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

Authors: Robin Thompson[1] and Ana Xavier[2]

1: Centre for Health Economics, University of York, UK
2: LICOS Centre for Transition Economics, Belgium

1. Background

In many countries of the former Soviet Union and Central and Eastern Europe undergoing economic transition patients are routinely asked to pay unofficially for medicines and medical supplies required for their medical treatment. They are also frequently expected to supplement state health worker salaries with unofficial monetary or non-monetary payments (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, M. 2000). The legality of these payments is not always clear. In some cases they are clearly illegal whilst in others their legality is ambiguous. 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; Chawla *et al.*, 1998; Ladbury, 1997). In Tajikistan, 70% of survey respondents stated that they expected to have to pay for health care (Mirzoev, 1999). Thompson and Witter (2000) and Ensor (2000) present typologies of these payments.

These payments are likely to be inequitable because a patient's access to services or quality of care may depend on their ability to pay. In Bulgaria the average unofficial payment made to health workers represented between 3-14% of a patient's average monthly income and the cost of an surgical procedure was 83% of average income (Delcheva *et al.*, 1997). In Kazakhstan it is common for relatives of patients requiring treatment to advertise in newspapers for monetary support. The Russia Longitudinal Monitoring Survey found that lack of money was the main reason for inability to obtain prescribed medicines and was cited by respondents twice as often in 1996 as in 1994 (Liu *et al.*, 1998). Bognar *et al.* (2000) find that income appears to be unimportant in determining payments. Some people may thus be delaying seeking 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 are estimated to represent a significant slice of the total health care expenditure and they represent a wages supplement for health workers whose wages were kept at very low levels before and after transition started (Kornai, 2000). In Kazakhstan, in 1996, they may have constituted 25-30% of the state budget considering medicines alone (Thompson and Witter, 2000).

Unofficial payments are rooted in systems of bargaining and connections inherited from the socialist system (Smith, 1973). According to Kornai (2000) and Gaal (1999a,b) the planned and rigid nature of health care provision led patients to search for mechanisms to obtain better and faster services than what they would obtain as the basic state services, more attention and a preferred doctor. Payments would buy patients a little freedom. At the same time, medical activity, although a demanding and intellectual activity, received very little financial reward, which could be increased with the unofficial payments.

The widespread existence of unofficial payments for health care is closely related to the impact of economic restructuring which has included the closure of uncompetitive state and private enterprises and increasing unemployment. The resulting decline in tax revenue and subsequent reductions in government health sector funding have meant that patients contribute to the shortfall in 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, creating a climate for unofficial payments. Chronic shortages, coupled with inadequate equipment mean that patients or their relatives are routinely asked to pay, through unofficial channels, for the medicines and other supplies required for their medical treatment and that are often scarce in hospitals as a result of tighter budget constraints. The purchasing of drugs constitutes a common source of unofficial expenditure although inpatient care appears to be the most costly item (Lewis, 2001). In Kazakhstan it is estimated that patients might be paying around US$50 for in-patient medicines (Ensor and Savelyeva, 1998).

Unofficial payments feature in countries where health workers salaries are low relative to other state and private sector professions, where delays are commonplace and the private sector (which could provide extra income) is practically non-existent. In Lithuania and Ukraine, for example, workers have waited up to three months to be paid, with reports of longer delays in Russia (Healy and McKee 1998). In Estonia, 60% of physicians reported receiving at least one non-cash gift each week and some received a monetary tip (Barr, 1996). Of those who had received a monetary tip, the average amount was around 18.5% of their monthly salary. Unofficial payments double the average gross salary of physicians in Poland (Chawla *et al.*, 1998), whereas specialist doctors in Albania can earn five times their salary through unofficial payments (Healy and McKee, 1998). In Hungary, unofficial payments constitute 62% of the net income of physicians (Kornai, 2000). The majority of unofficial payments are given to physicians. In Poland 81% of payments were paid to physicians with the rest being paid to other health workers (Chawla *et al.*, 1998). Payments may also constitute gifts or tips to health workers, which in some cases may have just a voluntary character. Nonetheless, in the Czech Republic whilst over 27% of patients gave gifts to obtain better treatment, 7% gave gifts out of fear of receiving no treatment (Masopust, 1989).

The presence of widespread corruption and minimal sanctions, for those who are caught taking such payments, fuels unofficial payments. Non-reporting by patients, weak corruption monitoring and weak enforcement of sanctions by the government and physicians' lack of accountability to a higher authority help maintain the system. Patients' lack of information may also help maintain the system of unofficial payments.

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 might 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) suggests that payments are made so as to change the attitude of providers towards the patient and adapt treatment to the patient's convenience. Kornai (2000) argues that these payments are "bribe" given by patients to doctors to ensure extra-attention, moving them up the queue thus obtaining a shorter period of waiting, a better bed or a chosen doctor. 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".

In this paper we consider a simple economic model of physician behaviour in which physicians adjust the quality of care to the level of unofficial payment paid by the patient. We examine the behaviour of the state salaried physician employed in a monopoly state acute hospital setting. Poorly paid and demotivated physicians are seen to 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. Physicians exploit their powerful position (*e.g.* better information concerning care) and the demand for higher quality by offering differing levels of service quality to paying and non-paying patients.

The model is motivated by a perception that the general quality of the health care provided by the state is poor and some patients are willing to pay unofficially in an attempt to improve the quality of care. We look at the context where there are two types of care offered to the patient – low quality and high quality care. Demand is generated as a result of the perceptions of quality and associated patients' preferences for the low and basic quality services and prices charged. The physician's maximisation problem is that of choosing the payment for the low and the high quality services given the patients' demands for those two types of treatment (in other words, the physician chooses the quality - payment combination), and in the process he obtains a monopolistic rent (Kornai, 2000).

The theory is then tested using data obtained from a survey of 1508 discharged acute hospital (surgery and trauma specialty) patients in Kazakhstan. Empirical studies on unofficial formal payments tend to provide anecdotal reports of physician behaviour with estimates of spending collected through primary surveys. There are few, if any published English language studies, which formally attempt to model and test physician behaviour within an unofficial payments context. Many anecdotal reports do however suggest that physicians exploit their monopoly position by engaging in discriminatory unofficial pricing and service differentiation.

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 and regressions of duration data. Section 6 discusses the results and concludes.

This paper contributes to the existing literature in various ways. First, 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 (measures using process and subjective measures of quality were undertaken). Previous studies were limited to answer the "whom, how much, when and to whom" of the matter. Second, we were able to gather detailed data on discharged acute hospital patients that are in general very difficult to obtain. The data are gathered for patients who had an intervention that was supposedly free of charge fact that allow us to identify clearly the amount paid unofficially and look at this only. Some of the previous studies did not distinguish between official and unofficial payments for care thus providing only a rough idea for 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 more formal health care systems elsewhere.

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 oblast/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. Incentives focused on penalties for failure rather than incentives for success (Ensor & 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. Finally, confidence in the fund collapsed 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 in-patient basis, the subject of this paper, payment was made to providers 11% of the time and 12% for surgeons. In addition, 25-42% of those who were hospitalised had to provide their own bedding, clean laundry and food, and 57% had to provide their own medicines (Sari *et al.* 2000). A decree formalising formal 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 by the World Bank in 1996 found 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 to benefits 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 with many individuals having 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, 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. Yet as the empirical results in this paper show, in reality the vast majority of patients pay for hospital care. Given that the health care system (i.e. funding) is dominated by hospital provision the number of days a patient spends in hospital is an important issue. In countries such as the UK post-hospital follow-up care has become increasingly important as length of hospital stays are reduced. In Kazakhstan, post-hospital follow-up care is weak. Anecdotal reports suggest that patients are willing to pay to stay in hospital, as they perceive that once they leave hospital follow-up care is weak.

An important issue is whether unofficial payments are made for entitled services or some enhanced level of care. Frequently the specific reason for payment is unclear. Patients may be asked to supply medicines and supplies required for their treatment because the hospitals do not have these. Or, in some cases, 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. 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 the entitled level of care or some enhanced level.

Information asymmetry coupled with endemic unofficial payments places acute hospital patients in an extremely vulnerable position. They do not know what services should be provided as part of their entitlement to state health care. Physicians though are fully aware of this entitlement. Medical standards define the scope of services that must be provided for each diagnostic category and include the scope and scale of diagnostic tests, medicines and medical supplies. Medical standards also state how many days that a patient should stay in hospital. The health worker can exploit his knowledge to obtain unofficial payments without a significant cost to himself. He can allocate state medicines and medical supplies to patients who pay unofficially. The physician has the power to keep patients in hospital or discharge early. He can dictate queues in the accident and emergency department.

3. A model of physician and patient's behaviour: an unofficial market for health care

In this section we model patients' and physicians' behaviour looking at 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, state salaried physicians might well adopt similar 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 provided by the state is perceived to be poor. As a consequence, some patients are willing to pay unofficially for services (*e.g.* medicines, surgeon) that are supposedly free in an attempt to improve the quality of care they receive. We assume that patients have different preferences for health care quality, which result in demand for quality of care that is 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 that they can exploit to obtain unofficial payments without significant cost to them (*e.g.* 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 given the lack of follow-up service provision outside of 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. Indeed, there are many anecdotal reports of state physicians adopting differential unofficial pricing strategies in the FSU and considerable evidence to suggest that patients are willing to pay unofficially for an improvement in the quality of health care (Thompson and Witter 2000).

3.1. A short review of the physician literature

The physician agency literature provides some useful insights into the behaviour and motivation of state salaried physicians employed in the Former Soviet Union (FSU) setting. Whilst the literature is predominantly written within the North American context there are a number of parallels with salaried

[1] We do not examine the decision making process of the hospital as a whole, or the hospital management team, but solely the behaviour of physicians.

state physicians working within endemic unofficial payment systems of the FSU, as state salaried physicians adopt patterns of market behaviour within state hospitals and explore some monopoly power.

Based on the existing literature the profit maximising assumption does provide a useful context where patients are willing to pay unofficially for extra quality and physicians are willing to exploit their power to provide it.[2] McGuire (2000) argues that there are not many alternatives to a profit maximising model subject to a demand. The author reviews the physician behaviour literature highlighting that many papers present no formal conception or model of the behaviour of the physician firm. Previous contributions have looked at physicians as profit maximisers setting prices for their services. In several of these models an element of monopoly power is present and explored by the physician in the context of complete information. Other studies look at supply or physician induced demand whilst some address information issues and physician motivation and objectives.

Within the profit maximising literature, several authors (*e.g.* Phelps, 1997 and Dranove and Satterthwaite, 2000) argued that location, specialty, and care quality imply that physicians are imperfect substitutes and, as such, there is an element of monopoly power with the demand curve sloping downwards. Gaynor and Gertler (1995) and Ma and McGuire (1997), in the context of perfect information, examine patient demand as a response to health care quality or some physician input. McGuire (2000) presents a model of monopolistic competition in which the physician has some market power but the patient has some alternatives. The price and quantity of physician services are found by maximising the physician's profit, subject to the constraint on patient net benefit imposed by competition with alternative physicians. McGuire's model considers an all or nothing offer to the patient, extracting all available consumer's surplus. With market power and the non-retradability of healthcare, the physician possesses the prerequisites for the exercise of first-degree price discrimination.

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 and thus support price discrimination. Focusing on the physician's self interest, Kessel (1958), in his analysis of the market for physician services in the US, 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 a consequence of utility maximisation by the individual doctor. Feldstein (1979) uses a simple monopoly model to analyse physician's pricing behaviour. Eisenberg (1986) argues that physicians are motivated by their self-interest although they may also be concerned with their patients' health. The insights of these models constitute our departing point. We believe this literature fits well with the context under analysis.

Another strand of the literature, which provides useful insights concerning unofficial payments, is that looking at corruption (see Bardhan, 1997 for a useful review). Lui (1985) presents 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 is linked to the opportunity costs of time for the individual. Myrdal (1968) has argued that corrupt officials may, instead of speeding up queues, actually cause administrative delays in order to attract more bribes. There is a strand in the corruption literature suggesting that, in the context of pervasive and cumbersome regulations in developing countries, corruption may actually improve efficiency (Bardhan 1997). This might be viewed at in terms of a coasean bargaining process in which a bureaucrat and the private agent may negotiate their way to an efficient outcome. Galasi and Kertesi (1989) model bribes for quality within socialist countries. They show how all consumers may end up

[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.

worse off when some of them pay bribes to obtain higher than official quality care. When inputs are fixed, bribery reduces the quality available to those paying fixed prices (in this case no price) and induces more corruption. In the end everybody may be paying bribes yet obtaining quality no higher than the official level (see also Kornai, 2000). We compare the results obtained in our model with those obtained by these authors.

3.2. Setting the quality - unofficial payment combination

Consider now, in the context of the parallel and unofficial market for health care described above, the demand for two competing health care interventions or processes used to treat the same condition but differing in some quality characteristic. 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 or hospital input such as patient contact time or "effort" devoted to the patient so that the low quality treatment corresponds to basic (sub-basic) consultation time whereas high quality treatment means additional (basic) doctor's *"effort"* or consultation time.

Another possible definition of treatment quality 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.[3] Shorter lengths of stay in this context therefore might be recognised as some basic (sub-basic) or low level of health care quality with longer length of stay perceived as a enhanced (basic) or high level of quality.[4]

Finally, time spent waiting before admission may also be perceived as a quality measure and the longer the wait the lower the quality of care as perceived by patients (*e.g.* Propper, 2000).

Assume now that the physician knows that the demand for his services is composed of heterogeneous consumers and some may have stronger preferences for the higher quality good.[5] The good being traded is treatment. Each consumer consumes one unit of the good in that a patient consumes the treatment only once at a time (*e.g.* one operation only). The treatment offered can be of two different qualities, or two different goods. There is the low quality treatment and there is the high quality treatment.

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.[6] We assume that the utility derived with treatment is separable in price and quality. The rationale behind it is that all consumers prefer a higher quality for a given price but a consumer with a higher θ is willing to pay more to obtain higher quality of care (measured or perceived). The utility function is thus:

[3] There are a number of reports to suggest this is the case, particularly in rural areas (Ensor and Thompson, 1999)
[4] Of course, shorter length of stay might be perceived by patients to be of higher quality, particularly for non-acute conditions. The theoretical model captures the multidimensional nature of health care by introducing a flexible quality parameter)
[5] Alternatively, some consumers may have different marginal rates of substitution between income and quality of treatment.
[6] Insights of vertical differentiation and price discrimination can be found in Tirole (1988).

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

where φ is a positive parameter describing quality. There are two types of treatment quality $\varphi = (\varphi_L, \varphi_H)$ with L and H referring respectively to the low quality and to high quality treatments. Moreover, $\varphi_L < \varphi_H$, that is, the quality associated with the basic (sub-basic) treatment is lower than the quality of the enhanced (basic) (at least as perceived by patients).[7] θ is a positive real number that can describe the taste for quality.[8] The parameter θ is distributed according to some density function, $f(\theta)$, which reflects the variation in tastes among patients, and to 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 price of treatment is represented by p. Given the two treatment types we assume that $p = (p_L, p_H)$ and $p_L < p_H$, that is, the low quality treatment is charged a lower price. If it were more expensive than no one would buy it.

A patient chooses the high quality treatment rather than the low quality treatment if

$$U_H^P \geq U_L^P > 0 \Rightarrow \theta\varphi_H - p_H \geq \theta\varphi_L - p_L > 0 \quad (2)$$

when getting a high quality treatment provides a higher utility than obtaining a low quality treatment, which provides a higher utility than no treatment ($U^P = 0$). This holds for all those for whom

$$\theta^{Hc} \geq \frac{p_H - p_L}{\varphi_H - \varphi_L} > 0 \quad (2a)$$

Patients obtain the low quality treatment whenever

$$U_L^P > 0 \Rightarrow \theta\varphi_L - p_L > 0 \quad (3)$$

which rearranging implies

$$\theta^{Lc} \geq \frac{p_L}{\varphi_L} > 0 \quad (3a)$$

Thus,

- All the patients who have preferences for quality higher than threshold θ^{Hc}, $\theta > \theta^{Hc}$, buy the high quality treatment (extra (basic) doctor's attention or longer length of stay (normal discharge) or shorter admission time).
- All the patients who have preferences for quality higher than threshold θ^{Lc}, $\theta > \theta^{Lc}$, buy the low quality treatment (basic (sub-basic) doctor's attention or shorter length of stay (early discharge) or longer admission time)

[7] For analytical simplicity we use only two discrete levels of quality. We believe this is enough for illustrative purposes. This context can be easily extended to one of a continuous range of quality levels.

[8] 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 story 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. See this context later.

- All the other patients for whom the threshold $\theta < \theta^{Lc}$ do not buy any treatment and are exclude 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. To obtain these demands one has to integrate the density function using the boundaries defined by the above critical levels of the taste parameter, θ. Thus, we have

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

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

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

Note that the demands depend on the level of unofficial payment.

The physician chooses the unofficial payments p_L and p_H so as to maximise his utility bearing in mind the above demand functions for each of the two treatment types. There is also a cost involved in the production of treatment (*e.g.* cost of physician time or the potential sanction imposed on the physician if found to be charging unofficial payments). For simplicity we assume that the costs are separable and linear. Thus, the doctor's maximisation problem is

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

And the first order conditions are:

$$\frac{\partial U^D}{\partial p_L} = D_L(p_L, p_H, .) + (p_L - c_L)\frac{\partial D_L(p_L, p_H, .)}{\partial p_L} + (p_H - c_H)\frac{\partial D_H(p_L, p_H, .)}{\partial p_L} = 0$$

$$\frac{\partial U^D}{\partial p_H} = D_H(p_L, p_H, .) + (p_H - c_H)\frac{\partial D_H(p_L, p_H, .)}{\partial p_H} + (p_L - c_L)\frac{\partial D_L(p_L, p_H, .)}{\partial p_H} = 0$$

Rearranging the terms we have that

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

with $i,j = L,H$ and $i \neq j$. This just shows that the physician as a monopolist chooses the unofficial payments so that marginal revenue equals marginal costs $mr = mc$.

Further rearranging (see Appendix 1) gives us

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

as

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

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; and 2) the cross-elasticity and the mark-up for the other good. If the treatments are substitutes (quite realistic as patients can have only one type of treatment), then the cross-elasticity, ε^D_{ji}, is negative. Thus, the mark-up price for good *i* is greater than 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 define a specific density function. We assume that patients' preferences over quality are distributed uniformly, that is, θ follows a uniform distribution.

Using this distribution 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 consequently the demands simplified to

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

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

Using the expressions just obtained for D_H and D_L to compute the first order conditions of before we obtain:

$$\frac{\partial U^D}{\partial p_L} = \frac{N}{\theta^M}\left(\frac{p_H - p_L}{\varphi_H - \varphi_L} - \frac{p_L}{\varphi_L}\right) + (p_L - c_L)\left(-\frac{N}{\theta^M(\varphi_H - \varphi_L)} - \frac{N}{\theta^M \varphi_L}\right) + (p_H - c_H)\left[\frac{N}{\theta^M(\varphi_H - \varphi_L)}\right] = 0$$

(A)

$$\frac{\partial U^D}{\partial p_H} = N\left(1-\frac{p_H-p_L}{\theta^M(\varphi_H-\varphi_L)}\right)+(p_H-c_H)\left(-\frac{N}{\theta^M(\varphi_H-\varphi_L)}\right)+(p_L-c_L)\left[\frac{N}{\theta^M(\varphi_H-\varphi_L)}\right]=0 \quad \text{(B)}$$

Solving (B) and (A) with respect to the unofficial payments p_H and p_L we find the following system of equations:

$$\begin{cases} p_L\left(\frac{2N}{\theta^M(\varphi_H-\varphi_L)}+\frac{2N}{\theta^M\varphi_L}\right)=\frac{2Np_H}{\theta^M(\varphi_H-\varphi_L)}+\frac{c_LN}{\theta^M(\varphi_H-\varphi_L)}+\frac{c_LN}{\theta^M\varphi_L}-\frac{c_HN}{\theta^M(\varphi_H-\varphi_L)} \\ p_H=p_L+\frac{\theta^M(\varphi_H-\varphi_L)}{2}+\frac{c_H-c_L}{2} \end{cases}$$

and the optimal values of p_L and p_H are

$$\begin{cases} p_L=\frac{\theta^M\varphi_L+c_L}{2} \\ p_H=\frac{\theta^M\varphi_H+c_H}{2} \end{cases} \quad (9)$$

which is equivalent to (rearranging and dropping the subscripts)

$$\varphi=\frac{2p-c}{\theta^M} \quad (9a)$$

Expression (9a) shows that the quality dimension is positively related to the payment and negatively related to the cost of providing the treatment type. To higher quality activities corresponds a higher price and is associated a higher cost. This relationship is the basis for the empirical analysis that follows.[9] If our results show a positive association between payment and quality then the above model can be a good representation of patients and physicians' behaviour in what concerns unofficial payments.

3.3. Measuring quality

Defining "quality" in health care is generally unproblematic. Campbell *et al.* (2000) suggest that there are two principal dimensions of quality of care for individual patients: access and effectiveness. The measurement of quality in healthcare is however 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". This feature of healthcare means leads to contractibility issues (McGuire 2000). One of the most difficult measurable indicators of quality is that of health worker "effort". It is known that many unofficial payments are given simply to "motivate" physicians to provide more

[9] 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).

"effort". McGuire (2000) suggests that "the care or effort that a doctor puts into a decision or treatment matter to the patient but it is difficult to incorporate into a payment system." The problem is that physician effort may not be observable.

A more tangible notion of quality might therefore be *time*. The problems with contracting on outcomes have meant that process indicators, such as "*time*", have traditionally been used to monitor and pay providers. Lengths of stay and waiting times for inpatient and outpatient appointments are both commonly used to monitor hospital performance. In recommending how patients should judge their doctors, McCall (1996), a physician, states "the amount of time a doctor spends interviewing you, examining you, and explaining things reflects how genuinely concerned that doctor is for your welfare". In summary, the study of unofficial payments illustrates the problems of contractibility in any health care system because of the difficulties associated with measurement of outcomes. In what follows we make use of two measures of health care quality: 1) admission waiting time faced by the patient and 2) a patient's length of stay in hospital. They are explained below.

3.3.1. Waiting time as a measure of quality

The demand for health care sketched previously is shaped by patients' perceptions of and preferences for quality of health care. A number of authors model demand where public and private care exist (Goddard *et al.*, 1995; Martin and Smith, 1999; Propper, 2000) examining the impact of income, price and quality on decisions to demand/use public or private care. Waiting time is frequently used as a quality proxy in studies of health service demand and long waits have been seen by the general population as an unsatisfactory characteristic of the NHS (Bosanquet, 1988). For example, Propper (2000) models quality using a waiting time parameter, with individuals varying in their valuation of this quality parameter. Such models provide a useful starting point in the analysis of unofficial payments for quality in a transition country such as Kazakhstan. Whilst little official private hospital provision exists in Kazakhstan, so that patients' decision is not between the private and the public as in those models, state health care workers are engaged in quality enhancing health care activities within the state hospital structure and patients entering the state structure therefore have a choice of differing quality services.

Moreover, our choice is based on previous studies of unofficial payments. Gaal (1999a,b) suggests that patients used such payments to obtain faster services than what they would obtain otherwise. Also Lewis *et al.* (2001) find that patients paid to save time. Kornai (2000) argues that patients make these payments so as to move up the queue and face a shorter waiting. In other words patients are paying an extra fee for immediate referral. Waiting time could thus be seen as a measure of health care service quality (the shorter the time the higher the quality of care). We test this hypothesis in this paper. A negative coefficient for payment is expected if waiting time is the inverse of a higher quality of care.

3.3.2. Length of hospital stay (LOS)

The empirical analysis discussed in this paper also focuses on 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 that may not necessarily agree with patients perceptions): lengths of stay have 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 care in hospitals and a much wider availability of community care and local facilities to provide follow up care have made this possible.[10] Still, concerns are raised about early discharge and post-surgical complications and hospital readmission. In the transition world the situation is quite

[10] In this context, and for reimbursement purposes, LOS has been used as a proxy for resource use and a longer LOS, "other things being equal", may indicate technical inefficiency.

different. Health facilities are limited in number and often located in cities often far from great part of the population. Post-hospital follow-up is poor or inexistent and transport to hospital is very 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 readmission as doctors monitor the patient condition for longer. Thus, a longer LOS may be perceived by patients in the transition world as better quality of care.

In order to isolate the association between payment and LOS it is important to understand and control for other factors. Martin and Smith, writing in 1996, highlighted that little is known about the causes in LOS variations in England, due to the shortage of publicly available data about patients. They conclude from the international literature that there are two broad determinants of length of stay: patient characteristics and hospital characteristics. Studies typically find that patient age and severity are very important determinants of length of stay (Godfarb *et al*., 1983). Patients of lower socio-economic status have longer lengths of stay (Epstein *et al.*, 1990) and DRG status makes a significant contribution towards explaining length of stay (Cairns and Munroe, 1992). We will therefore take these into consideration in the empirical analysis that follows.

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). Xiao *et al.* (1997) find that organisation of discharge and unplanned admissions to be both significant determinants of length of stay. Westert *et al.* (1993) confirmed the importance of the hospital, finding that variations in length of stay between doctors in the same hospital were much smaller than variations between hospitals. Burns and Wholey (1991) found that, although both patient and hospital characteristics were important in explaining length of stay amongst acute inpatients, length of stay was also positively associated with physician workload.

Also the means by which inpatient services are financed can have a direct effect on average length of stay. If a hospital is reimbursed for the costs of an inpatient stay on the basis of a constant *per diem* fee, they have an incentive to keep patients for lengthy periods. Alternatively, hospital reimbursement policies can serve to limit length of stay, for example prospective payments for inpatient stay in the US. Ensor and Thompson (1999) highlight the perverse funding arrangements in the countries of the FSU whereby hospitals were funded according to normative criteria of beds and bed-days. This encouraged long lengths of stay to maintain high occupancy rates and bed-days. However, the traditional method of normative funding has been replaced in many countries of the FSU and CEE by case-based payments, with many countries introducing prospective payment systems based on diagnostic related groups. These payment systems have introduced new incentives to discharge patients promptly. Medical standards supporting the system however specify the number of days a patient should stay in hospital.

4. Data and methods

The data used in this analysis come from a randomly selected survey of 1508 discharge surgical and trauma patients treated in three hospitals in Almaty City, Kazakhstan in 1999. Patients were asked about their experience in hospital and related expenditure, as well as their socio-economic status. Given the sensitivity of the survey patients were surveyed in their homes.

Two indicators – number of minutes spent in the Admission Department (AD) and number of days spent in hospital - were chosen to explore the relationship between unofficial payments and the process of care. 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). These codes were also chosen because of the entitlement of individuals suffering from one of these codes to free care. 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.

4.1. Dependent Variables

4.1.1 Time spent in Admission Department

The first process indicator tested as the dependent variable was the total time that patients spent in the Admission Department. This includes waiting time prior to any clinical intervention. Table 1 shows the mean number of minutes that patients spent in the Admission Department by hospital department and resource group. 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 the distribution of patients by resource group and department. A large number of patients fall into resource groups two and three. There are few patients coded as group 1 in the trauma departments.

Table 1: Minutes spent in admission department by ward and resource group (mean, s. d., n)

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

As the time for admission is positively skewed a log transformation was performed on the variable. Furthermore, the variable was standardised by ICD10 code. A T-test of time spent in the Admission Department (lnadmwait) by Admission Department payment shows no significant difference between those making a payment and those making no payment in the Admission Department.

4.1.2. Length of hospital stay (days)

Table 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 & 2 (surgery specialty) and Hospital 3 (trauma specialty). In the Hospitals 1 & 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. Furthermore, the variable was standardised by ICD10 code. A T-test of number of days spent in hospital (lnlos) by Admission Department payment shows that those patients who make a payment in the Admission Department have significantly higher length of stays than do those patients who make no payment.

Table 2: Days spent in hospital by ward and resource group (mean, s.d ,n)

Hospital	1	2	3	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

4.2. Independent Variables

The independent variables include illness severity (proxied by diagnostic resource groups), age, occupation, exemption status and income controlling for disease severity and social-economic variables as suggested in previous models looking at demand for care and length of stay in hospital. Due to data limitations we are unable to use information on hospital characteristics such as number of doctors and number of beds. Nevertheless, we control for differences across hospitals (*e.g.* differences in the number of doctors) using a hospital dummy. We also define a department - unofficial payment interaction variable. As medical standards apply equally to all the hospitals and the reimbursement method is similar we believe we do not need to control for these.

As we wish to study the relationship between quality of care obtained and unofficial payment we obtained information on payment. The general idea developed through the interviewing process is that payment negotiation takes place as soon as the patient arrives to the hospital in the Admission Department (AD) and before admission and treatment take place (e.g. patients seek to reduce admission time by paying). Negotiation takes place and the patients agree to a certain amount for a certain quality level. However, although negotiation and agreement takes place in the AD and before treatment, some patients do not (potentially because they cannot afford) pay all at once and while in the AD so that some pay after admission takes place. As a result and given the information gathered with questionnaires we consider two unofficial payments variables. The first (**Pay1**) is the amount of unofficial payment made by the individual before admission in the AD. The second (**Pay2**) is the amount of unofficial payment made after admission takes place and when already on the ward. Table 3 provides a list of the dependent and independent variables used in the models and Table 4 presents some summary statistics.

The econometric models described and tested below explore the relationship between unofficial payment and (1) number of minutes spent in the Admission Department and (2) length of stay (number of days spent in hospital). We first explore this relationship by undertaking simple linear regression analysis (OLS regressions). Two models are tested using the two different process indicators as the dependent variable. In the first model the dependent variable used is (logged) minutes spent in the admission department. The second model specifies the (logged) number of days spent in hospital as the dependent variable. Note that the relationship between the two measures of quality and the two unofficial payments is studied (see explanation further on).

Table 3: Variables used in the empirical analysis

Variable code	***Description***
Dependent	
Lnadmwait	Number of (ln) minutes an individual spends in the Admission Department
Lnlos	Number of (ln) days an individual spends in hospital
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	Household adjusted (ln) monthly consumption expenditure (income proxy) in local currency (KZT)
Payment variables:	
Lnpay1 Lnpay2 Pay_1 Pay_2	Log of amount of KZT paid in the Admission Dept Log of amount of KZT paid in the ward Pay_1 and Pay_2 (binary variables: 1=patient paid and 0=patient did not pay)
Hospital specific variables	
CCH, HAC, Trauma	Hospitals 1, 2 and 3 (Dummy variables)
HAC_1, Trauma_1 HAC_2, Trauma_2	Hospital payment interactions

Table 4: Descriptive statistics

	No. Obs.	**Mean**	**Standard. Deviation**	**Min**	**Max**
Dependent variables					
Admwait	1368	555.117	5.276.498	3	720
Los	1496	1.355.615	13.081	1	90
Lnadmwait	1368	3.782.068	.7491177	1.791.759	6.583.409
Lnlos	1496	2.591.974	.6216512	1.386.294	4.532.599
Independent variables					
Age	1508	4.298.939	1.800.365	5	89
Male	1508	.505305	.5001377	0	1
Income	1494	20683.01	14745.39	1.130.348	144000
Lnincome	1494	9.725.804	.6598959	7.030.281	1.187.757
Pay1	1452	2.949.345	6.145.867	0	52000
Pay2	1483	1.796.129	4.743.295	0	35000
Lnpay1	1452	3.397.425	3.528.209	1.098.612	1.085.906
Lnpay2	1483	2.537.555	3.025.274	1.098.612	1.046.319
Student	1508	.1332891	.3399997	0	1
Unemploy	1508	.183687	.3873572	0	1
Statwork	1508	.1637931	.3702105	0	1
Privwork	1508	.117374	.3219722	0	1
Selfwork	1508	.0364721	.187524	0	1
Retired	1508	.2533156	.4350544	0	1
Houswife	1508	.102122	.3029092	0	1
Exempt	1508	.2838196	.4509999	0	1

The regression diagnostics used to assess the OLS models include firstly, the calculation of variance inflation factors (VIF) to assess the predictors for collinearity. Second, we address potential

heteroskedasticity by specifying the Huber/White/sandwich estimator of variance instead of the traditional calculation. We compute the Ramsey reset test for each of the models, which uses the predicted values of the dependent variable to the power of 2, 3 and 4 in the regression and tests the significance of the coefficient estimates of those three extra regressors. This test amounts to testing $y = xb + zt + u$ (where z stands for the three powers of the predicted values of y) and then testing $t=0$. This is a test for the general specification of the relationship estimated namely in terms of omitted variables.

Given that both process indicators we regress on payments are classified as "*time*" data we develop the exploration by focusing on duration models.[11] These models are increasingly being used in health econometrics to study a range of issues such as the impact of tax on starting and quitting smoking (Forster and Jones, 2001) and the impact of hospital volume and cost on length of stay (Hamilton and Hamilton 1997). When pursuing the duration analysis we take heteroskedasticity into account in a similar way and a Wald test for omitted variables is used. Where appropriate we also conduct a test for unobserved heterogeneity and compared the different models using the log-likelihood and akaike criteria (see below).

5. Results

5.1. OLS Regressions

5.1.1. Time spent in the Admission Department (minutes)

The first model examines the relationship between the dependent variable defined as time spent in the admission department and the unofficial payments. More specifically, time spent in the admission department is defined as "*the total time a patient spends in the admission department, from the time of admission to hospital to transfer to theatre or the ward."*

The model can formally be specified, in its most general form, as:

$$\ln admwait = \beta_0 + \beta_1 RG + \beta_2 age + \beta_3 age^2 + \beta_4 gender + \beta_5 \ln income + \beta_6 \ln payment + \beta_7 occupation + \beta_8 Exemption + \beta_9 hospital + \beta_{10} hospital * \ln payment + e$$

The reference hospital is hospital 1, a surgical provider, and RG2 the reference resource group. We begin by modelling the unofficial payment as a binary variable thus exploring whether and the extent to which, the act of paying is associated with the time waiting to be admitted (see Table 5 – all tables are presented in Annex 2). Given the high correlation between **Age** and **Age_sq** we also run a model specification dropping the latter (Table 6). This significantly improves the VIFs without any changes in the coefficient estimates and their significance. We then proceed by specifying the continuous variable (amount of payment given to the medical staff) so as to check whether and in which way the amount of payment related to waiting time for admission (Table 7 and 8 – without **Age_sq**). We specify the model using pooled and unpooled data for each of the hospitals to assess differences across hospitals. We also specify a department interactions model across the pooled data using the continuous payment variable (Table 7).

Note that we use both **Pay1**, the payment in the admission department, and **Pay2** the amount paid on the ward. As said, the reality of unofficial payments is such that some patients, although agreeing to pay while in admission department and before admission, decide to pay / pay for admission later, while in the ward, as they may not have the required amount ready at the moment of admission. We keep the type of payment separate so as to distinguish each payment's effect and for endogeneity reasons (see below).

[11] See Jones (2000) for a brief introduction to duration models in health econometrics

We would expect to see a negative association between payment and admission time

The robust coefficient estimates in Table 5 show that for the pooled model specification the act of paying in the admission department (AD) – **Pay_1** - does not appear to be related to the admission time. However, when we run the hospital specific regressions payment in the AD is negatively and significantly associated with a shorter time for admission in Hospital 1, which provides surgery. This model also appears to be well specified according to the Ramsey Reset test. Patients from hospital 1 spend on average more time in the AD than do patients from hospital 3 (**Trauma**) and hospital 2 (**Hac**). If surgical patients wait longer in general than trauma patients do, then they may perceive it to be worth making a payment in an attempt to decrease admission time.

Paying in the hospital ward – **Pay_2** - appears to be positively associated with waiting time for admission when in the pooled model although this does not pass the Ramsey specification test. Further inspection shows that this is the case in hospital 3, the trauma hospital. Either those patients that agree to pay but pay after admission "loose" time in the AD in the process of bargaining. Or, because this variable may potentially include payments made in the ward for reasons either than admission time, it is not necessarily the case that a negative relationship should be observed. Further, both the pooled and the trauma hospital models may be misspecified so that this positive association must be seen with caution.

Examination of the amount paid in the AD (**lnpay1**) and in the ward (**lnpay2**) and its relation with admission time (Tables 7 and 8) indicates that the amount of unofficial payment paid in both places is significantly associated with a lower waiting time for admission in the case of hospital 1, the surgical provider (see Table 8 below). This model is well specified. Analysis of the association between payments and admission time in hospital 3, the trauma provider, shows a reversal of the sign with those making an unofficial payment spending more time in the admission department. Nonetheless the model appears to be misspecified so that again this positive association must be seen with caution.

When looking at socio-economic factors, we find that **retired** individuals (reference group state workers) wait longer for admission to surgery in hospital 1 and less for admission to the trauma units in hospital 3. **Students** have shorter admission waits for surgery in hospital 1. Individuals coded as part of resource group 3 – **rg3** - also appear to wait longer to be admitted to the trauma hospital. Income – **lnincome** - is positively associated with waiting time for surgery in hospital 1 and negatively associated with waiting time for trauma. If income is a proxy for health status then one could expect that those richer because healthier thus wait longer.

Finally, we test for potential endogeneity of the payment variables. [12] Endogeneity problem may arise from the fact that patients once experiencing the wait in the AD (defined as above) change their preferences concerning waiting and therefore the payment they make. In other words **Pay2** may be endogenous.[13] In order to test for endogeneity we regress the potential endogenous variable on all the other explanatory variables plus any other variables they may explain payment. We compute the predicted residuals from this regression and introduced them into the original regression so as to establish the significance of the corresponding coefficient estimate. An estimate that is statistically significantly different from zero may suggest that endogeneity is in place in which case the variable has to be instrumented for and a two stage least squares (2SLS) regression run. Endogeneity does not appear to constitute a problem when admission time is analysed as we find that payment is not endogenous. As such we can conclude that higher payment appears to reduce admission time for surgery.

[12] 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. Nevertheless, for diagnostic rigour and to go a step further we test for potential of endogeneity.

[13] The definition of Pay1 makes it exogenous.

Table 8: Admission time regressed on amount paid as unofficial payment (without age_sq)

	Pooled Continuous Pay	Hospital 1 Continuous Pay	Hospital 2 Continuous Pay	Hospital 3 Continuous Pay	Pooled Continuous Pay Interactions
Admission Time	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Trauma	-0.0495				-0.3359*
Hac	-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
Lnpay1	0.0115***	-0.0436*	0.0060	0.0382*	-0.0270***
Hac*lnpay1					0.0322***
Trauma*lnpay1					0.0641*
Lnpay2	0.0153**	-0.0218***	0.0142	0.0360*	-0.0189***
Hac*lnpay2					0.0354**
Trauma*lnpay2					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	F(3,226)=0.27	F(3,534)=0.75	F(3,491)=2.77	F(3,1283)=2.85
	Prob>F=0.0097	Prob>F=0.8443	Prob>F=0.5238	Prob>F=0.0412	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. VIFs are not far from 1.

Therefore, we can conclude that in acute urban hospitals paying in the AD is associated with shorter waits in the Admission Department for surgery so that the higher the payment the shorter the admission times for surgery. Moreover, patients admitted to Hospital 1 appeared to have paid both in the AD and later so as to reduce the time for admission to surgery.

Hence, if time spent in the admission department is considered a proxy for quality of care - the lower the time spent in the admission department the higher the quality - then the results just obtained support the theoretical model developed previously: patients pay unofficial 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.

5.1.2. Length of hospital stay (days)

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

$$\ln los = \beta_0 + \beta_1 RG + \beta_2 age + \beta_3 age^2 + \beta_4 gender + \beta_5 \ln income + \beta_6 \ln payment + \beta_7 occupation + \beta_8 Exemption + \beta_9 hospital + \beta_{10} hospital * \ln payment + e$$

Again we make use of both types of unofficial payments: in the AD and on the ward as explanatory variables. The rationale is that of before: patients pay / agree to pay for all the care quality (waiting time and LOS) they are to receive in their first encounter with the hospital staff, that is, whilst in the AD and before admission, but some patients may however need / wish to pay in instalments. Thus, we decided to take both payment variables into consideration also when looking at hospital LOS.

We also consider the model without **Age-sq** due to the high correlation between age and the latter and this improves the VIFs without impacting on the significance of the coefficient estimates of the model.

Tables 9 and 10 show the robust results of the regression of LOS in hospital and the act of paying unofficial payments (binary) in the AD and in the ward. Patients admitted to hospital 1 stayed in hospital less time than those going to the other hospitals (**Hac** and **trauma**) with the trauma hospital registering the longer LOS. Patients that pay unofficially (in the AD or in the ward) have a longer stay in hospital, that is, **pay_1** (especially in hospital 3) and **pay_2** (especially for hospitals 1 and 3) are positively and significantly associated with a longer LOS.

Table 12: LOS regressed on amount of payment (continuous) without age squared.

	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.**
Trauma	0.6653*				0.5586*
Hac	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*
Lnpay1	0.0199*	0.0029	0.0001	0.0503*	0.0103***
Hac*lnpay1					-0.0095
Trauma*lnpay1					0.0383*
Lnpay2	0.0418*	0.0496*	0.0198*	0.0599*	0.0554*
Hac*lnpay2					-0.0342*
Trauma*lnpay2					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	F(3,263)=1.03	F(3,567)=5.05	F(3,537)=1.46	F(3,1399)=1.92
	Prob>F=0.5431	Prob>F=0.3784	Prob>F=0.0019	Prob>F=0.2240	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. Vifs are not far from 1.

When looking at hospitals 1 and 2 it is mainly the ward payment that appears to be related to LOS as would be expected. Examination of the amount paid in the AD or in the ward and its relation to LOS (Tables 11 and 12) shows that the amount paid is positively and significantly associated with a longer stay in hospital. Both **Lnpay1** and **Lnpay2** are positively related to LOS (in all hospitals but with a lower association registered for hospital 2 and a stronger association found for the trauma hospital).

We also find that **age** is positively associated with length of stay. Housewives also spend a longer time in hospital. Men appear to stay longer in hospital than women do but s**tudents** stay less long. Patients belonging to diagnostic groups **RG3** and **RG4** spend the longest in hospital. Income, which may reflect health status, is negatively related to LOS so that the less healthy stay longer in hospital.

Note that the R^2 values are quite high and most models pass the Ramsey Reset test for general specification.

Finally, we check for potential endogeneity of the unofficial payment made by patients when in the ward – **pay2**. Once in the hospital ward, already experiencing a stay in hospital, patients may change their perceptions over quality and as a consequence the payment they may decide to make in the ward. By definition **pay1** is exogenous as it takes place before the patients experiences hospital stay.

Testing for potential endogeneity of the payment in the ward variable (**Lnpay2**) as previously explained we find that it may take place in the context of the pooled model and the trauma hospital. Using a two-stage least squares estimation and instrumenting lnpay2 with all 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 length of stay in the trauma hospital (see Table 13 in appendix 2).[14]

We can therefore conclude that in acute surgery and trauma hospitals in urban Kazakhstan, paying unofficially in the admission department and especially paying in the ward are related to a longer length of stay in hospital. And the bigger the payment made the longer is the stay, especially in trauma hospital. Therefore, if length of say is a proxy for quality, which may clearly be the case in the context of Kazakhstan where post-hospital treatment is virtually non-existent, transport to hospital is quite limited and expensive and thus increased stay in hospital is reassuring, then it can be suggested that patients are indeed paying to improve the quality of care they receive. Or, similarly, if they do not pay they may be discharged too early and thus patients pay so as to have the correct LOS for their condition.

5.2. Duration analysis

We now re-examine the relationship between LOS and unofficial payments using spell duration data. Detailed examination of the data on LOS shows this to be positively skewed and reaching 90 days, a rather large number of hospital bed days so that duration analysis may provide a good understanding of the data. Hence, we analyse patients' discharge, between 1999 and 2000, looking at the hazard of leaving hospital at each point in time given that one is still in hospital.

At first sight, because LOS is measured in days and the maximum number of days is 90 the data may be considered discrete time data and so we use discrete time analysis estimating a Cloglog hazard function. Nevertheless, we believe that the data may also be approximated by a continuous time distribution. We therefore establish the necessary comparisons with the hazard function using a set of functions

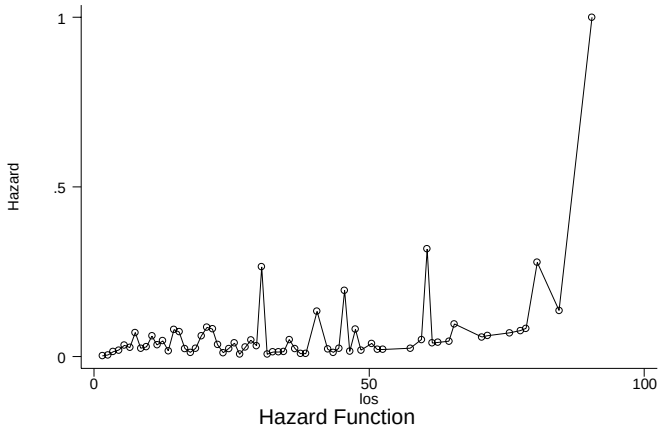
[14] Note that using instruments in a cross-section context is not straightforward: instruments may be limited especially when using a survey questionnaire. Therefore, some authors argue that the initial OLS analysis is still the "first-best" analysis.

(Exponential, Gompertz, Weibull, Cox, and Generalised Gamma[15]). Briefly, the exponential, the Weibull and the Gompertz are parameterised as proportional hazards. The exponential assumes the hazard rate is constant over time whereas the other two assume the hazard rate is either monotonically increasing or decreasing. The Cox also assumes a proportional hazard function but is semi-parametric hazard function. The Loglogistic or lognormal function assumes hazard rates that are not always increasing or decreasing. The Generalised Gamma function assumes a more flexible expression that can be reduced to the Weibull or the lognormal depending on a parameter kappa.

We start by examining the plot of the non-parametric hazard (Figure 1 below and Figures 2 and 3 in appendix 2), defined as $h_t = d_t / n_t$ (and assuming the covariates are zero), where d stands for the number of those discharged after t periods (days in our case) and n represents the number of those still in hospital after t days.

At first sight the Hazard function just plotted may be considered either constant or increasing. With this in mind we should choose a specification of the baseline hazard, which allows for one of these two characteristics. As such, the log-logistic model may not be appropriate and as a consequence we disregard this specification. We also look at whether the hazard function is different according to gender and found no evidence of this.

Figure 1: Hazard function for length of hospital stay



The results for the continuous and discrete time models are presented in Tables 14 to 19. With the exception of the generalised gamma and the discrete time models all the results are coefficients rather than hazard ratios. We account for potential heteroskedasticity by using the Huber/White/sandwich estimator of variance instead of the traditional variance matrix and produce robust estimations. Finally, we also take into account observation level frailty or heterogeneity. The estimated model (with frailty or heterogeneity) will, in addition to the standard parameter estimates, produces an estimate of the variance of the frailties (parameter theta on the tables) and a likelihood-ratio test of the null hypothesis that this variance is zero. If the null hypothesis is true, the model reduces to the model without frailty.

It can be seen that in the hazard of leaving hospital is lower in the trauma hospital as compared to hospital 2 and as compared to hospital 1. Age decreases the hazard of leaving hospital. Housewives and retired individuals (in the continuous time models) also have a lower hazard. Males have a lower hazard than women do. Privwork and selfwork (in the continuous time models) increase the hazard of leaving hospital.

[15] See Stata 7 Reference Manual, 2001 for a detailed description of these duration models.

With respect to unofficial payments and their relation with the hazard of leaving hospital we find that paying in the admission department, as well as paying in the ward, is associated with a reduced hazard of leaving hospital.

The gamma parameter in the Gompertz and the p-parameter in the Weibull are positive and significant indicating a positively increasing hazard ratio. Looking at the coefficient of **logd** it can be seen that this is positive and significant meaning that the hazard ratio is positive and increasing over time. The **kappa** parameter of the generalised gamma is positive and significantly different from zero suggesting that the log-normal specification is, probably, not an adequate model to fit the data.

Note that the Cox model does not pass the test for proportional hazard as the *Prob>chi2=0,* and both the Cox and the Gompertz do not pass the Wald test for the specification of the model namely for the omission of variables. The generalised gamma, the Weibull and the Exponential all pass the Wald test. This last test consists of computing the predicted values of the dependent variable using a linear combination of the covariates, and introducing the squared values of these into the models and testing their significance. If they are significant the model may be badly specified.

The log-likelihood ratio test for the existence of unobserved heterogeneity suggests that this does exist (the coefficient estimates for the parameter theta are significant) and thus should be accounted for in the estimation process.

We then compare the models using two methods: 1) using the log-likelihood criteria and choosing the one with the highest log likelihood (because it is a negative value we must choose the one closest to zero) and 2) the Akaike criterion whereby we compute a value given by the following expression - *2(log-likelihood) + 2(c + p + 1)*, where *c* is the number of covariates and *p* the number of ancillary parameters, and choose the model with the lowest Akaike value.

We rank the continuous time models from 1 to 5 with 1 being the best in terms of the two criteria as well as the two discrete models using the symbol *. Looking at Table 20A without accounting for observation heterogeneity (frailty) the Generalised gamma model would be preferred model both according to the log likelihood and the AIC criteria, followed closely by the Weibull, then the Gompertz, the Exponential and the Cox. The generalised gamma model and the Weibull also pass the Wald specification test (whereas the Gompertz and the Cox do not). The discrete time Cloglog model does however perform better than any of the continuous time models do according to the same criteria. When frailty is taken into account we find that the Weibull model ranks first followed by the Generalised gamma, the Gompertz and the Exponential. It does however appear to be misspecified according to the Wald test. Hence, the Generalised gamma may still be the preferred specification. The discrete time Cloglog model with frailty still ranks best.

Finally, it can be seen that the results obtained with the duration analysis are consistent with those obtained with the simple OLS analysis namely when we consider the association between payment and stay in hospital.

6. Preliminary Discussion and conclusions

In this model we use theoretical and empirical analysis to investigate unofficial payments for health care in the context of transition namely in Kazakhstan. In the theoretical model discussed above we present an informal market for health care quality that takes place within state facilities. Patients given their preferences for quality make a payment to a health worker to improve health care quality to be received (e.g. reduce the time spent in the AD). On the supply-side physicians may have enough ability to manipulate queues, decide upon resource use (e.g. theatre and bed use) and treatment. They exploit their

monopoly position the demand for higher quality 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. The existence and magnitude of payments is also dependent on the level of corruption in the department, and more specifically the individual physician. In a fluid situation, such is the hospital AD, patients and physicians are likely to have to make rapid decisions based on the information they have.

We thus explored whether, other things being equal, patients in Kazakhstan are paying unofficially to see the quality of care they receive improved. We investigate whether patients would want to spend less time in the admission department and whether they would want to spend more time in hospital. It is likely that the acutely ill patient relies on the physician to address his health care needs promptly and effectively (more than likely not knowing what this entails). Yet, one might also assume that patients would want to be processed quicker in the admission department for example. We might also expect those with a higher value of time to be willing to pay to reduce time-spent waiting in the admission department and LOS. This view is supported by empirical studies (e.g. Propper 2000; Bishai *et al.* 2000.). Also it is likely 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 and the hospital stay.

We conduct both OLS and duration analysis. The results presented show 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. Indeed, the empirical analysis suggests that: 1) unofficial payments are associated with longer length of stay in all hospitals and 2) paying is associated with a reduction in the admission time for surgery in hospital 1 (surgery). 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. 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 are not paying specifically to reduce time spent in the AD or time spent in hospital in that they may be paying for reassurance and increased physician "effort". In that case these two indicators are proxies for quality (although anedoctal reports suggest that patients may be paying to reduce the wait for admission). The problems of measurement of health care quality are well known (McGuire 2000). We do however find evidence of a strong association between payment and process more specifically that patients that make an unofficial payment receive different health care than that of those that do not pay anything.

Some further questions are left for future research. Further analysis could use subjective measures of health care quality (e.g. a categorical variable used by patients to classify the quality of care received and compare the results with the ones of this paper. Further theoretical analysis of unofficial payments using bargaining models might also proved to be useful.

Reading

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

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

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

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

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

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

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

Cannoodt, L., 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., Kawiorska, D. (1998) *Financing health services in Poland: new evidence on private expenditures.* Health Economics 7: 337-346.

Delcheva, E. Balabanova, D. 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 willingness to pay for laproscopic cholycystectomy.* Social Science and Medicine. Vol 44:5. 699-707.

Dranove, D., & 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., Rittmann, J. *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., 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. Vol 66: (2), 182-201.

Folland, S., Goodman, A., Stano, M. (1993). *The economics of health and health care.* Macmillan Press.

Forster, M., Jones, A. (2001) *Starting and quitting smoking.* Journal of the Royal Statistical Society.

Gaal (1999b) *Under-the-table payment and health care reform in Hungary.* Unpublished paper. Health Services Management Training centre, Semmelweis University of medicine, Budapest, Hungary.

Gaal, P. (1999a). *Informal payments in the Hungarian health services.* Unpublished paper. Health Services Management Training centre, Semmelweis University of medicine, Budapest, Hungary.

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

Gaynor, M. (1994). *Issues in the industrial organisation of the market for physician services.* The Journal of Economics and Management strategy 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., 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., 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 and McKee (1997) *Health sector reform in Central and Eastern Europe.* Health Policy and Planning 12 (4) 286-295.

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

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.* Unpublished paper for the Festschrift in honour of George Soros.

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

Lewis, M. (2000), *Who is paying for health care in Eastern Europe and Central Asia?* World Bank Publication.

Liu, Y., Rao, K., 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. & McGuire, T. (1997). *Optimal health insurance and provider payment.* American Economic Review. 87(4): 685-704.

Martin, S. and Smith, P. (1995). *Modelling waiting times for elective surgery.* Occasional Paper, Centre for Health Economics, University of York.

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*. Chapter 9 in Handbook of Health Economics, Volume 1, edited by Culyer, A.J. and Newhouse, J.P. Elsevier Science.

Mirzoev, T. (1999). *Corruption in Tajikistan as seen by the private sector.* Unpublished paper. Budapest, Hungary.

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

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

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

Ruffin, R., Leigh,D. (1973). *Charity, competition, and the pricing of doctors' services.* The Journal of Human resources, VIII: 2.

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

Sari, N., Langenbrunner, J., 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 manual: Q-St.* Stata Corporation

Thompson, R., Witter, S. (2000), *Informal payments in transition economies: implications fir 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 (1987) Intermediate Economics. Norton, New York.

Westert, G., Niebor,A., 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. Unpublished paper

World Bank (2000b) "Health" chapter in Czech Republic: Public Expenditure Review. Unpublished paper.

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

Appendix 1

The doctor's maximisation problem is

$$\max_{p_E,p_B} U^D = p_B D_B(p_B,p_E,..) + p_E D_E(p_B,p_E,..) - c_B D_B(p_B,p_E,..) - c_E D_E(p_B,p_E,..) =$$
$$= (p_B - c_B) D_B(p_B,p_E,..) + (p_E - c_E) D_E(p_B,p_E,..)$$

And the first order conditions are:

$$\frac{\partial U^D}{\partial p_B} = D_B(p_B,p_E,..) + (p_B - c_B)\frac{\partial D_B(p_B,p_E,..)}{\partial p_B} + (p_E - c_E)\frac{\partial D_E(p_B,p_E,..)}{\partial p_B} = 0$$

$$\frac{\partial U^D}{\partial p_E} = D_E(p_B,p_E,..) + (p_E - c_E)\frac{\partial D_E(p_B,p_E,..)}{\partial p_E} + (p_B - c_B)\frac{\partial D_B(p_B,p_E,..)}{\partial p_E} = 0$$

Rearranging the terms we have that

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

with $i,j = L,H$ and $i \neq j$ and which represents the fact that the doctor as a monopolist chooses prices so that marginal revenue equals marginal costs $mr = mc$.

dividing both side 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 (the derivative of demand with respect to its own price goes to the right hand side) 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 putting the demand on the first term in the denominator we have

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

which is equivalent to

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

as the first term on the right hand side corresponds to the inverse of the own elasticity of demand, that is,

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

Diving and multiplying the denominator of the second term of the right hand side of the equation 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: Regression Estimates

Note: Payment variables are all assumed to be unofficial payments made in either the admission department (pay1) or on the ward (pay2). 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.

Tables 5 – 13 present OLS regression results. Tables 14 – 18 present duration model results

Table 5: 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	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Trauma	-0.0662			
Hac	-0.0796			
Rg1	0.0852	-0.0631	0.0936	0.3301*
Rg3	0.0685	-0.1024	-0.0090	0.1941*
Rg4	0.1091	-0.0297	0.1375	0.2020
Age	0.0123	-0.0026	0.0142	0.0209
Age_sq	-0.0002**	0.0000	-0.0002	-0.0002***
Male	-0.0028	-0.1074	-0.0246	0.0611
Lnincome	-0.0297	0.2091**	0.0352	-0.1967*
Pay1	0.0443	-0.2091**	0.0090	0.1681**
Pay2	0.1390*	-0.0748	0.1019	0.3001*
Student	-0.0261	-0.3555***	-0.0078	0.1093
Unemploy	0.0678	0.1975	0.0909	0.0181
Privwork	-0.0686	0.2062	-0.2607***	-0.0324
Selfwork	0.0965	-0.1356	0.2220	0.0183
Retired	0.2239	0.7395*	0.4035	-0.3468
Houswife	0.0557	0.0995	0.0633	0.0018
Exempt	-0.1073	-0.2290	-0.3281	0.3437
_cons	3.8377*	2.0149**	3.1325*	4.9986*
No of observations	1358	253	567	538
R2	0.0196	0.1046	0.0357	0.0825
Ramsey Reset test	F(3,1336)=2.45 Prob>F=0.0621	F(3,233)=0.56 Prob>F=0.6399	F(3,547)=0.49 Prob>F=0.6864	F(3,518)=1.11 Prob>F=0.3438
Mean VIF	6.76	8.33	7.43	7.41

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

Table 6: Admission time regressed on whether or not an unofficial payment was made (without age squared)

	Pooled Binary Pay	**Hospital 1 Binary Pay**	**Hospital 2 Binary Pay**	**Hospital 3 Binary Pay**
Admission time	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Trauma	-0.0659			
Hac	-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*
Pay1	0.0445	-0.2068**	0.0042	0.1766**
Pay2	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	F(3,234)=0.58	F(3,548)=0.49	F(3,548)=3.27
	Prob>F=0.0247	Prob>F=0.6275	Prob>F=0.6869	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.

	Pooled Continuous Pay	**Hospital 1 Continuous Pay**	**Hospital 2 Continuous Pay**	**Hospital 3 Continuous Pay**	**Pooled Continuous Pay Interactions**
Admission Time	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Trauma	-0.0493				-0.3323*
Hac	-0.0746				-0.2433**
Rg1	0.0686	-0.1212	0.0864	0.1332	0.0650
Rg3	0.0735	-0.1263	-0.0160	0.2116*	0.0624
Rg4	0.1015	-0.0368	0.1129	0.2103	0.1120
Age	0.0140***	0.0023	0.0155	0.0230	0.0130
Age_sq	-0.0002**	-0.0001	-0.0002	-0.0003***	-0.0002**
Male	-0.0039	-0.1368	-0.0195	0.0573	-0.0046
Lnincome	-0.0201	0.2637*	0.0341	-0.1905*	-0.0146
Lnpay1	0.0115***	-0.0445*	0.0067	0.0368*	-0.0278***
Hac*lnpay1					0.0342**
Trauma*lnpay1					0.0636*
Lnpay2	0.0146**	-0.0224***	0.0143	0.0346*	-0.0197***
Hac*lnpay2					0.0366*
Trauma*lnpay2					0.0476*
Student	-0.0075	-0.3181	-0.0090	0.2052	-0.0137
Unemploy	0.0581	0.2369	0.0731	0.0107	0.0670
Privwork	-0.0978	0.2184	-0.2853***	-0.0629	-0.0887
Selfwork	0.0735	-0.0436	0.1865	-0.0087	0.0916
Retired	0.2206	0.8897*	0.3966	-0.3838	0.2381
Houswife	0.0461	0.1023	0.0437	-0.0086	0.0304
Exempt	-0.1232	-0.2924***	-0.3475***	0.3466	-0.1292
_cons	3.6742*	1.5080	3.0829*	4.8050*	3.8284*
No of observations	1308	245	553	510	1308
R2	0.0187	0.1345	0.0356	0.0911	0.0311
Ramsey Reset test	F(3,1286)=2.02	F(3,225)=0.27	F(3,533)=1.16	F(3,490)=1.6	F(3,1282)=2.84
	Prob>F=0.1087	Prob>F=0.8484	Prob>F=0.326	Prob>F=0.1895	Prob>F=0.0366
Mean VIF	6.76	8.44	7.4	7.44	7.37

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

Table 8: Admission time regressed on amount paid as unofficial payment (without age squared).

	Pooled Continuous Pay	Hospital 1 Continuous Pay	Hospital 2 Continuous Pay	Hospital 3 Continuous Pay	Pooled Continuous Pay Interactions
Admission Time	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Trauma	-0.0495				-0.3359*
Hac	-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
Lnpay1	0.0115***	-0.0436*	0.0060	0.0382*	-0.0270***
Hac*lnpay1					0.0322***
Trauma*lnpay1					0.0641*
Lnpay2	0.0153**	-0.0218***	0.0142	0.0360*	-0.0189***
Hac*lnpay2					0.0354**
Trauma*lnpay2					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	F(3,226)=0.27	F(3,534)=0.75	F(3,491)=2.77	F(3,1283)=2.85
	Prob>F=0.0097	Prob>F=0.8443	Prob>F=0.5238	Prob>F=0.0412	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 9: Length of stay 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	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Trauma	0.6812*			
Hac	0.1789*			
Rg1	-0.0375	0.0288	-0.0257	-1.0088*
Rg3	0.1875*	0.1786*	0.1973*	0.1490**
Rg4	0.2582*	0.2744*	0.4070*	0.0498
Age	0.0216*	0.0221**	0.0288*	0.0095
Age_sq	-0.0002*	-0.0002**	-0.0003*	-0.0001
Male	0.0693*	0.1459*	0.0349	0.0221
Lnincome	-0.0587**	0.0179	-0.0686*	-0.0714
Pay_1	0.1325*	-0.0563	0.0102	0.3237*
Pay_2	0.2643*	0.3424*	0.1473*	0.3360*
Student	0.0379	0.1213	0.0874	-0.1346
Unemploy	0.0332	0.1041	0.0154	0.0237
Privwork	-0.0455	-0.0118	-0.0091	-0.0966
Selfwork	-0.0030	0.2373***	0.0196	-0.1446
Retired	0.1562***	0.3220**	0.1649	0.0249
Houswife	0.1005**	0.0928	-0.0011	0.0848
Exempt	0.0153	-0.0059	-0.1082	0.1830
_cons	2.0653*	1.2399*	2.2663*	3.1792*
No of observations	1482	290	600	592
R2	0.3545	0.2559	0.254	0.1092
Ramsey Reset test	F(3,1460)=2.39	F(3,270)=2.76	F(3,580)=0.87	F(3,572)=7.16
	Prob>F=0.0675	Prob>F=0.0425	Prob>F=0.4554	Prob>F=0.0001
Mean VIF	6.75	8.4	7.42	7.25

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

Table 10: LOS regressed on whether or not an unofficial payment was made (without age squared).

	Pooled Binary Pay	**Hospital 1 Binary Pay**	**Hospital 2 Binary Pay**	**Hospital 3 Binary Pay**
LOS	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Trauma	0.6828*			
Hac	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	F(3,271)=1.66	F(3,5811)=4.65	F(3,573)=6.47
	Prob>F=0.0861	Prob>F=0.1763	Prob>F=0.0032	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 11: Length of stay 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.**
Trauma	0.6635*				0.5591*
Hac	0.1686*				0.3051*
Rg1	-0.0338	0.0429	-0.0277	-0.9660*	-0.0334
Rg3	0.1791*	0.1802*	0.1924*	0.1613**	0.1739*
Rg4	0.2462*	0.2741*	0.3848*	0.0638	0.2613*
Age	0.0232*	0.0201**	0.0288*	0.0141	0.0206*
Age_sq	-0.0002*	-0.0002**	-0.0003*	-0.0001	-0.0002*
Male	0.0658**	0.1480*	0.0420	0.0027	0.0551**
Lnincome	-0.0695*	0.0039	-0.0739*	-0.0822	-0.0616*
Lnpay1	0.0198*	0.0010	0.0012	0.0496*	0.0091
Hac*lnpay1					-0.0070
Trauma*lnpay1					0.0379*
Lnpay2	0.0411*	0.0484*	0.0200*	0.0593*	0.0542*
Hac*lnpay2					-0.0325*
Trauma*lnpay2					0.0006
Student	0.0359	0.0925	0.0733	-0.1311	0.0088
Unemploy	0.0105	0.1003	-0.0088	-0.0105	0.0107
Privwork	-0.0542	-0.0210	-0.0385	-0.0926	-0.0570
Selfwork	-0.0200	0.2170***	0.0078	-0.1629	-0.0083
Retired	0.1535***	0.2711*	0.1535	0.0378	0.1411
Houswife	0.0907***	0.0765	-0.0069	0.0778	0.0665
Exempt	0.0057	0.0220	-0.1006	0.1607	0.0232
_cons	2.0634*	1.3434*	2.3114*	2.9991*	2.0424*
No of observations	1424	282	586	556	1424
R2	0.3584	0.2677	0.2536	0.1341	0.3758
Ramsey Reset test	F(3,1402)=0.2614	F(3,262)=2.35	F(3,566)=1.27	F(3,536)=2.19	F(3,1398)=2.71
	Prob>F=0.2614	Prob>F=0.0725	Prob>F=0.2836	Prob>F=0.0878	Prob>F=0.044
Mean VIF	6.75	8.38	7.39	7.37	7.33

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

Table 12: LOS regressed on amount of payment (continuous) without age squared.

	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.**
Trauma	0.6653*				0.5586*
Hac	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*
Lnpayc21	0.0199*	0.0029	0.0001	0.0503*	0.0103***
Hac*lnpay1					-0.0095
Trauma*lnpay1					0.0383*
Lnpay2	0.0418*	0.0496*	0.0198*	0.0599*	0.0554*
Hac*lnpay2					-0.0342*
Trauma*lnpay2					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	F(3,263)=1.03	F(3,567)=5.05	F(3,537)=1.46	F(3,1399)=1.92
	Prob>F=0.5431	Prob>F=0.3784	Prob>F=0.0019	Prob>F=0.2240	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 13: LOS regressed on (continuous) payment taking into account the potential endogeneity of the payment in the ward.

	Pooled Continuous pay	**Hospital 3 Continuous Pay**
LOS	**Coef**	**Coef.**
Lnpay2	0.2020*	0.2758*
Trauma	0.6321*	
Hac	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*
Lnpay1	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 14: Discrete time Cloglog models for the hospitals using the binary payment variable and with and without unobserved heterogeneity.

Binary payment	Discrete time PH Cloglog Hospital 1	Discrete time PH Cloglog Heterogeneity Hospital 1	Discrete time PH Cloglog Hospital 2	Discrete time PH Cloglog Heterogeneity Hospital 2	Discrete time PH Cloglog Hospital 3	Discrete time PH Cloglog Heterogeneity Hospital 3	Discrete time PH Cloglog Pooled	Discrete time PH Cloglog Heterogeneity Pooled
Hazard	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Logd	0.8953*	1.8971*	0.9631*	2.4434*	0.3438*	0.4545*	0.5654*	1.1098*
Trauma							-1.6746*	-2.0739*
Hac							-0.3943*	-0.5754*
Rg1	-0.0049	-0.2618	0.3508*	-0.0017	2.5084**	2.6047**	0.2914*	0.1241
Rg3	-0.6382*	-0.8603*	-0.2564*	-1.0266*	-0.1265	-0.1875	-0.2658*	-0.5532*
Rg4	-0.7658*	-1.3886*	-0.7407*	-2.0515*	-0.1678	-0.1426	-0.4534*	-0.8120*
Age	-0.0525**	-0.0989**	-0.0924*	-0.1274*	-0.0038	-0.0037	-0.0409*	-0.0555*
Age_sq	0.0005**	0.0010**	0.0008*	0.0012*	0.0000	0.0000	0.0004*	0.0005*
Male	-0.4684*	-0.5760**	-0.1913**	-0.1682	-0.0442	-0.0448	-0.1969*	-0.2263*
Lnincome	-0.0585	-0.0920	0.2944*	0.3463*	0.0977	0.1229	0.1093**	0.1711*
Pay_1	0.3483**	0.2054	-0.0477	-0.0766	-0.4936*	-0.5458*	-0.2284*	-0.3049*
Pay_2	-1.0165*	-1.4013*	-0.4504*	-0.6853*	-0.4510*	-0.5110*	-0.5221*	-0.6753*
Student	-0.3934	-0.6787	-0.4307**	-0.2686	0.2839	0.2677	-0.0939	-0.1598
Unemploy	-0.2890	-0.4794	-0.1529	0.0166	-0.2204	-0.2085	-0.1786***	-0.1556
Privwork	0.0935	-0.0657	-0.0697	0.1168	0.0940	0.1052	0.1066	0.0957
Selfwork	-0.5960***	-1.2361**	0.0482	-0.1469	0.4312***	0.4181	0.0976	-0.0190
Retired	-0.6658**	-1.3865**	-0.6725**	-0.8031***	0.0213	0.0394	-0.2302	-0.3472
Houswife	-0.2987	-0.3632	-0.1060	0.0389	-0.1060	-0.1242	-0.2064***	-0.2782***
Exempt	-0.1595	-0.1750	0.3840	0.6617***	-0.3073	-0.3073	-0.0960	0.0178
_cons	-0.4305	0.4870	-4.0267*	-5.5841*	-4.2704*	-4.6113*	-2.2655*	-2.7665*
Gamma var.		0.7593		0.8996*		0.1160*		-0.8212*
exp(ln_varg)								
Z		3.2170		6.4488		1.4484		6.5593
No of observations	2057	2057	5427	5427	12605	12605	20089	
Wald test for 17	Chi2(17)=151.29	chi2(17)=151.29	chi2(17)=357.69	chi2(17)=357.69	chi2(17)=95.75	chi2(17)=95.75	chi2(19)=776.72	chi2(19)=776.72
Variables	Prob>chi2=0	Prob>chi2=0	Prob>chi2=0	Prob>chi2=0	Prob>chi2=0	Prob>chi2=0	Prob>chi2=0	Prob>chi2=0
LR Test for existence of		chi2(1)=18.67		chi2(1)=100.75		chi2(1)=2.38		chi2(1)=74.37
Unobs. Heterogeneity		Prob>chi2=0.00		Prob>chi2=0.00		Prob>chi2=0.12		Prob>chi2=0.00
Log-likelihood	-762.8259	-753.4891	-1710.1028	-1659.7259	-2370.5549	-2369.3643	-4931.1352	-4893.9503

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

Table 15: Discrete time Cloglog models for the hospitals using the continuous payment variable and with and without unobserved

Continuous payment	**Discrete time PH Cloglog Hospital 1**	**Discrete time PH Cloglog Heterogeneity Hospital 1**	**Discrete time PH Cloglog Hospital 2**	**Discrete time PH Cloglog Heterogeneity Hospital 2**	**Discrete time PH Cloglog Hospital 3**	**Discrete time PH Cloglog Heterogeneity Hospital 3**
Hazard	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Logd	0.9092*	2.0466*	0.9755*	2.4609*	0.3391*	0.4737*
Rg1	-0.0401	-0.3462	0.3560*	0.0119	2.4195**	2.5358**
Rg3	-0.6639*	-0.9076*	-0.2428**	-1.0018*	-0.1723	-0.2438***
Rg4	-0.7492*	-1.4692*	-0.6965*	-1.9567*	-0.2501	-0.1979
Age	-0.0456*	-0.0893**	-0.0919*	-0.1289*	-0.0158	-0.0157
Age_sq	0.0004***	0.0009**	0.0008*	0.0012*	0.0001	0.0001
Male	-0.5072*	-0.6341**	-0.2249**	-0.1993	-0.0336	-0.0261
Lnincome	-0.0025	-0.0017	0.3183*	0.3722*	0.0908	0.1256
Lnpay1	0.0231	-0.0206	-0.0021	-0.0112	-0.0751*	-0.0856*
Lnpay2	-0.1399*	-0.2207*	-0.0537*	-0.0955*	-0.0718*	-0.0850*
Student	-0.3372	-0.4908	-0.3914**	-0.2242	0.2490	0.2403
Unemploy	-0.2873	-0.4737	-0.0613	0.1077	-0.1527	-0.1326
Privwork	0.0882	-0.0038	0.0090	0.2390	0.1011	0.1145
Selfwork	-0.5465	-1.1310***	0.0651	-0.1048	0.4544***	0.4454
Retired	-0.5902***	-1.2633**	-0.6279**	-0.7688***	0.0655	0.0672
Houswife	-0.2999	-0.2839	-0.0923	0.0629	-0.1184	-0.1251
Exempt	-0.1861	-0.3126	0.3157	0.6447***	-0.3516	-0.3261
_cons	-0.9430	-0.3309	-4.3041*	-5.7798*	-3.6291*	-4.0579*
Gamma var. exp(ln_varg)		0.8529		0.8859		0.1396*
Z		3.3714		6.3157		1.6850
No of observations	2005	2005	5235	5235	11676	11676
Wald test for 17	chi2(17)=151.29	chi2(17)=151.29	chi2(17)=325.46	chi2(17)=325.46	Chi2(17)=9.25	chi2(17)=9.25
Variables	Prob>chi2=0	Prob>chi2=0	Prob>chi2=0	Prob>chi2=0	Prob>chi2=0	Prob>chi2=0
LR Test for existence of		chi2(1)=21.37		chi2(1)=101.18		chi2(1)=3.3
Unobserved heterogeneity		Prob>chi2=0.00		Prob>chi2=0.00		Prob>chi2=0.07
Log-likelihood	-738.6618	-727.9788	-1660.9703	-1610.3781	-2214.0601	-2212.4112

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

Table 16: Comparison between discrete and continuous time models using continuous payment variable (different specifications of the Hazard function).

Continuous Payment	**Continuous time Exponential**	**Continuous time Gompertz**	**Continuous time Weibull**	**Continuous time Cox**	**Continuous time Gener. Gamma**	**Discrete time PH Cloglog Model**
Hazard	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Logd						0.5771*
Trauma	-1.1208*	-1.6809*	-2.0232*	-1.8329*	1.1145*	-1.6514*
Hac	-0.1812*	-0.1956*	-0.2882*	-0.3012*	0.1900*	-0.3811*
Rg1	0.4045*	0.5519*	0.9040*	0.8979*	-0.3606*	0.2758**
Rg3	-0.0189***	0.0606*	0.0339***	-0.0022	0.039*	-0.2704*
Rg4	-0.1653*	-0.2283*	-0.2525*	-0.2950*	0.1777*	-0.4586*
Age	-0.0238*	-0.0396*	-0.0486*	-0.0475*	0.0226*	-0.0468*
Age_sq	0.0002*	0.0004*	0.0005*	0.0004*	-0.0002*	0.0004*
Male	-0.1104*	-0.1795*	-0.2250*	-0.2163*	0.1037*	-0.1971*
Lnincome	0.0468*	0.0693*	0.0840*	0.0772*	-0.0496*	0.1271*
Lnpay1	-0.0282*	-0.0459*	-0.0499*	-0.0416*	0.0286*	-0.0336*
Lnpay2	-0.0424*	-0.0590*	-0.0756*	-0.0646*	0.0429*	-0.0757*
Student	-0.0059	-0.0122	-0.0982**	-0.1152*	-0.0091	-0.1248
Unemploy	-0.1003*	-0.1783*	-0.1709*	-0.1784*	0.0996*	-0.1115
Privwork	0.0767*	0.1147*	0.1397*	0.1480*	-0.0750*	0.1274
Selfwork	0.2665*	0.4811*	0.5681*	0.4612*	-0.2434*	0.1238
Retired	-0.0318	-0.1511*	-0.1223*	-0.1592*	0.0202	-0.1963
Houswife	-0.0884*	-0.1217*	-0.1564*	-0.1473*	0.0876*	-0.1961***
Exempt	-0.1958*	-0.3321*	-0.3792*	-0.2767*	0.1847*	-0.1111
_cons	-1.8565*	-1.8658*	-3.5254*		1.8330*	-2.1835*
Gamma		0.0279*				
P			1.8058*			
ln_sigma					-0.5209*	
Kappa					0.3811*	
No of observations	18906	18906	18906	18906	18906	18916
Wald test for 18 (19) variables	chi2(18)=20983 Prob>chi2=0	chi2(18)=16512.8 Prob>chi2=0	chi2(18)=13486.2 Prob>chi2=0	chi2(18)=12029.3 Prob>chi2=0	chi2(18)=19029.7 Prob>chi2=0	chi2(19)=759.1 Prob>chi2=0
Wald test omitted variables	chi2(1)=0.28 Prob>chi2=0.591	chi2(1)=96.72 Prob>chi2=0.000	chi2(1)=0.18 Prob>chi2=0.674	chi2(1)=22.47 Prob>chi2=0.000	chi2(1)=0.48 Prob>chi2=0.489	
Test proport. Haz. assumption (global test)				chi2(18)=715.04 Prob>chi2=0.000		
Wald test for kappa=1					Chi2=9.467 Prob>chi2=	
Log-likelihood	-22233.195	-19930.734	-17930.166	-163754.66	-17435.913	-4697.7418

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

Table 17: Comparison between discrete and continuous time models using continuous payment variable (different specifications of the Hazard function) accounting for heterogeneity (frailty).

	Continuous time	**Continuous time**	**Continuous time**	**Continuous time**	**Discrete time**
	Exponential Frailty	**Gompertz Frailty**	**Weibull Frailty**	**Gener. Gamma Frailty**	**PH Cloglog Model Frailty**
Hazard	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Logd					1.1563*
Trauma	-1.1208*	-1.6808*	-2.7625*	1.0840	-2.0562*
Hac	-0.1812*	-0.1956*	-0.4549*	0.1718	-0.5562*
Rg1	0.4046*	0.5519*	0.8007*	-0.3941	0.1074
Rg3	-0.0188***	0.0606*	-0.1147*	-0.0023	-0.5482*
Rg4	-0.1653*	-0.2283*	-0.4716*	0.1289	-0.7952*
Age	-0.0238*	-0.0396*	-0.0480*	0.0226	-0.0598*
Age_sq	0.0002*	0.0004*	0.0004*	-0.0002	0.0005*
Male	-0.1105*	-0.1795*	-0.2484*	0.1016	-0.2271*
Lnincome	0.0468*	0.0693*	0.1335*	-0.0496	0.2004*
Lnpay1	-0.0282*	-0.0459*	-0.0712*	0.0286	-0.0476*
Lnpay2	-0.0424*	-0.0590*	-0.1028*	0.0426	-0.1040*
Student	-0.0060	-0.0122	0.1121**	-0.0075	-0.1373
Unemploy	-0.1003*	-0.1783*	-0.2379*	-0.0180	-0.0964
Privwork	0.0767*	0.1147*	0.2107*	-0.1599	0.1251
Selfwork	0.2665*	0.4811*	0.5924*	-0.3159	0.0288
Retired	-0.0318	-0.1511*	0.0711	-0.0136	-0.3689
Houswife	-0.0884*	-0.1217*	-0.2034*	0.0041	-0.2530***
Exempt	-0.1958*	-0.3321*	-0.4193*	0.1514	0.0606
_cons	-1.8564*	-1.8658*	-4.6964*	1.8331	-2.8187*
Gamma		0.0279*			
P			2.4516*		
ln_sigma				-0.9971	
Kappa				0.7643	
Gamma variance indiv. heterogeneity	-16.3422*	-14.9501*	-0.5685*	0.4517*	0.4645*
Z					6.6161
No of observations	18906	18906	18906	18906	18916
Wald test for 18 (19) variables	chi2(18)=20983.7 Prob>chi2=0	chi2(18)=16512.5 Prob>chi2=0	chi2(18)=6496.3 Prob>chi2=0		chi2(19)=759.1 Prob>chi2=0
Wald test for omitted variables	chi2(1)=0.29 Prob>chi2=0.59	chi2(1)=96.73 Prob>chi2=0.00	Chi2(1)=5.99 Prob>chi2=0.00		
LR test existence unobs. heterogeneity	Prob>chi2=0.00	Prob>chi2=0.00	Prob>chi2=0.00	Prob>chi2=0.00	chi2(1)=75.86 Prob>chi2=0.00
Log-likelihood	-22233.2	-19930.73	-17543.76	-17569.61	-4659.81

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

Table 18: Comparison between discrete and continuous time models using continuous payment variable (different specifications of the Hazard function). Interaction terms.

Continuous Payment	**Continuous time Exponential**	**Continuous time Gompertz**	**Continuous time Weibull**	**Continuous time Cox**	**Continuous time Gener. Gamma**	**Discrete time PH Cloglog**
Hazard	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef**	**Coef.**
Logd						0.6106*
Trauma	-1.1052*	-1.6665*	-2.1255*	-1.9662*	1.0867*	-1.5919*
Hac	-0.4165*	-0.4861*	-0.7420*	-0.7886*	0.4215*	-0.7048*
Rg1	0.3564*	0.4795*	0.8063*	0.8084**	-0.3273*	0.2462**
Rg3	-0.0109	0.0799*	0.0574*	0.0153*	0.0281**	-0.2586*
Rg4	-0.1989*	-0.2792*	-0.3156*	-0.3646*	0.2089*	-0.5117*
Age	-0.0216*	-0.0367*	-0.04738*	-0.0472*	0.0204*	-0.0451*
Age_sq	0.0002*	0.0004*	0.00058*	0.0004*	-0.0002*	0.0004*
Male	-0.0967*	-0.1623*	-0.1908*	-0.1807*	0.0930*	-0.1871*
Lnincome	0.0472*	0.0782*	0.1038*	0.0946*	-0.0470*	0.1190**
Lnpay1	0.0187*	0.0277*	0.0459*	0.0501*	-0.0160*	0.0037
Hac*lnpay1	-0.0125*	-0.0256*	-0.03858*	-0.0357*	0.0094**	-0.0017
Trauma*lnpay1	-0.0672*	-0.1030*	-0.13078*	-0.1247*	0.0654*	-0.0901*
Lnpay2	-0.0864*	-0.1176*	-0.17198*	-0.1654*	0.0843*	-0.1337*
Hac*lnpay2	0.0647*	0.0823*	0.12968*	0.1358*	-0.0627*	0.0937*
Trauma*lnpay2	0.0457*	0.0642*	0.10488*	0.1083	-0.0418*	0.0549**
Student	0.0391***	0.0621**	0.00528	-0.0324*	-0.0485**	-0.0771
Unemploy	-0.1026*	-0.1841*	-0.1843*	-0.1847*	0.1016*	-0.1130
Privwork	0.0577*	0.0777*	0.0756*	0.0917*	-0.0617*	0.1097
Selfwork	0.2382*	0.4350*	0.5028*	0.3989*	-0.2235*	0.0874
Retired	-0.0280	-0.1649*	-0.1381*	-0.1588*	0.0149	-0.1947
Houswife	-0.0633*	-0.0988*	-0.1268*	-0.1166*	0.0607*	-0.1545
Exempt	-0.2277*	-0.3851*	-0.4516*	-0.3545*	0.2186*	-0.1738
_cons	-1.8643*	-1.9828*	-3.7131*		1.8354*	-2.1167*
Gamma		0.0287*				
P			1.8470*			
Ln_sigma					-0.5428*	
Kappa					0.4114*	
No of observations	18906	18906	18906	18906	18906	18916
Wald test for 22 (23) variables	chi2(22)=20983 Prob>chi2=0	chi2(22)=17281. Prob>chi2=0	chi2(22)=14201.5 Prob>chi2=0	chi2(22)=13106.3 Prob>chi2=0	chi2=21172.15 Prob>chi2=0	chi2(23)=803.31 Prob>chi2=0
Wald test omitted variables	chi2(1)=4.57 Prob>chi2=0.03	chi2(1)=53.65 Prob>chi2=0.00	chi2(1)=1.35 Prob>chi2=0.25	chi2(1)=5.63 Prob>chi2=0.02	chi2(1)=6.99 Prob>chi2=0.01	
Test proport. Haz. (global test)				chi2(22)=989.76 Prob>chi2=0.00		
Wald test for kappa=1					Chi2=-5.8785 Prob>chi2=	
Log-likelihood	-22115.059	-19701.606	-17541.944	-163372.86	-17096.418	-4675.624

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

Table 19: Comparison between discrete and continuous time models using continuous payment variable (different specifications of the Hazard function) accounting for heterogeneity (frailty). Interaction terms.

	Continuous time Exponential	**Continuous time Gompertz**	**Continuous time Weibull**	**Continuous time Gener. Gamma**	**Discrete time PH Cloglog heterogeneity**
Hazard	**Coef.**	**Coef.**	**Coef.**	**Coef.**	**Coef.**
Logd					1.1444*
Trauma	-1.1052*	-1.6665*	-2.6193*	1.0789	-1.8725*
Hac	-0.4165*	-0.4861*	-1.0330*	0.4256	-0.9520*
Rg1	0.3563*	0.4795*	0.7228*	-0.4396	0.1029
Rg3	-0.0109	0.0798*	-0.1097*	-0.0151	-0.5305*
Rg4	-0.1989*	-0.2792*	-0.5720*	0.1923	-0.8338*
Age	-0.0216*	-0.0367*	-0.0419*	0.0204	-0.0517*
Age_sq	0.0002*	0.0004*	0.0003*	-0.0002	0.0004*
Male	-0.0967*	-0.1623*	-0.2351*	0.0247	-0.2120*
Lnincome	0.0473*	0.0782*	0.1265*	-0.0470	0.1816*
Lnpay1	0.0187*	0.0277*	0.0357*	-0.0160	-0.0202
Hac*lnpay1	-0.0125*	-0.0256*	-0.0137	0.0096	0.0175
Trauma*lnpay1	-0.0672*	-0.1030*	-0.1692*	0.0645	-0.1050*
Lnpay2	-0.0864*	-0.1176*	-0.1966*	0.0843	-0.1639*
hac*lnpay2	0.0647*	0.0823*	0.1497*	-0.0622	0.1046*
Trauma*lnpay2	0.0457*	0.0642*	0.0832*	-0.0485	0.0424
Student	0.0391**	0.0622**	0.2033*	0.0379	-0.0473
Unemploy	-0.1026*	-0.1841*	-0.2287*	0.0262	-0.0882
Privwork	0.0577*	0.0777*	0.2062*	-0.1127	0.1431
Selfwork	0.2382*	0.4350*	0.5678*	-0.2931	0.0271
Retired	-0.0280	-0.1648*	0.1093***	0.0149	-0.3099
Houswife	-0.0633*	-0.0988*	-0.1082**	-0.0665	-0.1701
Exempt	-0.2277*	-0.3851*	-0.5047*	0.2186	-0.0177
_cons	-1.8644*	-1.9828*	-4.8164*	1.8354	-2.7546*
Gamma		0.0287*			
P			2.4933*		
Ln_sigma				-1.0847	
Kappa				0.8231*	
Gamma var. (ln_varg)	-16.7245*	-15.4754*	-0.5947*	0.7326*	0.4214* e(ln)
Z					6.2544
No of observations	18906	18906	18906	18916	18916
Wald test for 22 (23) variables	chi2(22)=23587.78 Prob>chi2=0	chi2(22)=17282.3 Prob>chi2=0	chi2(22)=7227.01 Prob>chi2=0		chi2(23)=803.31 Prob>chi2=0
Wald test omitted variables	chi2(1)=4.58 Prob>chi2=0.03	chi2(1)=53.65 Prob>chi2=0.00	chi2(1)=2.86 Prob>chi2=0.09		
Test proport. Haz. (global test)					
Test for existence of unobs. heterogeneity					chi2(1)=68.52 Prob>chi2=0.0
Log-likelihood	-22115.059	-19701.606	-17157.066	-17462.632	-4641.366908

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

Table 20A: Comparisons between models using the log-likelihood and the Akaike information criterion.

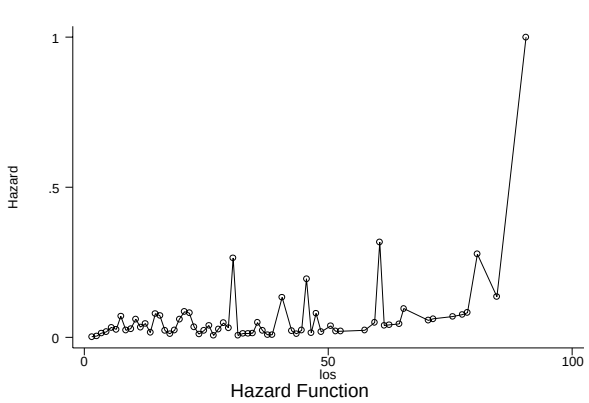
Continuous payment	Exponential	Gompertz	Weibull	Cox	Gen. Gamma	Cloglog
Log-likelihood	-22,233.19	-19,930.73	-17,930.17	-163,754.66	-17,435.913	-4,697.7418
Rank	4	3	2	5	1	1*
AIC	44,504.39	39,899,46	35,900.33	327,547.32	34,913.826	9,437.4836
Rank	4	3	2	5	1	1*

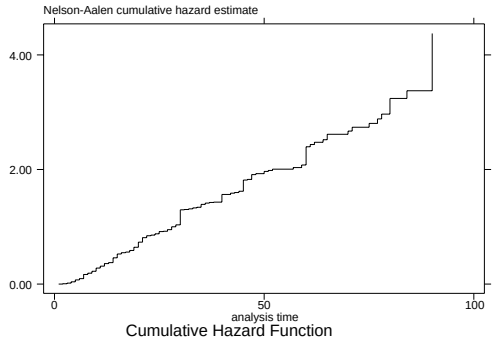
Note that the generalised gamma model would be chosen both by the log likelihood and the AIC criterion. This model also passes the Wald specification test (the Gompertz and the Cox do not). In this sense the generalised gamma model should be preferred in the continuous case. According to the criteria the discrete time Cloglog model performs better.

Table 20B: Comparisons between models using the log-likelihood and the Akaike information criterion accounting for heterogeneity.

Continuous payment	Exponential heterogeneity	Gompertz heterogeneity	Weibull heterogeneity	Gen. Gamma heterogeneity	Cloglog heterogeneity
Log-likelihood	-22,233.2	-19,930.73	-17,543.76	-17,569.913	-4,659.81
Rank	4	3	1	2	1**
AIC	44,504.39	39,899,46	35,127.52	35,181.82	9,361.626
Rank	4	3	1	2	1**

Figure 2: Nelson-Aalen hazard function estimates





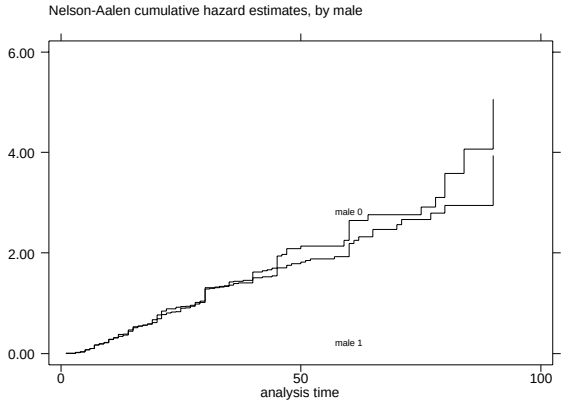
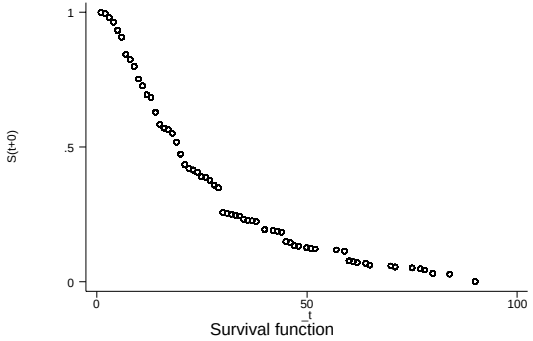
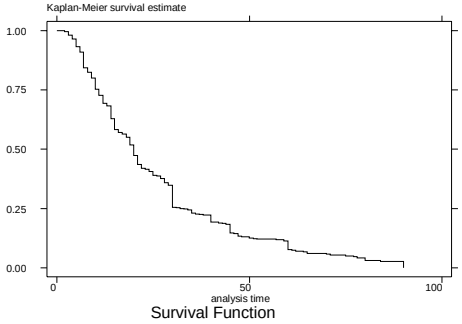
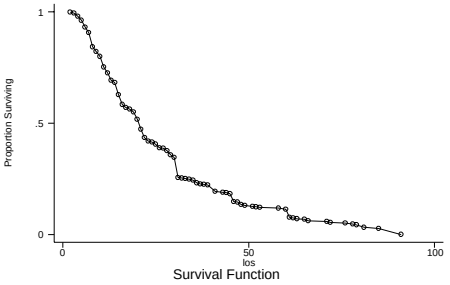


Figure 3: Kaplan Meier survival function estimates







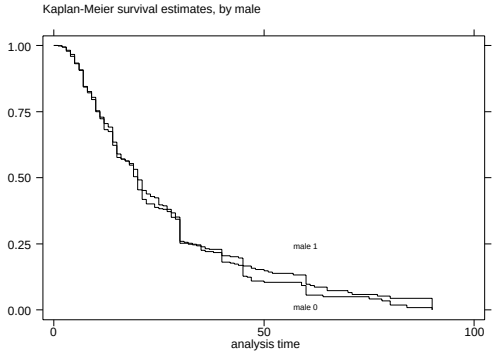


Figure 4: Scaled Schoenfeld Residuals