
RG4	1508	0.074	0.261	0	1
Hospital 1	1508	0.199	0.399	0	1
Hospital 2	1508	0.400	0.490	0	1
Hospital 3	1508	0.401	0.490	0	1

Regression analysis

Note: Payment variables are all assumed to be unofficial payments made in either the admission department (pay_1 and lnpayment1) or on the ward (pay_2 and lnpayment2). The tables presented here show estimates using binary payment variables (yes/no response) and continuous payment variables. The nature of the variable transformation is defined at the head of each table.

Table 6: Admission time regressed on whether or not an unofficial payment was made.

	Pooled Binary Pay	**Hospital 1 Binary Pay**	**Hospital 2 Binary Pay**	**Hospital 3 Binary Pay**
Admission time (lnadmwait)	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Hospital 3 (Trauma)	-0.0659			
Hospital 2	-0.0807			
Rg1	0.0804	-0.0594	0.0887	0.3262*
Rg3	0.0705	-0.0998	-0.0085	0.1860**
Rg4	0.1159	-0.0270	0.1467	0.2013
Age	-0.0028	-0.0067	0.0001	-0.0020
Male	-0.0030	-0.1100	-0.0218	0.0524
Lnincome	-0.0327	0.2086**	0.0339	-0.2024*
Pay_1	0.0445	-0.2068**	0.0042	0.1766**
Pay_2	0.1433*	-0.0739	0.1015	0.3104*
Student	-0.1167	-0.3778**	-0.0805	-0.0551
Unemploy	0.0538	0.1929	0.0742	0.0160
Privwork	-0.0739	0.2045	-0.2642***	-0.0378
Selfwork	0.0902	-0.1360	0.2184	0.0092
Retired	0.1517	0.7162*	0.3143	-0.4110***
Houswife	0.0567	0.0978	0.0655	-0.0034
Exempt	-0.1087	-0.2307	-0.3248	0.3347
_cons	4.1802*	2.1032**	3.4215*	5.5740*
No of observations	1358	253	567	538
R2	0.0163	0.1044	0.0332	0.0752
Ramsey Reset test	F(3,1337)=3.13 Prob>F=0.0247	F(3,234)=0.58 Prob>F=0.6275	F(3,548)=0.49 Prob>F=0.6869	F(3,548)=3.27 Prob>F=0.0212
Mean VIF	2.16	2	2.21	2.28

Notes: *, **, and *** stand for significance level of 1%, 5% and 10% respectively. Estimations are robust.

Table 7: Admission time regressed on amount paid as unofficial payment (continuous).

	Pooled Continuous Pay	**Hospital 1 Continuous Pay**	**Hospital 2 Continuous Pay**	**Hospital 3 Continuous Pay**	**Pooled Continuous Pay Interactions**
Admission Time (lnadmwait)	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Hospital 3 (Trauma)	-0.0495				-0.3359*
Hospital 2	-0.0766				-0.2350**
Rg1	0.0631	-0.1125	0.0819	0.1293	0.0598
Rg3	0.0757	-0.1198	-0.0150	0.2026*	0.0641
Rg4	0.1073	-0.0325	0.1212	0.2089	0.1174
Age	-0.0023	-0.0074	0.0012	-0.0015	-0.0024
Male	-0.0044	-0.1418	-0.0170	0.0468	-0.0059
Lnincome	-0.0239	0.2611*	0.0330	-0.1973*	-0.0179
Lnpayment1	0.0115***	-0.0436*	0.0060	0.0382*	-0.0270***
Hosp2*lnpayment1					0.0322***
Hosp3*lnpayment1					0.0641*
Lnpayment2	0.0153**	-0.0218***	0.0142	0.0360*	-0.0189***
Hosp2*lnpayment2					0.0354**
Hosp3*lnpayment2					0.0483*
Student	-0.1052	-0.3708***	-0.0821	0.0255	-0.1063
Unemploy	0.0433	0.2266	0.0559	0.0101	0.0529
Privwork	-0.1035	0.2145	-0.2895***	-0.0691	-0.0946
Selfwork	0.0669	-0.0471	0.1835	-0.0183	0.0860
Retired	0.1438	0.8352*	0.3069	-0.4479***	0.1648
Houswife	0.0477	0.0994	0.0458	-0.0138	0.0302
Exempt	-0.1243	-0.2964***	-0.3441***	0.3375	-0.1292
_cons	4.0490*	1.7260**	3.3745*	5.4275*	4.1766*
No of observations	1308	245	553	510	1308
R2	0.0149	0.1334	0.033	0.083	0.0277
Ramsey Reset test	F(3,1287)=3.82 Prob>F=0.0097	F(3,226)=0.27 Prob>F=0.8443	F(3,534)=0.75 Prob>F=0.5238	F(3,491)=2.77 Prob>F=0.0412	F(3,1283)=2.85 Prob>F=0.0363
Mean VIF	2.18	2.04	2.21	2.34	3.65

Notes: *, **, and *** stand for significance level of 1%, 5% and 10% respectively. Estimations are robust.

Table 8: LOS regressed on whether or not an unofficial payment was made.

	Pooled Binary Pay	**Hospital 1 Binary Pay**	**Hospital 2 Binary Pay**	**Hospital 3 Binary Pay**
LOS (lnlos)	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Hospital 3 (Trauma)	0.6828*			
Hospital 2	0.1798*			
Rg1	-0.0449	0.0386	-0.0347	-1.0097*
Rg3	0.1911*	0.1917*	0.1968*	0.1464**
Rg4	0.2702*	0.2870*	0.4255*	0.0525
Age	0.0019	0.0016	0.0058*	-0.0016
Male	0.0707**	0.1372*	0.0406	0.0198
Lnincome	-0.0623*	0.0165	-0.0706*	-0.0742
Pay_1	0.1333*	-0.0424	0.0023	0.3267*
Pay_2	0.2690*	0.3504*	0.1472*	0.3385*
Student	-0.0814	0.0064	-0.0313	-0.2157***
Unemploy	0.0147	0.0824	-0.0118	0.0195
Privwork	-0.0523	-0.0226	-0.0140	-0.0992
Selfwork	-0.0137	0.2283	0.0100	-0.1488
Retired	0.0615	0.2054	0.0209	-0.0073
Houswife	0.1028**	0.0832	0.0043	0.0826
Exempt	0.0134	-0.0125	-0.1034	0.1776
_cons	2.5066*	1.6664*	2.7366*	3.4584*
No of observations	1482	290	600	592
R2	0.3463	0.2393	0.2317	0.107
Ramsey Reset test	F(3,1461)=2.2 Prob>F=0.0861	F(3,271)=1.66 Prob>F=0.1763	F(3,5811)=4.65 Prob>F=0.0032	F(3,573)=6.47 Prob>F=0.0003
Mean VIF	2.15	2.05	2.21	2.25

Notes: *, **, and *** stand for significance level of 1%, 5% and 10% respectively. Estimations are robust.

Table 9: LOS regressed on amount of payment (continuous).

	Pooled Continuous Pay	**Hospital 1 Continuous Pay**	**Hospital 2 Continuous Pay**	**Hospital 3 Continuous Pay**	**Pooled Continuous Pay Interactions**
LOS	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Hospital 3 (Trauma)	0.6653*				0.5586*
Hospital 2	0.1692*				0.3188*
Rg1	-0.0419	0.0524	-0.0363	-0.9668*	-0.0404
Rg3	0.1834*	0.1936*	0.1925*	0.1586**	0.1775*
Rg4	0.2575*	0.2827*	0.4024*	0.0670	0.2714*
Age	0.0027***	0.0020	0.0054*	0.0000	0.0025***
Male	0.0666**	0.1424*	0.0473	-0.0018	0.0547***
Lnincome	-0.0740*	0.0013	-0.0755*	-0.0869***	-0.0653*
Lnpayment1	0.0199*	0.0029	0.0001	0.0503*	0.0103***
Hosp2*lnpayment1					-0.0095
Hosp3*lnpayment1					0.0383*
Lnpayment2	0.0418*	0.0496*	0.0198*	0.0599*	0.0554*
Hosp2*lnpayment2					-0.0342*
Hosp3*lnpayment2					0.0009
Student	-0.0879***	-0.0072	-0.0470	-0.2359***	-0.1009**
Unemploy	-0.0084	0.0806	-0.0371	-0.0133	-0.0063
Privwork	-0.0614	-0.0300	-0.0447	-0.0965	-0.0639
Selfwork	-0.0307	0.2072***	-0.0009	-0.1684	-0.0167
Retired	0.0563	0.1664	0.0079	0.0009	0.0544
Houswife	0.0938***	0.0697	-0.0016	0.0750	0.0672
Exempt	0.0045	0.0163	-0.0957	0.1549	0.0234
_cons	2.5307*	1.7258*	2.7887*	3.3650*	2.4472*
No of observations	1424	282	586	556	1424
R2	0.49938	0.255	0.2297	0.1307	0.3688
Ramsey Reset test	F(3,1403)=0.71 Prob>F=0.5431	F(3,263)=1.03 Prob>F=0.3784	F(3,567)=5.05 Prob>F=0.0019	F(3,537)=1.46 Prob>F=0.2240	F(3,1399)=1.92 Prob>F=0.1251
Mean VIF	2.18	2.08	2.21	2.32	3.61

Notes: *, **, and *** stand for significance level of 1%, 5% and 10% respectively. Estimations are robust.

Table 9A: LOS regressed on (continuous) payment taking into account the potential endogeneity of the payment in the ward (lnpayment2).

	Pooled Continuous pay	**Hospital 3 Continuous Pay**
LOS	**Coef**	**Coef.**
Lnpayment2	0.2020*	0.2758*
Hospital 3 (Trauma)	0.6321*	
Hospital 2	0.0691	
Rg1	0.0385	-0.5919*
Rg3	0.1949*	0.2400*
Rg4	0.2448*	0.2249
Age	0.0039**	0.0024
Male	0.0639	-0.0379
Lnincome	-0.2135*	-0.2389*
Lnpayment1	0.0490*	0.0775*
Student	-0.0324	-0.1245
Unemploy	0.0540	0.1260
Privwork	0.0254	0.0023
Selfwork	-0.1967***	-0.3770
Retired	0.1917***	0.1023
Houswife	0.1495***	0.1263
Exempt	0.0533	0.3353
_cons	3.2949*	3.9398*

Instrumented: Lnc22
Instruments: rg1-rg4, age, male, lnincome, lnc21, occupation, exemption, university, type of referral, surgery.

Table 10A: Perceived quality (1=not good, 2=satisfactory, 3=good and 4=very good) regressed on amount paid as informal payments

	Pooled Continuous Pay	**Hospital 1 Continuous Pay**	**Hospital 2 Continuous Pay**	**Hospital 3 Continuous Pay**	**Pooled Continuous Pay Interactions**
Perceived quality	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Hospital 3 (Trauma)	0.2063**				0.1450
Hospital 2	-0.0464				-0.1341
Rg1	0.1578	-0.1579	0.1883	8.8740*	0.1569
Rg3	0.0304	-0.3045***	0.1218	0.0756	0.0303
Rg4	-0.1811	-0.6061*	-0.0356	0.0890	-0.1747
Age	-0.0020	0.0015	-0.0003	-0.0068	-0.0020
Male	-0.1163***	-0.0236	-0.1993***	-0.0640	-0.1108***
Lnincome	0.1883*	0.0665	0.3463*	0.0936	0.1890*
Lnpaymet1	-0.0172**	-0.0260	-0.0143	-0.0269***	-0.0353***
Hosp2*lnpayment1					0.0252
Hosp3*lnpayment1					0.0153
Lnpayment2	-0.0044	-0.0023	-0.0127	0.0020	-0.0101
Hosp2*lnpayment2					0.0051
Hosp3*lnpayment2					0.0103
Student	0.0433	0.1254	0.1148	-0.2375	0.0486
Unemploy	0.0298	-0.0367	0.1590	-0.0416	0.0369
Privwork	-0.0316	-0.1969	-0.0122	0.0750	-0.0260
Selfwork	0.4824*	0.2476	0.6671*	0.3040	0.4808*
Retired	0.3870**	0.0758	0.4747	0.5037	0.3999**
Houswife	-0.0820	-0.4519	-0.1838	0.1744	-0.0721
Exempt	-0.3031***	0.0394	-0.3319	-0.4134	-0.3076***
Cut 1	0.2637	-1.1426	2.0163	-1.1032	0.2200
Cut 2	1.1786	-0.1002	2.8977	-0.1709	1.1350
Cut 3	2.8595	1.6426	4.4575	1.6540	2.8169
Prob(y=1|x)	0.0781	0.0709	0.109	0.0496	0.0781
Prob(y=2|x)	0.2197	0.2482	0.2419	0.1823	0.2197
Prob(y=3|x)	0.5690	0.5674	0.5179	0.6230	0.5690
Prob(y=4|x)	0.1332	0.1135	0.1312	0.1451	0.1332
No Observations	1434	282	587	565	1434
Pseudo R2	0.0183	0.0293	0.0273	0.0158	0.0187
Wald test for all variables	Chi2(17)=60.27 Pr>chi2=0.000	Chi2(15)=20.17 Pr>chi2=0.165	Chi2(15)=34.99 Pr>chi2=0.003	Chi2(15)=3345.72 Pr>chi2=0.000	Chi2(21)=61.94 Pr>chi2=0.000
Log-likelihood	-1578.69	-301.68	-680.66	-575.02	-1578.09
Wald test for omitted variables	Chi2(1)=6.81 Pr>chi2=0.009	Chi2(1)=0.07 Pr>chi2=0.785	Chi2(1)=0.75 Pr>chi2=0.387	Chi2(1)=0.03 Pr>chi2=0.857	Chi2(1)=6.35 Pr>chi2=0.012

Notes: *, **, and *** stand for significance level of 1%, 5% and 10% respectively. Estimations are robust.

Table 10B: Perceived quality (1=not good, 2=satisfactory, 3=good and 4=very good) regressed on amount paid as informal payments controlling for endogeneity of payment2

	Pooled Continuous Pay	Hospital 1 Continuous Pay	Hospital 2 Continuous Pay	Hospital 3 Continuous Pay	Pooled Continuous Pay Interactions
Perceived quality	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Hospital 3 (Trauma)	0.1487***				0.1814
Hospital 2	-0.1770**				-0.6362**
Rg1	0.2396***	-0.0574	0.3035**	9.4633*	0.2689**
Rg3	0.0510	-0.2563	0.0784	0.0906	0.0615
Rg4	-0.1919	-0.5859*	-0.4058**	0.1166	-0.1796
Age	0.0000	-0.0017	0.0057	-0.0064	0.0000
Male	-0.1314**	0.0415	-0.1674	-0.0730	-0.1162***
Lnincome	-0.0021	-0.1927	-0.0872	0.0663	-0.0244
Lnpayment1	0.0208	0.0187	0.0786*	-0.0228	0.0026
Hosp2*lnpayment1					0.0437**
Hosp3*lnpayment1					0.0078
Lnpayment2	0.2087*	0.2713*	0.3947*	0.0413	0.1968*
Hosp2*lnpayment2					0.1136
Hosp3*lnpayment2					-0.0198
Student	0.1340	-0.2521	0.5132**	-0.2129	0.1276
Unemploy	0.1160	-0.1132	0.3902**	-0.0145	0.1282
Privwork	0.0882	-0.2684	0.4323**	0.0915	0.1101
Selfwork	0.2590	-0.3842	0.4109	0.2617	0.1976
Retired	0.5872*	0.3245	0.9145*	0.5477***	0.6022*
Houswife	-0.0109	-0.3295	0.1050	0.1628	0.0173
Exempt	-0.2872***	0.0570	-0.3413	-0.4151	-0.2834***
Cut 1	-0.8011	-3.0157	-0.1988	-1.2168	-1.0709
Cut 2	0.1152	-1.9502	0.6901	-0.2815	-0.1517
Cut 3	1.8073	-0.1770	2.2672	1.5480	1.5439
Prob(y=1|x)	0.0784	0.0709	0.11	0.0492	0.0784
Prob(y=2|x)	0.2191	0.2482	0.2403	0.1828	0.2191
Prob(y=3|x)	0.5693	0.5674	0.5178	0.6239	0.5693
Prob(y=4|x)	0.1331	0.1135	0.1320	0.1441	0.1331
No of observations	1442	282	591	569	1442
Pseudo R2	0.0227	0.0435	0.0368	0.0160	0.0247
Wald test for all variables	Chi2(17)=74.04 Prob>F=0.000	Chi2(15)=31.82 Prob>F=0.007	Chi2(15)=49.98 Prob>F=0.000	Chi2(15)=4197.61 Prob>F=0.000	Chi2(21)=79.65 Prob>F=0.000
Log-likelihood	-1580.37	-297.27	-679.39	-578.01	-1577.06
Wald test for omitted variables	Chi2(1)=9.39 Pr>chi2=0.002	Chi2(1)=2.43 Pr>chi2=0.119	Chi2(1)=0.08 Pr>chi2=0.772	Chi2(1)=0.05 Pr>chi2=0.831	Chi2(1)=3.31 Pr>chi2=0.069

Notes: *, **, and *** stand for significance level of 1%, 5% and 10% respectively. Estimations are robust.

Table 11: Marginal effects for the five different probabilities

	Mean values of variables					Baseline (all dummies = 0)					Attributing 1 or 0 to a dummy if mean closer to 1 or 0 or more representative group				
	X	**dF(y=1)/dx**	**dF(y=2)/dx**	**dF(y=3)/dx**	**dF(y=4)/dx**	**X**	**dF(y=1)/dx**	**dF(y=2)/dx**	**dF(y=3)/dx**	**dF(y=4)/dx**	**X**	**dF(y=1)/dx**	**dF(y=2)/dx**	**dF(y=3)/dx**	**dF(y=4)/dx**
Hospital 3 (trauma)	0.395	-0.0236	-0.0376	0.0225	0.0388	0	-0.0284	-0.0370	0.0322	0.0333	1	-0.0237	-0.0376	0.0227	0.0386
Hospital 2	0.410	0.0827	0.1320	-0.0788	-0.1359	0	0.0996	0.1299	-0.1129	-0.1166	0	0.0831	0.1320	-0.0796	-0.1355
Rg1	0.072	-0.0350	-0.0558	0.0333	0.0575	0	-0.0421	-0.0549	0.0477	0.0493	0	-0.0351	-0.0558	0.0337	0.0573
Rg3	0.513	-0.0080	-0.0128	0.0076	0.0131	0	-0.0096	-0.0126	0.0109	0.0113	1	-0.0080	-0.0128	0.0077	0.0131
Rg4	0.074	0.0234	0.0373	-0.0222	-0.0384	0	0.0281	0.0367	-0.0319	-0.0329	0	0.0235	0.0373	-0.0225	-0.0383
Age	43.244	0.0000	0.0000	0.0000	0.0000	43.244	0.0000	0.0000	0.0000	0.0000	43.244	0.0000	0.0000	0.0000	0.0000
Male	0.502	0.0151	0.0241	-0.0144	-0.0248	0	0.0182	0.0237	-0.0206	-0.0213	1	0.0152	0.0241	-0.0145	-0.0248
Lnincome	9.719	0.0032	0.0051	-0.0030	-0.0052	9.719	0.0038	0.0050	-0.0043	-0.0045	9.719	0.0032	0.0051	-0.0031	-0.0052
Lnpayment1	3.410	**-0.0003**	**-0.0005**	**0.0003**	**0.0006**	3.410	**-0.0004**	**-0.0005**	**0.0005**	**0.0005**	3.410	**-0.0003**	**-0.0005**	**0.0003**	**0.0006**
Lnpayment2	2.556	**-0.0271**	**-0.0433**	**0.0259**	**0.0446**	2.556	**-0.0327**	**-0.0426**	**0.0370**	**0.0383**	2.556	**-0.0273**	**-0.0433**	**0.0261**	**0.0445**
Hosp2*lnpayment1	1.749	**-0.0057**	**-0.0091**	**0.0054**	**0.0093**	0	**-0.0068**	**-0.0089**	**0.0078**	**0.0080**	0	**-0.0057**	**-0.0091**	**0.0055**	**0.0093**
Hosp3*lnpayment1	1.177	**-0.0010**	**-0.0016**	**0.0010**	**0.0017**	0	**-0.0012**	**-0.0016**	**0.0014**	**0.0014**	1.177	**-0.0010**	**-0.0016**	**0.0010**	**0.0017**
Hosp2*lnpayment2	1.118	**-0.0148**	**-0.0236**	**0.0141**	**0.0243**	0	**-0.0178**	**-0.0232**	**0.0202**	**0.0208**	0	**-0.0148**	**-0.0236**	**0.0142**	**0.0242**
Hosp3*lnpayment2	0.965	**0.0026**	**0.0041**	**-0.0024**	**-0.0042**	0	**0.0031**	**0.0040**	**-0.0035**	**-0.0036**	0.965	**0.0026**	**0.0041**	**-0.0025**	**-0.0042**
Student	0.129	-0.0166	-0.0265	0.0158	0.0273	0	-0.0200	-0.0260	0.0226	0.0234	0	-0.0167	-0.0265	0.0160	0.0272
Unemploy	0.182	-0.0167	-0.0266	0.0159	0.0274	0	-0.0201	-0.0262	0.0227	0.0235	0	-0.0167	-0.0266	0.0160	0.0273
Privwork	0.114	-0.0143	-0.0229	0.0136	0.0235	0	-0.0172	-0.0225	0.0195	0.0202	0	-0.0144	-0.0229	0.0138	0.0235
Selfwork	0.037	-0.0257	-0.0410	0.0245	0.0422	0	-0.0309	-0.0404	0.0351	0.0362	0	-0.0258	-0.0410	0.0247	0.0421
Retired	0.260	-0.0783	-0.1250	0.0746	0.1287	0	-0.0943	-0.1230	0.1068	0.1104	1	-0.0787	-0.1250	0.0754	0.1283
Houswife	0.103	-0.0022	-0.0036	0.0021	0.0037	0	-0.0027	-0.0035	0.0031	0.0032	0	-0.0023	-0.0036	0.0022	0.0037
Exempt	0.291	0.0368	0.0588	-0.0351	-0.0606	0	0.0444	0.0579	-0.0503	-0.0520	0	0.0370	0.0588	-0.0355	-0.0604

Table 12: Marginal effects of payment variables for various cases

	HOSPITAL 1	dF(y=1)/dx	dF(y=2)/dx	dF(y=3)/dx	dF(y=4)/dx
if male	**lnpayment1**	-0.0005	-0.0005	0.0006	0.0004
	lnpayment2	-0.0376	-0.0410	0.0452	0.0333
if female	**lnpayment1**	-0.0004	-0.0005	0.0004	0.0005
	lnpayment2	-0.0322	-0.0427	0.0362	0.0387
if RG1	**lnpayment1**	-0.0003	-0.0005	0.0001	0.0006
	lnpayment2	-0.0215	-0.0424	0.0118	0.0521
if RG3	**lnpayment1**	-0.0004	-0.0005	0.0004	0.0005
	lnpayment2	-0.0296	-0.0432	0.0310	0.0417
if RG4	**lnpayment1**	-0.0005	-0.0005	0.0006	0.0004
	lnpayment2	-0.0406	-0.0395	0.0496	0.0305
if privworker	**lnpayment1**	-0.0003	-0.0005	0.0003	0.0005
	lnpayment2	-0.0275	-0.0433	0.0267	0.0441
if unemploy	**lnpayment1**	-0.0003	-0.0005	0.0003	0.0006
	lnpayment2	-0.0268	-0.0433	0.0251	0.0450
if retired	**lnpayment1**	-0.0001	-0.0004	-0.0003	0.0008
	lnpayment2	-0.0117	-0.0357	-0.0206	0.0681
if exempt	**lnpayment1**	-0.0006	-0.0004	0.0007	0.0003
	lnpayment2	-0.0458	-0.0362	0.0558	0.0262
	HOSPITAL 2	**dF(y=1)/dx**	**dF(y=2)/dx**	**dF(y=3)/dx**	**dF(y=4)/dx**
if male	**lnpayment1**	-0.0007	-0.0003	0.0008	0.0002
	lnpayment2	-0.0591	-0.0239	0.0660	0.0169
if female	**lnpayment1**	-0.0007	-0.0004	0.0008	0.0003
	lnpayment2	-0.0533	-0.0300	0.0626	0.0207
if RG1	**lnpayment1**	-0.0005	-0.0005	0.0006	0.0004
	lnpayment2	-0.0399	-0.0399	0.0485	0.0312
if RG3	**lnpayment1**	-0.0006	-0.0004	0.0007	0.0003
	lnpayment2	-0.0502	-0.0328	0.0601	0.0228
if RG4	**lnpayment1**	-0.0008	-0.0003	0.0008	0.0002
	lnpayment2	-0.0621	-0.0202	0.0672	0.0151
if privworker	**lnpayment1**	-0.0006	-0.0004	0.0007	0.0003
	lnpayment2	-0.0477	-0.0348	0.0578	0.0247
if unemploy	**lnpayment1**	-0.0006	-0.0004	0.0007	0.0003
	lnpayment2	-0.0468	-0.0355	0.0569	0.0254
if retired	**lnpayment1**	-0.0003	-0.0005	0.0003	0.0006
	lnpayment2	-0.0252	-0.0432	0.0213	0.0471
if exempt	**lnpayment1**	-0.0008	-0.0002	0.0008	0.0002
	lnpayment2	-0.0669	-0.0137	0.0682	0.0124
	HOSPITAL 3	**dF(y=1)/dx**	**dF(y=2)/dx**	**dF(y=3)/dx**	**dF(y=4)/dx**
if male	**lnpayment1**	-0.0004	-0.0005	0.0004	0.0005
	lnpayment2	-0.0298	-0.0431	0.0316	0.0414
if female	**lnpayment1**	-0.0003	-0.0005	0.0003	0.0006
	lnpayment2	-0.0251	-0.0432	0.0211	0.0472
if RG1	**lnpayment1**	-0.0002	-0.0005	-0.0001	0.0008
	lnpayment2	-0.0159	-0.0396	-0.0050	0.0606
if RG3	**lnpayment1**	-0.0003	-0.0005	0.0002	0.0006
	lnpayment2	-0.0228	-0.0428	0.0153	0.0503
if RG4	**lnpayment1**	-0.0004	-0.0005	0.0005	0.0005
	lnpayment2	-0.0326	-0.0426	0.0369	0.0384
if privworker	**lnpayment1**	-0.0003	-0.0005	0.0001	0.0007
	lnpayment2	-0.0210	-0.0423	0.0106	0.0528
if unemploy	**lnpayment1**	-0.0003	-0.0005	0.0001	0.0007
	lnpayment2	-0.0204	-0.0421	0.0088	0.0537
if retired	**lnpayment1**	-0.0001	-0.0004	-0.0004	0.0009
	lnpayment2	-0.0082	-0.0307	-0.0358	0.0748
if exempt	**lnpayment1**	-0.0005	-0.0005	0.0006	0.0004
	lnpayment2	-0.0374	-0.0410	0.0449	0.0335