Political Science Department
Central European University

Payments for Services Free by Law in the Russian Public Healthcare: Analysis of a Hidden Institution

Ph.D. Dissertation
submitted to Political Science Department, the Central European University

by Maxim Ryabkov
supervisors: Balázs Varadi and Iván Csaba
defense board members: Paul Anand, Loránd Ambrus-Lakatos, Tamás Meszerics

Budapest
May 2004

Acknowledgements

This work would not have been possible without advice and support of quite a number of people and institutions. Iván Csaba (formerly of Political Science Department at the CEU) has helped me through the research stage of the project, teaching me consistency and precision. Balázs Varadi has greatly helped with completion of the project, patiently listening to my arguments and bettering them with many new ideas, only few of which I explicitly acknowledge in the text. Loránd Ambrus-Lakatos has taught me formal modeling and commented on several drafts of this text. Tamás Meszerics helped hone my story of private payments in public healthcare at several key junctions. Tamás Rudas has been my teacher of statistics. Paul Anand has kindly agreed to be an external referee for this dissertation. Him together with many others I thank for fruitful discussions. Alexander Maxwell edited the text and made a lot of useful suggestions.

During my field research in Russia, I greatly benefited from discussions with Sergey Shishkin of Independent Institute for Social Policy, Igor Sheiman, Elena Potapchik, Larisa Popovich, Tatyana Makarova, and many others, whose research jump started mine and who showed me the way through the intricacies of Russian healthcare.

My special thanks go to the Central European University for financial and logistical support and to the University of Boston and Institute for Social Research, Moscow for permission to use primary data from their surveys.

My heartfelt thanks go to my wife, Maria Khovanskaia, whose support and encouragement cannot be overestimated.

The remaining factual errors, misinterpretations, and conceptual blunders are all mine.

Abstract

This dissertation presents the empirical and theoretical findings of multi-pronged research into the nature and causes of payments for services which are free by law in Russian public healthcare. This research rejects the received wisdom of explaining the phenomenon by insufficient public funding. Instead, such payments are viewed as informational rent to medical professionals. I further hypothesize that private money plays the double role of the purchase of services and of establishing a special relationship with the medical professional. A mechanism of quality assurance exists which is based on the professional and reputation concerns of the doctor, who is willing to favor paying patients over non-paying ones. A regulatory gap creating a conflict of interest for the providers creates a favorable environment for such a mechanism. Limited public financing can hypothetically be viewed as a form of rent extraction.

Contents

Chapter One. Introduction

This introductory chapter pursues three objectives. The first is to present the aims and methods of the research undertaken. The second is to describe the statutory institutions of the Russian public healthcare, so as to set the scene for the main story of this research. The third objective is to place this research within the existing literature, both on the same and related subjects. Hence, section 1.1 is a statement of purpose with a description and justification of methodology. Section 1.2 describes the current system of public healthcare of Russia and places the research topic in the context of recent developments in health policies. Section 1.3 discusses the relevant literature and embeds this research into a number of broader theoretical topics.

1.1 Research Purpose and Methods

1.1.1 Statement of purpose. Consider the following two stylized human interest stories about the Russian public healthcare. Someone visits a doctor with a complaint and receives a suggestion of hospitalization. Hospitalization and treatment are free of charge by law. The truth of the matter, however, is that hospitalization is unlikely to happen very soon because of queues. It is desirable for the patient to be hospitalized as soon as possible, and, moreover, it is desirable to end up in a good hospital. The current rules limit the choice of hospital. Queue jumping is possible if the patient goes home, waits for the next crisis, and then calls the emergency services. A good hospital however is not guaranteed. Money can solve both problems, says the doctor. The patient pays and signs a document saying that he declines the opportunity to stand in the queue for free hospitalization.

The other story is about a nurse able to slacken her performance and go unpunished. An elderly patient lies with a dropper under his clavicle. Failing to attend to such a patient (to change the dropper) may result in clogging up of the blood vessel and death. Extra money paid or expected from relatives of the elderly patient may mean the difference between life and death.

Both stories can be enriched with more details from a wide range of choices. Across all the divergent paths one can choose from, one thing remains constant. A patient pays for something that has been officially promised to be free of charge possibly receiving additional quality or quantity of service.

The subject matter of this research is payments for services free by law in the Russian public healthcare. The next subsection formulates the exact research questions. Answering them will lead to an understanding of the modes of relevant transactions, of motivations behind them and of the resulting distribution of costs and benefits.

1.1.2 Research questions and methods. This dissertation answers the following three research questions. The first is the most general: Why does a system of payments for services and goods free by law emerge and persist without receiving much political or administrative attention, as if not only the state but the general public, patients themselves, preferred the current state of affairs? The literature (see Section 1.3 for references) suggests hat the answer lies in the financial situation of the public healthcare. It is very intuitive that low pay can weigh in doctors' decision-making, while a patient's being informed about the miserable salaries enhances patient willingness to pay. The true story, however, is more complex than that, as will be seen shortly.

My second research objective is to place the issue of the payments into the context of past and current efforts by public authorities to remodel and govern the system of public healthcare in Russia. How do payments for services free by law fit within the current institutional set up? How do general health policies react to the dubious practice? What interests might be vested in the existence of the payments?

The third question touches on the micro-level of interactions behind payments for services free by law. What distribution of benefits and costs do the transactions of interest bring about? One particularly interesting sub-question is: Do patients or their relatives aim *to buy a service* from a provider or professional, whenever this service is in excess of what should have been accorded them for free? In more theoretical terms, can provider-patient relations be modeled in terms of relations of buyer-seller interactions, possibly with further complications due to informational asymmetry? The simplest answer would be an affirmative, but this dissertation contends that a different type of relations appears to dominate the transactions of interest.

The concept of leverage of a theory has been defined in the literature: a theory has high leverage if it explains much with little (King et al. 1994; see also Dowding 2001). Obviously, this definition means that a large and complex theory explaining a single simple event is a bad one, since it has small leverage. In this case, the event is, as it were, 'over-explained', even if explained well. The other extreme is when a theory 'under-explains' a complex phenomenon, leaving many gaps in reconstruction of relevant causal links. Formalization helped by many strong assumptions could serve as an example. Too many, or too strong, assumptions mean small leverage.

This research is a case study in institutional analysis. Its starting point is the question why an institution persists despite its apparent perversity. In order to maximize the leverage of my explanation I will seek a minimal balance of factors that explains stability of the institution. For that purpose, I will find a place for this institution first of all, within a system of official structures and relations; secondly, in terms of micro-level interactions among individual agents.

Speaking about the appropriate methods, a few stylized features of the object of the research are to be taken into account. The institution is clandestine; the perceptions of the actors involved are fraught with idiosyncrasies and alleged self-deception. Also, the research covers many instances of a phenomenon, the instances being rather heterogeneous. Given the fact that the actual surveys were not intended to deal with the said intricacies, the latter make exclusive reliance on statistical analysis rather inappropriate. Instead, I propose to keep in balance three different approaches: quantitative statistical description, qualitative institutional analysis at micro- and macro-levels, and, finally, speculative modeling.

First of all, I place the institution of payments for services free by law within the framework of official institutions and policies. This is the macro-analysis, the first empirical part of the research. Then I consider the interaction of providers and patients in order to identify micro-level factors explaining the stability of the institution of interest. This is the second empirical part of the research. Finally, I attempt to construct a broader speculative picture to analyze payments for healthcare services free by law from the point of view of cost-benefit analysis. The last component is a formal economic model, demonstrating the consistency of some speculative concepts.

1.1.3 Structure of the Dissertation and limits to its ambition. Aiming to provide a consistent and convincing combination of empirical and theoretical arguments, I will employ the following structure. The dissertation takes the reader from hard facts to not-so-hard facts, to yet softer ones, and then to speculative inquiries, which build on the empirical picture.

Chapter Two describes and discusses the legal status of free and paid healthcare in Russia and the institutional set-up behind free care endowment. I highlight and conceptualize opportunities for the institutional provider and individual medical professional to charge the patient for services within the free care endowment. Chapter Three describes and analyses quantitative survey-based and qualitative interview-based data. A majority of the respondents in the qualitative interviews had accepted money for their services, and special care is taken to ensure the validity of conclusions in light of the potential unreliability of the source of data. Appendix A supports Chapter Three with a detailed account of sources of data.

Chapter Four presents a theory of payments for services free by law. The theory draws on the empirical evidence and suggests a balance of various effects the payments have on distribution of costs and benefits for patients, services providers and the public institutions financing and controlling healthcare. Chapter Five formalizes some of the insights from Chapter Four in form of a simple incentive-contract model. Appendix B extends Chapter Five with some more formalism.

One limit to the research was inevitably set by access to data. Data aggregation across space and time results in a largely static picture of payments for services free by law. The picture is also averaged across the many regions of a huge country. Comparisons between post-reform Russia and the Soviet Union are not feasible at the moment, though it remains likely that the nature and major effects of charges for services free by law have not changed since the Soviet times.

The subject matter of this research certainly belongs to a broader context of administrative practices in Russia and corruptive elements therein. Co-existence of paid and free services is found, beside healthcare, in education. Comparison and extrapolation across ‘close-to-state’ sectors is therefore a legitimate exercise, which does not, however, belong to the ambit of this research. It would rather be its natural extension.[1]

1.2 Russian Public Healthcare: The General Setting

In this second part of the Introduction, the scene is set. The recent history of Russian public healthcare is briefly described, with emphasis on the overall institutional setting.

1.2.1 The country. Russia (the Russian Federation) is normally considered as an economic and political unity in transition from an authoritarian rule with a state-controlled (socialist) economy to a more democratic and market-oriented society. The transition officially started in January 1992 with the lifting of price controls and various reforms, including that of the healthcare system. The border between socialism and capitalism is, however, rather blurred. Much of the retail service and many services were gradually leaving state control in the second half of the 1980s. Adding general erosion of central control over the economy, one could consider the transition as having started years before the official collapse of socialism. This notion of gradual transformation as opposed to an abrupt regime change is certainly true of healthcare, if anything else. For a strong opinion on this topic as well as a detailed description of economic policies see Hough 2001.[2]

[1] An example might however show the relevance of a wider picture in explaining its constituent parts. It is well known, if on anecdotal level, that police officers, secondary school and higher education teaching staff, and, obviously, medical professionals, tend to rely on arguably illegal income from charging their customers officially entitled to free services. Assuming this, it seems possible that corruption in one sector encourages corruption in others. A school teacher conditions good marks for a student on the student's participation in paid tutoring sessions, even if the latter are not warranted by considerations of the student's performance. That teacher, however, must pay a doctor for treatment in hospital to a doctor, who bribes a road policeman. This road policeman can, for the sake of argument, be the father of that student. Money flows in circles.

[2] Jerry Hough traces manifold unhealthy forms of governmental involvement in the economy in Russia. He believes that "it can be confidently asserted that the Russian economy was centrally regulated and directed from January 1992 through the end of 2000. " (Hough 2001, 31)

As of now, Russia is a country of 145 million people, a federal republic, with a very strong presidency and a relatively weak two-tier parliament, consisting of the State Duma (elected under a mixed proportional and majority system) and the *Sovet Federatsii* (Senate, consisting of regional representatives).[3] There is a consensus around the notion that the "federal center" is virtually equivalent to the presidential power, exercising control over a large part of bureaucracy. Regions, as members of the federation, are the main counterbalance to this "federal center". Regions are represented by elected regional senators in the Senate. Federal units include:

- 49 non-ethnic federal units or *oblasts*;
- 32 ethnic federal units (republics with particularly high level of autonomy and *autonomous okrugs*);
- 6 *krais*;
- 2 federal cities (Moscow and Saint Petersburg)

Further subdivision includes municipal units (*rayons*). A common name for a head of a federal unit is *gubernator* ("governor", normally for an *oblast* or *krai*, also Saint Petersburg), sometimes "president" (for an ethnic federal unit) or "mayor" (for example, in Moscow) or "head of administration" (for example, Khabarovsk *krai*). These are executive power; regional legislatures (sometimes called *Duma*) exist, as well. The executive branch is normally the most powerful: the federal balance of powers mirrors the regional forms of governance.

Russia has a market economy with major elements of a regular market-oriented legal system in place (see International Monetary Fund 2003 for data, Brown 2001

[3] Here is a widespread view on the results of the recent parliamentary elections on December 7, 2003: "Russia's experiment with parliamentary democracy, never full-hearted, is more or less dead." (*Economist*, Dec 11th 2003).

and Hough 2001 for analysis). Russia is not yet a member of the WTO (May 2004) though negotiations are under way.

Some indicators are presented in Table 1-1. The choice of indicators is such as to give the reader a broad notion of Russian public finances within the country's economy.

Table 1.1 General characteristics of the public finances of the Russian Federation

Variable	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**
GDP year-on-year change, %	-4.1	-3.4	0.9	-4.9	5.4	9.0	5.0	4.3
Federal government spending, % GDP		20.9	19.8	16.7	17.1	14.6	14.9	15.8
Regional and local government spending, % GDP		16.7	19.7	17.1	15.1	14.6	15.1	15.7
Transfers and loans from the federal into regional budgets, % GDP		3.0	3.4	1.9	1.7	1.7	2.7	2.7
Off-balance funds[4] (spending) % GDP		8.2	10.0	8.5	8.2	8.0	8.0	8.3[5]
Health-related expenditure by regional and local governments, % GDP		2.4	2.7	2.1	2.0	1.9	1.8	2.1
Expenditures by Mandatory Health Insurance Fund, % GDP		0.7	1.2	1.1	1.0	1.0	1.0	1.2[6]

Source: International Monetary Fund 2003

Relationships between the federal center and regions are a major political issue and a major source of political intrigue. The balance of powers is fluid and comprises both formal and informal channels of influence. The actual distribution of power through formal and informal mechanisms is very complex. Archie 2001 is a collection of papers that address the issues pertaining to the recent developments of the Russian polity and its federal structure in particular.

[4] These are: Pension, Mandatory Health Insurance, Employment, Social Insurance. Another name for them is extrabudgetary funds. The nature of Pension fund is however such that it may be a hidden source of budget financing, because it can lend money to the government.

[5] Average for the first three quarters.

This short overview of the country has not given its due to social problems, policies and public perceptions of them. Ryan (1994) provides a comprehensive collection of social problems-related data from the early period of transition. Rose (2001a and 2001b) gives a lot of material regarding Russian citizens' assessment of their situation. Health status issues are covered in some detail. Statistics suggest that standard health parameters, such as life expectancy and illness incidence have recently declined. However, as is often the case in healthcare research, such variables are difficult to link to changes in healthcare provision (see Rose 2001b, especially page 29 onwards; Carr (1982) discusses the political and ideological role of hypothetical links between national health parameters and national medicine). Chernichovsky et al. 1998 is an attempt to model links between health of the nation and economic and environmental situation. Their conclusion (put crudely) is that variation in financing cannot have much effect on the health status of people.

1.2.3. Public healthcare. The Russian public healthcare providers are hospitals, policlinics and their combinations, so-called territorial medical units. Hospitals render in-patient services, while policlinics are integrated out-patient units. Providers have not been privatized, nor even turned into autonomous economic entities. The legal status of hospitals and policlinics is *uchrezhdenie,* in other words a branch of the state. This means that medical providers have roughly the same status as schools or the police. Like schools and the police they enjoy some financial autonomy and can earn extra money through provision of paid services subject to strict limitations. In particular, providers may sell services in excess of what is required from them by

[6] Average for the first three quarters.

their participation in statutory healthcare, receive money for those services, and distribute proceeds according to certain rules (see Chapter Two on the subject).

Public healthcare has undergone some changes since the collapse of socialism. The reform of Russian public healthcare started at the end of the 1980s with introduction of something similar to the British GP fund-holding. In Saint Petersburg (then Leningrad) and Kemerovo, policlinics were allowed to keep money that they transferred to hospitals for patients for whom they were gatekeepers. The scheme collapsed, because of lack of accountability and the reluctance of policlinics to send patients to hospitals.

The Russian Constitution (see Constitution 1993, Art 41) establishes a general right to healthcare, while the 1993 Health Protection Law (Health Protection Law 1993, notably Art. 17, 19) develops this notion. The basis for the current system of Russian public health financing is the 1991 framework Health Insurance Law (Health Insurance Law 1991, Art. 6 in particular).[7] This Law establishes the institute of Mandatory Health Insurance (henceforth, MHI). The legal framework developed over three years and by October 1993 its major elements, down to bureaucratic procedures of money collection and disbursement, were finally adopted (see Federal Fund of Mandatory Health Insurance 1995).

The main players in the public healthcare field are the Federal Ministry of Health, the Mandatory Health Insurance Fund with its regional structures, private insurance companies contracted within the Mandatory Health Insurance, healthcare departments of the executive power in the regions, and medical professional organizations.

Russian public care is financed from:

[7] As regards early developments in post-socialism healthcare, see Chernichovsky et al. 1996. In particular, the authors indicate that in the last year of socialism public spending on healthcare increased and went down with market reforms and deepening economic crisis.

- mandatory insurance premia, paid as a payroll tax for the working (3.6% of the payroll, of which 0.1% goes to the Federal Fund and 3.5% to the regional funds) and local budget appropriations for the non-working; this is the MHI component;
- appropriations from both the regional (including municipal and other sub-regional) and federal budgets.

The MHI component generally covers payroll, medication and some other operating expenses of hospitals and policlinics. Commercial overhead, capital investment, sometimes emergency care, and some medical needs are financed from regional or Federal budgets. Municipal budgets are sometimes involved, too. Sometimes dentistry is financed from a regional budget. Treatment of tuberculosis and cancer and a number of special programs are financed from regional and the federal budgets. Mandatory Health Insurance financing amounts to slightly upwards of one-third of total public healthcare financing in the country. The figure increased steadily since the inception of the system to about 40% presently.

Public money comes into the system through two channels: regional (including municipal) and federal budgets and the MHI. The MHI financing is much more transparent than the budgetary one. Independent experts (see TACIS 1998) habitually express discontent over the two-pronged nature of healthcare financing, because budgetary appropriations also cover salaries and medication, even though officially budgetary and insurance financing are designated for different purposes.

As regards MHI financing, note that there are two tiers of financing institutions within the MHI. The upper tier consists of the Federal and regional Funds of MHI. These are statutory organizations that manage extra-budgetary public finances. Healthcare providers are either financed directly by regional Funds or by private insurance companies, who contract, finance, and audit service providers. These private insurers constitute the officially state-independent lower-tier in the MHI pyramid.

The introduction of private insurance companies was meant to create a competitive environment. Private companies and regional MHI funds are, however, obliged to contract all the providers in the territory, which eliminates some of the potential for competition. Competition appears at another key juncture: contracting with enterprises for servicing their employees. Private companies do engage in limited competition over these corporate clients (World Bank 1999; Shishkin 2000b).

Schemes of financing differ across regions. In 1998, fee for service was used in 30.3% of all cases, while diagnostic groups were used in 33.9%. Cost reimbursement was used in 17.4%; while capitation in 14.4% of cases. Diagnostic groups were used in 58.7% of all hospitals, and day-beds in 29.1% (Shishkin 2000b, 291-293). Gleaning incentive effects is impossible, as data do not show the expected correlation between the power of an incentive scheme and output of the provider (Shishkin 2000, 293; Makarova 2000).[8] Diversity among regions, settlements, and care providers is high enough for relevant incentive effects to be concealed behind a host of uncontrollable factors.

Table 1-2 presents those characteristics of the country's healthcare system that are the most relevant in view of the discussion to follow.

[8] An additional reason is that providers have no legal control over saved money, which makes reducing supply and cost optimization not optimal strategies providers even under prospective reimbursement schemes.

Table 1-2. Some characteristics of the Russian healthcare system

Variable	1992	1995	1997	1998	1999	2000	2001	2002
Doctors per 10,000	43	44.5	46.1	46.7	47.1	47.2	47.3	
Nurses per 10,000	115	111	111	111	111	108	108	
Hospital beds per 10,000	131	126	121	118	116	116	115	
Out-patient capacity, visits per day per 10,000	224	236	238	239	241	245	248	
New diagnoses per 1000	615.6	678.8	674.2	670.4	710	735.7	725.2	747

Source: Goskomstat 2003a (table 9.1)

Russia is relatively densely populated with doctors. The EU (pre-Accession) average is 22 doctors per 10,000 inhabitants (1994; source: Mossialos and Le Grand 1999, 120-121). Only Spain and Italy have approximately the Russian density of doctors (41 and 45 per 10,000 inhabitants). The number of hospital beds is also higher in Russia than in the EU, but Luxembourg and the Netherlands approach the Russian level (113 and 111, respectively, Mossialos and Le Grand 1999, 122).

The public healthcare employs the majority of Russian doctors and renders the bulk of medical services. Some separate private clinics exist, mostly in Moscow, but most of the private medicine is incorporated within the public clinics in form of separate departments (sometimes called *khozraschet* departments). Some state institutions (Academy of Sciences, ministries, etc.) have their owned hospital and policlinic services available to their employees. Some large companies do, as well, but part of these is financed from local budgets and the current situation is rather murky (see Antipova et al. 1998 for more details on this). More along the lines of combination of private and public care is discussed in Chapter Two.

The initial idea was to apply the 'money follows the patient' principle (see Sheiman 1998 and Sheiman 1999 for a discussion). In general, patients have the right to choose their doctors, hospitals and even insurance companies (Health Insurance Law 1991, Art. 6). In fact, none of these freedoms was fully implemented, though some regions (e.g. Kemerovo *oblast*) appear to be more advanced in this sense than others. Choice of insurer is impeded by the rule that insurance companies, where they exist, contract enterprises, which pay premia on behalf of their employees.[9] Choice of hospital is impeded by queuing rules. Policlinics are normally large units that serve large territories, so it is difficult to change primary care providers, as well.

The most recent innovations proposed by the government and experts include a right to change out-patient service provider once a year. Together with a reimbursement schedule linked more tightly to performance, this is expected to implement finally the idea of money following the patient (see Dmitriev 2004). It is worth noting that Russia has experience with high-powered reimbursement schemes (see Makarova 2000), which has already earned harsh criticism from some experts (see Sheiman 1998).

Another central concept of the Russian MHI is the principle of regional autonomy. Federal provisions define minimum standards of care, while regions decide the details (see Chapter Two, sections 2.1 and 2.2). Regions also have large authority in year-on-year administration of healthcare, though federal powers (the Ministry of Health together with the Federal Fund of MHI) perform general oversight (see Chapter Two, section 2.1 for relevant details). Historic differences among regions are one important official reason for allowing regions a lot of freedom in designing their healthcare systems. Variation in institutional design across regions should help overcome cross-

[9] As of now, the premia are collected by the federal tax service, which then transfers the funds to the

regional inequality and compensate for inherited local financial deficits and institutional inadequacies. Since part of the insurance premium goes to the Federal Fund, the latter uses the proceeds to help troubled regions, in a bid to compensate further the said historic inequalities.[10]

While discussing the institutional set-up of Russian healthcare, one should not neglect the implementation side of the issue. Whatever laws and incentive mechanisms exist, these are bound to be implemented through bureaucratic or managerial mechanisms. These are policy choices on their own. One important element is that the local and regional authorities closely control care providers (which reflects their status of branches of the state).[11] TACIS 1998 and especially Antipova et al. 1998 (a position paper on one region) contain evidence that further supports this notion.[12]

Throughout this paper I use the expression *the state* to refer to a conglomerate of regional and federal institutions. Thus understood, the state engages in procurement of healthcare alongside with other activities, and faces tradeoffs due to budgetary and other constraints.

1.2.3 The issue of under-funding. It has long been a major contention of researchers and politicians alike that the Russian healthcare story is that of an acute imbalance between the state's commitments of free healthcare and the financing accorded for the

regional and Federal Funds of MHI.

[10] This is its primary chartered goal: see the definition on www.ffoms.ru (the official web site of the Federal Fund of Mandatory Health Insurance)

[11] The literature sometimes claims that there has been ill-prepared decentralization in governance of care delivery: "The general process of government decentralisation has resulted in the decentralisation of health care administration: the vertical structure of administrative subordination was destroyed and the public health care system was divided into Federal, regional and municipal systems." (Shishkin 2000b, 7). This research points in the opposite direction: the system remains stable.

purpose.[13] Shishkin (2000b, 101-108) makes this under-funding a cornerstone in his approach to Russian health politics. He believes that the commitments themselves are not worked out in sufficient detail. As to the notion of 'under-funding', it is difficult to operationalize the notion itself. Typically, authors find that financial means accorded to healthcare, either budgetary or from the mandatory health insurance fall below the standards set by the Ministry of Health for the given planned amount of services or, alternatively, that planned budgetary appropriations are not disbursed (see Popovich 2000). Table 4.2 in Shishkin 2000b (107) quantifies this under-funding by comparing actual appropriations to the recommendations of the Ministry of Health. By this measure, in 1999, public funding was 83% of the ideal.

Using funding in the Soviet Union as a benchmark, experts estimate the drop in financing between 1990 and 1999 between 30 and 70% (see Shishkin 2000b, 101 for an overview). The variance in estimates is due to variance in deflators (see Bogatova et al. 2002, Chapter 2 for a discussion of under-funding).

The fundamental measure of under-funding should be the gap between public funding and the cost of desirable or even actual service output. This can be reformulated as a question: What level of service could have been provided, had the financing been increased? The question whether under-funding takes place or to what extent is rather difficult to answer, since healthcare requires a qualified and motivated workforce. It is only too obvious that discrepancies between planned or recommended or historical and actual current disbursements cannot reliably approximate the fundamental cost-of-service meaning of under-funding.

[12] These papers are also interesting examples of multi-pronged research into administrative detail of Russian healthcare. A document library on www.zdravinform.ru contains these and other research reports, including English versions for some.

[13] Chernichovsky et al. 1996 provide an early round up of basic data about the Russian healthcare system prior to the reform as well as a description of the initial reform steps. His account shows increase in spending over the 1980s and subsequent drop in spending at the beginning of the 1990s.

The Presidential Administration and the Government acknowledge that there exists a discrepancy between entitlement for free care and money appropriated for public healthcare. It is not clear how the government experts measure the under-funding, the authorities have promised to address the problem. Improvement of financing, better administration and planning are proposed. Recently, the Presidential Administration and the Government have suggested imposing a stricter limit on free care and transforming providers into autonomous state-owned entities which could earn profit by selling services. More flexibility in combining compulsory and voluntary medical insurance is promised, too (see Dmitriev 2004 and also Government 2000a,b).

Both the government and independent experts maintain that the immediate and most important effect of the under-funding is that patients are forced to pay the difference between the cost of care and the public I argue that this link between private and public source of funding is at least overstated. Chapter Three discusses empirical evidence to this effect and Chapters Four and Five suggest an alternative to the under-funding paradigm, whose details I discuss below in the literature review.

1.3 Literature Review

This literature review pursues a double objective. Firstly, it purports to place the research in the context of economic and public policy research on such topics as the right to medical care (subsection 1.3.1); healthcare provision regulation (subsection 1.3.2); and corruption (subsection 1.3.3). Secondly, this section reviews the literature on the very subject of payment for services free by law in post-socialist public healthcare (subsection 1.3.4). I will identify the points in the received wisdom that I want to challenge.

1.3.1 Right to medical care. It is probably a staple assertion that 'right to healthcare' does not make sense without necessary stipulations about how much, who may receive care, how much care is granted, and who will pay for the service (see on the evolution of this right concept in Den Exter and Hermans 1999). Chapter Two (section 2.2 primarily) reviews such stipulations for the Russian context in detail, focusing on the use of volumes and queues to give meaning to free care entitlement. The overall effort of the government (see Chapter Two for details) is directed at assuring some minimal standards of care subject to budget constraints.[14] Normative alternatives here may involve quality-adjusted life years gained due to successful treatment (Blomqvist 2001).

In the Russian case, the research centers on provision of a certain volume of free service and minimization of the cost of production. Some efficiency considerations can be found in empirical research on excessive capacity that can be eliminated with no effect on volumes of service (Bogatova et al. 2002). In Chapter Two I will try to shift the focus to individual entitlement to free care by paying attention to rationing according to need and medical standard as it is used in statutory provisions. I will be focusing on the controversies such a notion creates in the context of a combining of free and paid care.

This dissertation thus contributes to understanding of the limits of free care entitlement in transition societies (see, for comparison with Central Europe, Tymowska 1999, Simek 1999, Krizova 1999). The ultimate goal is however to place the Russian free care entitlement in the context of conflicting motivations engendered by combination of paid and free care.

[14] For a discussion of the topic see the volume of European papers Den Exter and Hermans 1999 and Lo Schiavo 2000.

From a political point of view, free care entitlement obviously involves difficult choices. But on the top of it, implementation of such entitlement involves reliance on cooperation of the medical profession (see Harrison 1999, notably table 2; also Lo Schiava 2000 for a discussion of the European and Garfield 1978 for that of the American context).[15] Sections 2.3 and 2.5 below enlarge on the Russian situation, demonstrating the power of the profession.

1.3.2 Healthcare economics and policy. The second broader context within which this study is to be placed is healthcare economics and policy. There are three topics on which this research bears. The first is physician agency, as it directly relates to provider-patient relationships. The second is specifics of patient choice. The third is optimal regulation in face of informational and motivational peculiarities of healthcare.

Physician agency. It is a staple in healthcare economics that a patient is not a regular customer. As Molla Donaldson 1991 puts it, the old adage ‘you get what you pay for’ does not exactly apply to healthcare. Dazzlingly complex as an economic enterprise and fraught with informational asymmetry, healthcare production demands specific trust relations between physicians and patients. Kenneth Arrow spelt out the basic economic reasons for this in his seminal 1963 paper:”[…] medical care belongs to the category of commodities for which the product and the activity of production are identical[.] (Arrow 1963, 949).”

[15] George France emphasizes that Italian courts favor professional competence over freedom of choice of provider or doctor, except for GP’s (France 1999, 45).

Citing empirical evidence to the effect that medical professionals do not behave as profit maximizers, Arrow identifies some consequences of their specific professional motivations:

> The unusual pricing practices and attitudes of the medical profession are well-known: extensive price discrimination by income (with an extreme of zero prices for sufficiently indigent patients) and, formerly, a strong insistence on fee for service as against such alternatives as prepayment […] Price competition is frown on. (Arrow 1963, 953-954)

In this dissertation the 'benevolence of the doctor towards the patient' is used as a catch-phrase to denote the host of motivations that prevent professionals from profit maximizing and are more or less clearly different from fear of punishment for violating the terms of their job descriptions (Chapter Three). Dealing with motivations, one should be careful not to assume too much knowledge of this murky area. Kessel 1958 argued that the observed 'benevolent' behavior of the doctor was in fact related to a need to placate public opinion. In response, Arrow leaves open such a possibility. There may be a host of other different motivations behind what effectively is the doctor's benevolence. This dissertation recognizes complexities of motivations and abstracts away from psychological signature of economic decisions.

Specific details of professional behavior extend beyond the sphere of professional-patient relations. As Arrow himself says, such things as price fixing and other corporate behavior also belong to the realm of peculiarities of healthcare market. The differences between American paid care and Russian paid care are too large to make a direct comparison meaningful. But the very issue of corporate behavior will necessarily emerge, also in Chapter Three.

In general, the notion of profession as an economic phenomenon applies to healthcare. Mattheus 1991 discusses specifics of professional behavior, including monopoly practice, price fixing, and, certainly the responsibility that comes from superior intelligence available to a professional as compared with the professional's customer.[16]

Eisenberg 1986 (as quoted in McGuire 2000) linked psychological and professional satisfaction with the experience of functioning as the patient's advocate:

> A substantial part of the physician's satisfaction with practice is fulfilled by serving successfully as a patient's advocate. (McGuire 2000, 522)

The concept of advocacy will be used extensively in what follows. It will be shown to apply in a number of contexts, due to the circumstances of "gray" market.[17]

The literature on healthcare economics suggests that constraints on doctors' achieving certain income targets affect relations between them and their patients. McGuire notices:

> There is abundant evidence that in some circumstances physicians are prepared to trade off income against welfare of the patient. Furthermore, this tradeoff is affected by income effects, in a manner consistent with conventional views of labor supply. (McGuire 2000, 526)

In particular, increased regulation of prices and quantities may result in the physician's behavior approaching that of profit-maximization. The avenues for showing benevolence narrow and income objectives become the overriding concern. One way of demonstrating how this can happen is to let physicians pursue certain target incomes. Rizzo and Blumenthal 1996 supply empirical evidence in support of income targeting by physicians, as opposed to profit maximization. Then the regulator faces a choice between the following two policies:

[16] In legal terms, Mattheus suggests the customer cannot be expected to check quality of service or product bought and the principle *caveat emptor* is replaced by *uberrima fides*.

[17] See Garfield 1978: conscientious doctors may even shield the patient from institutional inefficiencies, thus assuring steady supply of quality care even in absence of suitable institutional

- Loose regulation and reliance on benevolence;
- Stringent regulation that turns a doctor into a profit-maximizer.

The non-commercial nature of medical enterprise also affects the practices of provider institution (though its basis may not necessarily lie in professional norms, but general social norms or even tax incentives). Non-profit institutions play a great role in providing care in all of the Western countries (see Newhouse 1991 for the U.S.). Non-governmental for-profit healthcare providers were found cross-subsidizing across patient groups, as well as providing emergency care for those incapable of paying, which may also involve cross-subsidy (Kovner and Neuhauser 1978, 422).

At the same time, there is evidence that providers will jeopardize patient health unless due professional diligence is rewarded with money. Wallet biopsy, as Carr (1982) affectionately calls the procedures, may lead to the transferring of a poor patient to a public hospital after providing some basic emergency treatment. This transfer may, Carr asserts, happen at the expense of the patient's life. This counterbalances the idea of benevolence of medical professionals or non-profit nature of medical enterprise in general. Closely related to this is the problem of Medicare patients in the U.S. Hospitals are known to avoid such patients, at least if they face sufficient demand from those who are able to pay more (Carr 1982)

The formal model in Chapter Five makes benevolence an important factor in policy choice, while Chapter Four embeds benevolence in a framework of provider-patient relations. Professional agency issues emerge throughout the dissertation, starting with Chapter Two and especially in Chapter Three, section 3.3, where I attempt a conceptualization of specific patient-professional trust relations for the Russian context.

support. If this is relevant for the U.S., it should also be relevant for Russia, where institutional

Patient choice. As Evans (1983) observes, it is inconsistent to maintain simultaneously that (1) a patient can make a meaningful choice regarding satisfaction of his or her health needs and that (2) regulation of healthcare provision is absolutely required. It is almost a consensus that the first proposition does not hold:

> Professional self-regulation, direct regulation of providers, and universal public insurance are all based on the belief that uninformed and unsubsidized consumer choices in an unregulated market will lead to patterns of utilization which do not correspond to needs. (Evans 1983, 359)

If patient choice is not guided by a free care entitlement or public subsidy, it will be governed by professional advice, possibly affected by professional self-regulation. Because professional benevolence is not absolute, there is a need to balance professional advice with external forms of regulation. Much of the policy debate both in Europe and North America deals with the exact shape of this balance. Immergut 1992 explains how the history of European healthcare systems is rooted in curbing independence of the profession, while Carr 1982 shows how American efforts in the same direction largely failed.

This notion of patients in need of a guiding hand has rarely been taken into account in the Russian policy discussion. The "money follows the patient" principle embodies exactly the opposite view. The notion that patient pays the doctor because the state does not also stands in substantive conflict with the view of patient as more or less guided by either doctors or regulators (see elaboration of the concept of under-funding in subsection 1.3.3 below), unless medical professionals are extremely benevolent. Limited nature of patient choice in healthcare will play a great role in this Dissertation (see sections 3.3 and 4.2).

framework is traditionally considered as wanting.

Principal-agent problem in healthcare. The literature contends that patient choice is limited by the nature of healthcare. The government's ability to successfully regulate and/or procure healthcare is also limited (see Kessler 1978 (421) on the American experience of hospital accountability to the government). The state cannot simply order the provider to deliver the first-best quantity and quality of care.

This research deals with a combination of private (direct) and public (third party) financing of healthcare. This means that provider incentives come from two sources: private and public. Moreover, these incentives are of two types: financial and administrative. Chapters Four and Five directly address the issue. Informational rent earned by providers of care due to informational asymmetries is in focus.

At the most abstract level, this is a principal-agent problem. Fudenberg and Tirole 1993 (Chapter 7) and Osborne and Rubinstein 1995 (Chapter 10) are classical references for the general theory of incentive contracting and the theoretical limits of efficiency. Laffont and Tirole 1993 (Chapter 2) is based on a procurement model that I use in Chapter Five of this dissertation.

Applying incentive contracting to healthcare systems faces sector-specific limitations. In particular, medical care can be measured in terms of quantity as well as quality of service. Alternative incentive schemes give opposite incentives along these two dimensions.[18] Pauly 2000 and Chalkey and Malcomson 2000 present an overview of findings on provider reimbursement in healthcare.

[18] At one end of the spectrum there is prospective global financing, where provider performance does not affect the amount of financing. The provider has incentive to minimize costs of service, but then may compromise quality. At the other, there is cost-plus retrospective financing, giving the provider an incentive to inflate the amount of services rendered disregardful of costs incurred, but possibly ensuring high quality of service. Of course, this is a gross simplification of the actual range of possible schemes and their effects.

Peculiarities of healthcare transpire in the fact that the actual institutions of procurement and/or regulation of care combine many elements (Freeman 2000, especially 86-103 on professional agency). Table 2 in Harrison 1999 (89) displays the mechanisms that the British National Health Service uses, from arm's length incentive management to auditing and quality standard enforcement. In the same table, a set of responses from the profession is presented. The responses include compliance, limited cooperation, and resistance. All this bears marks of agency problems and the power of the profession to exercise influence on policy implementation: "Doctors as gatekeepers and prescribers are key in the determination of health care costs." (Kanavos and Yfantopoulos 1999, 188). TACIS 1998, Antipova et al. 1998, and Ivánova et al. 1999 provide much Russian material on administrative and economic regulation of healthcare. A detailed description of actual administrative practices is particularly helpful for understanding the nature of payments for services free by law.

One general healthcare regulation topic that is especially close to the subject of this research is cost-sharing, widely used in Western Europe in many forms (see an overview in Mossialos and Le Grand 1999, 75-81). Another related way of regulating care provision is a policy of private financing of care, which exists in Portugal and, in a less systematic fashion, Greece (see Pereira et al. 1999; Sissouras et al. 1999, 365-366). As will be shown in Chapters Two and Three, the Russian case is effectively a specific mixture of the two ways of involving private resources in (co-)financing public care.

1.3.3. Corruption and petty corruption. The existence of an institution that is not supposed to exist by law is the subject matter of this research. This research therefore belongs in the field of corruption studies.

Giving the phenomenon of paying for services free by law the name of ‘corruption’ immediately implies that the proceeds illegal rent. In other words, private payments are not covering the minimal cost of service, but increase personal incomes of service providers. Such a rent, whose equivocal legal status follows from the discussion in Chapter Two is identified in Chapter Four, while Chapter Three supplies evidence that suggests domination of the rent component.

The literature on corruption suggests three topics of relevance in this context. First of all, collecting extra payments from patients appears to resemble what is often called petty or ‘street’ corruption. Indeed, a number of empirical researches on petty corruption included payments to doctors, nurses, and orderlies as one class of petty corruption. Miller et al. (2001, 259) provide statistical evidence to the effect that health services are more corrupted than policy, education, welfare system, and judicial system in Ukraine, Bulgaria, Slovakia, and the Czech Republic. To be more precise, Miller’s comparison is made on the following counts: readiness to accept small gift, experience of recently accepting a small gift, readiness to accept money or an expensive gift, experience of recently accepting money or an expensive gift. On all four accounts, healthcare appears to be the most corrupted. Certainly, the word ‘corrupted’ must be treated with caution, due to different nature of seemingly similar transactions in different institutions.

One alternative explanation of the statistical result is that doctors and others in medical establishments are more open about their true intentions and past deeds than police or university professors. This would be compatible with the intuitive picture of ‘corruption’ in healthcare as being little frowned upon or even totally excused in difference from charges for what is free by law in other institutions.

The research has also identified a number of peculiarities of petty corruption, including a fuzzy border between illegal rents and tips; payment without a promise of favors to come in exchange; and the apparent role of customs and common expectations. I will show that these features are relevant for payments of interest (Chapter Three, notably subsection 3.4.2).

The second topic is the economic function of corrupt dealings in particular institutions and environments. Rose-Ackermann (1975, 1999) suggests four main functions: market clearing, incentive effects, lowering costs of business or reducing uncertainty, and facilitating illegal activity. Payments in healthcare, if directly related to production of services, ostensibly clear the market and provide incentives for better work. There are also ways, to be discussed in Chapter 3, in which they reduce uncertainty. The problem of enforcing corrupt deals remains:

> Firms pay bribes to obtain certainty, but the certainty may be illusory, because they cannot enforce corrupt deals. (Susan Rose-Ackerman 1999, 17)

As all corrupt deals, illegal payments in healthcare should be linked to factors or mechanisms that suggest at least some level of enforcement. This is another problem discussed in Chapter Three, section 3.3.

This research is a description of how an institution functions, rather than a normative assessment of how it should function. Nonetheless, theoretical controversy over corruption's welfare effects carries over to payments in healthcare. This is the third topic linking this research with the literature on corruption.

Some authors claimed that corruption may increase economic efficiency, others created models where it did not. In fact, it is obvious that corruption can increase efficiency in the Pareto sense. When a law is imperfect, violating it can easily produce efficiency improvements (see Rashid 1981 for an intricate example of this).

Efficiency effects of corruption will generically depend on legal and institutional frameworks under consideration Khan 1996, Mehmet 1996). The effects of private payments (analyzed in Chapters Four and Five) may or may not increase welfare, depending on specific interactions between medical professionals and paying patients.

When assessing effects of private payments, I have much benefited from the rat-race model of inefficient equilibria (see Galasi and Kertesi 1989 for a rat-race scenario of corruption and Landers 1996 for a labor market application). Chapter Four, section 4.2 describes adverse effects exerted by paying patients on other patients through a similar mechanism.

1.3.3 Research on charges for legally free healthcare. The phenomenon itself has been mostly discussed in relation to socialist and transition economies. It would however be unjust to claim that it is geographically confined to Eastern and Central Europe and the former Soviet Union. There are two ways of defining the phenomenon, a ‘narrow’ one and a ‘broad’ one. The narrow version is based on the following scenario. Patients slip money into doctors’ pockets in expectation of additional quality or quantity of service. The broad version includes all scenarios of public healthcare patients being coaxed or forced to pay for healthcare. This Dissertation treats the problem in its broader version. Close attention paid to the status of legal paid care is one element of favoring the broad definition. The literature seems to favor the narrow version, but there are important exemptions from the rule, as will be seen shortly.

In its broad version, the problem of paid and free care supplied by a provider to same or different patients is known probably in all, even the most successful healthcare systems. Probably in every healthcare system, doctors or institutional providers face the problem of discriminating between groups of patients that differ as sources of income for the providers of care.[19] The British case of such a conflict used to be the subject of political tensions since the inception of the NHS. Since doctors were allowed to keep pay beds in hospitals, there was a suspicion, somewhat founded on facts, that the principle of care rationed by need may at times give way to the principle of rationing by payment.[20] Barbara Castle, Secretary of State for the Social Services in 1974 voiced the problem most emphatically:

> The issue before us is whether the facilities of the NHS, which are supposed to be available only on the principle of medical priority, should contain facilities that are available on the different principle of ability to pay. We say that those two principles are incompatible in the NHS[…] (communicated to Parliament in 1974, as quoted in Klein 2001, 90)[21]

Among other EU countries, in Greece patients entitled to free care often turn to private options. This may happen, for example, when a doctor transfers a patient to his or her private practice. Though currently non-legal, private practices at public hospitals still exist (Sissouras et al. 1999, 365-366).

Balance billing (charging a patient on the top of third party (public) reimbursement) is another way of combining private and public funding at the level of provider-patient interaction. Officially prohibited in Russia, it is legal in France. Pauly 2000 contains analysis of the practice of balance billing in the U.S. In a sense, the Russian practices of charging a patient for care free by law will be shown to resemble

[19] In the US, as has already been said, a hospital will make effort to get rid of Medicare patients. Senator Kennedy's 1978 national insurance was directed at making all HMO patients equal (Carr 1982, 413).

[20] Cullis and Jones (1984) consider NHS waiting lists in the context of patient choice between paid and free care.

[21] The ideological conflict here is further emphasized by an underlying religious metaphor. As Barbara Castle pointed out on another occasion, the NHS is much like a church in that it must be based on principles completely free of a conflict of interest of the kind described: "What would we say of a person who argued that he could only serve God properly if he had paid pews in his church" (Klein 2001, 90).

balance billing, though in this case there is no official reference price above which a patient will be billed.

What has been found in Greece, has direct relevant to the public healthcare in Russia. Replacing free care with paid services has profound consequences for the substance of all healthcare issues. The conflict of interest similar to the one identified for Britain is rooted in the current legislation, as Chapter Two argues. This legal dimension justifies considering the problem of payments for things free by law in its broad rather than only narrow version. And it is this legal dimension that has so far been missing from the analysis in the literature.

The problem of paid services crowding out free ones is mostly relevant to Central and Eastern Europe and the former Soviet Union. Studies on Russia mostly concentrate on proving that patients pay for services free by law (see Chapter Three for references to empirical studies). A few authors went beyond the existence claim in a bid to capture the essence of the institution. I will formulate their findings and controversial elements thereof that are to be challenged in the subsequent Chapters.

First of all I will consider Russian literature on the institution of payments for services free by law. It appears that the literature contains two main conceptualizations of the phenomenon.

The first is a characterization of it as tips or customary and voluntary 'gratitude payments'. This means that the payments are voluntary and have limited economic consequences per se. Services would have been discharged in absence of the payments and the latter create little distortion. The second conceptualization views the payment as a means of supplementing the inadequate public funding. In this sense, the payment itself does not change the distribution of costs and benefits as compared to the officially intended. The private payment only makes up for the means of achieving this intended distribution and associated levels of service production. This dissertation offers an alternative to both these concepts: the private payment affects production and distribution of services.

These two concepts are a simplification of what exactly the Russian authors say about the payments of interest. It is worthwhile seeing how research evolved in response both to the changes in the economy of the country and, apparently, to the conditions in which the research was being made.

The early literature on what happened in Soviet and Russian healthcare contained a ‘benevolent’ view of the phenomenon of payments. The latter are tips or gifts of appreciation, often forced on doctors by grateful patients.

The existence of an extensive tradition of gift giving was recognized in the Soviet literature. Suk (1984, 36-37) relates the following two ethical paradoxes. Accepting a gift demonstrates the physician’s placing value on his or her service. Refusing it signals the low quality of the professional service. The second ‘ethical paradoxical’ account of why it is even ethical for a physician to accept gifts is that refusing to accept them may signal to the patient that there is no hope of recovery. These two signaling stories must be balanced against the idea among the more conscientious Soviet doctors that a doctor shall never be remunerated in any form by the patient.

The reality was nastier, as evidence cited by Ryan (1990) suggests. Stories of extortion and exorbitant prices paid even for regular and simple services indicate that the reality of Soviet healthcare included more sinister things than gifts of appreciation, which one could find it even unethical to decline. Ryan (1990, 25-28) calls the payments for services that were definitely for free in Soviet Union “bribes, gifts, and extortion”. He even suggests that the authorities were concerned about extortion practiced by medical staff in late 1970s. Such practices were increasingly branded as a form of corruption. At the end of the day, Ryan's account balances the 'gift' and 'bribe' versions of conceptualizing the payments, which in any case were illegal at the time.[22] The low pay of the physician and gratitude of the patient are quoted together as explanatory factors.

As the economic crisis unfolded, together with the reform of public healthcare, another conceptualization emerged for payments for services free by law emerged. At the level of facts, research suggested increasing private contributions (see Chapter Three). At the level of interpretation, these allegedly increasing private contributions were claimed to play the role of full payments for services, since the state had reneged on its financing obligations. One particular feature of the situation, namely use of legal contracts to cover up illegal charges, reinforced this notion of the payments as full scale paid healthcare (Boiko et al. 1998, 2000a, 2000b). Sheiman (1999) claims the following:

> The package of medical benefits under mandatory health insurance (MHI) is not balanced with the financial resources available. Providers have no explicit obligation to the authorities to provide free care provision. This allows them to charge even for services, which they can and must provide free. Being unable to fulfil its financial obligations, the government has to accept these rules of the game. As a result, billing practices have gone out of control. (Sheiman 1999, 104)

On his view, the economic transition was accompanied by transition from tipping to full-scale paid care, and this was because of drop in public financing (see Shishkin 1998, Sheiman 1999, Shishkin 2000a, b). This picture, per se, looks simplistic and the authors hasten to acknowledge that links between public financing and private payments were more complex (see Shishkin 2000a). In particular, private payments in healthcare could be a reason for the authorities not to increase budgetary appropriations.[23]

[22] Most explicitly so since a 1981 amendment to the Penal Code.

[23] Thus allowing the creeping replacement of free care with paid care (Sheiman 2000, Shishkin 2000a).

The evolution has thus been made towards an acknowledgement of the existence of a complex phenomenon with uncertain welfare effects. And yet, under-funding has remained the centerpiece of the argument. It appears that the best expression of the resulting 'ideal type' construction covering many causal links is found in Klyamkin and Timofeeva 2000. Poor doctors and poor patients solve their problems by negotiating some flexible framework within which services free by law are effectively sold. According to Klyamkin and Timofeeva (2000, 165) this gray market operates without any common rules and norms. Bogatova et al. (2002, 19) disagree with the latter statement, but share the basic idea of replacement of free care with paid care in the circumstances of economic transition.

Klyamkin and Timofeeva's notion of private payments in healthcare as a response to inadequate public funding sits well with the general claim of their book that Russia's corrupt institutions have become a state unto themselves. Though healthcare payments are not exactly corruption, both phenomena stem from a failure of the official institutions, either to finance or control or both. With regard to healthcare, I shall call the approach *the under-funding paradigm*.

The ideal type of 'poor helping the poor' and under-funding as the primary explanatory factor represent the balance of concepts that is central to both theoretical and political discussions of payments in public healthcare. The above mentioned interview with a deputy prime-minister (Dmitriev 2004) shows that for the authorities the notion of alleged imbalance has acquired central importance, possibly because promising money is the easiest way to show concern and breed hope. Popovich 2000, Makarova 2000 (see page 23) and Chernez et al. 2003 (as the most recent reference) demonstrate the same for the expert community.

The under-funding paradigm ignores issues of professional agency in favor of 'macro' and policy factors. At the end of the day, according to the paradigm, paid care is a reaction to lack of free care. Thus this ideal type construction omits an important element of the reality, namely, professional agency. This dissertation pays close attention to how exactly professional agency changes the picture.

A related deficiency of the under-funding paradigm is that the absolute level of money flow is taken too literally as a major factor. The idea is apparently that had the patient been richer, the patient would have bought healthcare services on some open market. Similarly, had the doctor been richer, the doctor would not have charged the patient, possibly accepting some tip or gift of appreciation. These two elements appear quite questionable. It is not the mere existence of the causal links they purport to capture but the overriding significance ascribed to these links which is in doubt. Professional benevolence has been acknowledged in the literature, but likewise have monopolistic pricing, denial of care and exploitation of patient ignorance. The following Chapters purport to redress the apparent imbalance by taking into account not only the absolute level of public money, but also the exact reasons why patients may want to pay.

It is time to bring in some international perspective. The phenomenon is widespread in Central Europe as well, in particular Hungary. János Kornai and collaborators undertook a research in the matter. The picture drawn by Kornai putatively favors the broad notion of payments for health services free by law, though the basic story remains that of 'money in an envelope' (Kornai 2000). Kornai and Eggleston (2001) identified two basic forms of forcing the patient from the free into paid care. The first is making the patient pay a 'gratuity', or, using the Hungarian term, "gratitude money". Similar to tips, this gratuity can play the role of a genuine gift of appreciation. With the emergence of legal paid care in Hungary, the second strategy appeared, the more complex one:

> A patient arrives at a doctor's private office and pays the doctor, formally as a fee for the visit. In fact, the patient wishes to purchase a privilege by making the payment. He/she expects special attention from the doctor at the latter's main place of work, a state hospital, or clinic – help in jumping the queue […] and so on. " (Kornai and Eggleston 2001, 169).

As the authors immediately remark, this second strategy is bound to depend on the particulars of provider reimbursement arrangements. The Czech system does not leave space for the Hungarian type of illegal mixture between public and private care. Kornai 2000 suggests that the major cause of persistency of the phenomenon after the collapse of socialism is that "bureaucratic and market coordination, and public and private ownership, combine in a sometimes healthy and sometimes distorted way. Gratitude money fits into this ambiguous environment." (Kornai 2000, 8)

The Russian situation should also be viewed as a regime of cohabitation and interaction of free and paid care. The elements of this interactive regime are discussed in Chapters Two and Three. The under-funding paradigm does not pay its due to reflect such interaction.

Kornai (2000) shares the idea of low public financing being an important factor in making the system of payment persistent. However, he immediately acknowledges the possibility of reverse causation. For the socialist economy, he suggests that illegal charges were overlooked by the authority, for the charges were an alternative to raising wages (Kornai 2000, 7). Currently, there are vested interests in maintaining the system of payments (ibid., 16), which potentially effects the same reverse causation.

János Kornai makes a special effort in emphasizing that provider-patient relations over 'gratitude payments' are not the classical clearing of a market of services. He displays evidence to the effect that it is somewhat irrational for a patient to pay. The results quoted in Kornai 2000 (tables 2 and 3) include the following interesting facts: forty-one percent of doctors believe that personal connections are more important than money for the patient to get better treatment. Half of the doctors wholly or partly agree that 'gratitude money' does not make difference to treatment (as against thirty percent of the public). Finally, specialty, rank and reputation play the most important role in determining the price, and not the performance of the doctor.

Conceptualizing the peculiarity of gratitude payments, Kornai suggests that due to lack of transparency the market for services paid for with 'gratitude money' is extremely distorted:

> On the kind of concealed market in which gratitude payments are made, the lack of transparency means that the process fails to converge on an [efficient] equilibrium price. {Kornai 2000, 10)

It seems that this treatment of the problem is incomplete. Why do patients pay? And then, the process does converge somehow to a price, even if not in the sense of the microeconomic textbook clearing of a market of services. An attempt to answer this question on the basis of specific roles played by physicians is made in Chapter 3. Physician agency features most importantly in this research, and peculiarities of economic behavior of the medical profession lay ground for partial rationalization of the patient's behavior.

One problem discussed in both Hungarian and Russian literature is that of burden of charges for care on households. This problem is not of concern in this Dissertation. All issues of equity and fairness of price were consciously avoided. The reason for this is that substantive assessment of these issues cannot be done with the data available.

In addition, the research pointed out public complacency with the fact that medical professionals charge for things free by law. According to Kornai and Eggleston 2001 (173), in Hungary, gratuities are viewed by majority of both physicians and public as a necessary evil. The doctor has a moral right to accept them (according to 80% of the doctors and 70% of the public). These payments are not detrimental to doctor-patient relations (65% of both groups). About 50% of the physicians asked (60% of the public) believe that payments do not create differences in quality of treatment.

1.3.5 Conclusions to the literature review.

The theoretical and empirical research points out to many peculiarities of health care as an economic sector and are of public policy, notably specific agency relationships. These are to be taken into account when discussing payments for services free by law. Further insights are found in the literature on corruption, according to which payments in healthcare may be regarded as petty corruption.

The literature on payments for services free by law in Russia appears to view the phenomenon largely as a result of economic transition and associated crisis, as well as persistent poverty and under-funding of care by public sources. In this framework, paid care replaces free care through legal or illegal mechanisms. I shall challenge this notion in the chapters to follow.

Chapter Two. Rules And Procedures Regarding Combination of Free and Paid Care

2.1 Purposes and Methods

Russian providers of free medical services have the right to deliver paid services.[24] This general provision is accompanied by a number of restrictions and rules, among which there are requirements to acquire license, observe accounting rules, and pay taxes. Prices are also regulated. Proceeds are distributed according to certain rules that cap the portion available for personnel remuneration.

This chapter answers the following question:

How does the statutory delineation of a border between free and paid services affect payments for services free by law?

To answer this question I have closely examined rules and procedures that affect this border. This is not to say that a regular legal analysis is to follow. Statutory provisions that underlie the institutions of public healthcare are enforced and disputes are resolved not through courts of law, but through internal complaint mechanisms and bureaucratic procedures.

A patient comes to a clinic upon reading the published rules and procedures regarding his or her individual rights. What is this patient to expect, given that courts of law are excluded from the picture? This is the perspective from which I shall look at the material of interest.

[24] They are not autonomous business units and the state is fully responsible for their financial affaires. They cannot be bankrupted, as a result. But this consideration will not affect the discussion in this Chapter: all conclusions here will remain in force even if all these providers turn into autonomous state-owned or privately owned enterprises, or public companies.

Why is it important to review the rights of patients together with the regulation of paid care in research on payments for services free by law? Foreshadowing the discussion below, one will see that permitting delivery of both free and paid services creates a conflict of interest. A medical service provider has an incentive to charge for what is to be provided free of charge if that adds to the provider's profits. Upon describing the regulatory framework, I therefore focus on the conflict of interest as the formal institutional and legal background against which the phenomenon of payments for services free by law must be appreciated.[25]

Three components of the analysis to follow are:

1. provisions for free care;
2. provisions for paid care;
3. public policies that pertain to regulating the conflict of interests.

Once I explain the relevant rules and procedures, I shall attempt to formulate the ensuing balance of incentives for medical service providers. At this point, payments for services free by law enter the picture, because I show that providers of medical services enjoy enough leeway to force patients into paying for what is free by law. Whether they indeed do so is a question for Chapter Three.[26] The research question of this Chapter can be reformulated as follows:

Assuming that a conflict of interest is demonstrated, what is the precise balance of incentives the institution of public healthcare gives providers?

[25] The following Chapter Three argues that if patients are forced to pay for services free by law, it is not necessarily the case that they do it completely informally. This is an advance note of the empirical evidence making this query on formal rules relevant.

[26] A clarification that is in order here concerns the difference between institutional provider (hospital, policlinic, or 'local medical organization' that combines the two) and individual professional (doctor, nurse, orderly) as regards their participation in the alleged illegal charges. This Chapter deals strictly with the ability to charge improperly, while remaining immune to prosecution due to bad regulation of institutional providers. The above difference may not be of high importance because regulation of individual professional behavior is not developed to the point where it could have been considered as a factor here. The subsequent Chapter contains a putative picture of the payments of interest that partially accounts for relationships between institutional providers and salaried professionals.

A terminological remark is in order. It could be misleading to use the term 'illegal' to define payments for services free by law. It is easy to imagine a situation, when charges for something free by law are not illegal charges. If a restaurant provides a banquet dinner, its participants are not expected to pay, because the organizer or sponsor does. Yet there appears to be nothing wrong with a participant preferring to buy a certain dish from the restaurant instead of taking it free of charge. This is why the word *illegal* will be used with extreme caution. More reasons for the caution are forthcoming, for the rules and procedures to be reviewed add further vagueness to the border between legal and illegal.

Section 2.2 explains the statutory commitment to provide care free of charge at the point of use. Section 2.3 looks at the basic provisions for paid care. Section 2.4 looks at the last component, namely those policies that directly regard the issue of payments for medical services in violation of the free care entitlement. Conclusions are drawn in Section 2.5 regarding the balance of incentives providers of medical services have under the current institutional arrangements, and the definition of a regulatory gap in provisions for free and paid care.

2.2 Entitlement for Free Medical Care: Volumes and Rationing of Free Services

This section has the limited ambition of providing a guideline into the legal underpinnings of the right to free healthcare in the Russian Federation.[27] It is the general approach in Russian healthcare policy that patients can expect to receive a certain amount of care free at the point of need. In a world of scarce resources, the

[27] To the best of my knowledge, this rather elementary task has not yet been performed in the literature. The existing accounts of the legal framework do not clearly represent patient rights to care.

problem of defining the right to free healthcare is closely related to that of rationing. This means (Locock 2000, 92) that the following three questions are to be answered:

1. What services should be available?
2. How much of each service should be provided?
3. To whom shall the services be provided?

The issue is to determine how the overall financing, both in terms of total quantities and ways of distribution of funds, translates into satisfaction of individual claims to free healthcare.

The following paragraphs describe in more detail the composition of free healthcare as regards menus and volumes. The ‘framework’ mandatory healthcare insurance law provides a clarification on this point. It reads:

> Citizens have the right to a guaranteed volume of free healthcare, according to the programs of mandatory health insurance. (Health Insurance Law 1991, Art. 20)

The notion of volumes of free care is introduced, but reference is made to a further elaboration. The latter was provided in regional programs until 1998, when the Federal Government suggested a set of standard volumes of care per capita. The Federal Program of free healthcare (Government 1998, Appendix V) determines the mentioned volumes, both covered by MHI and regional and federal budgets

These standard volumes are volumes per 1000 inhabitants and serve as guidelines for regional authorities and MHI funds, which define volumes of care by possibly more refined categories of diagnostic cases, establishments or on a geographic basis. The federal program aggregates the volumes at the level of 'out-patient', 'in-patient', and 'emergency' care. Regional regulations follow the suit adding further details sometimes.

In the literature, the 1998 definition of guaranteed volumes of care is sometimes called the second reform of public healthcare (Linnakko 2002, 4). Yet it seems that this general definition of volumes only codified the practice that predated it. The following consideration shows that this notion of volumes is not very consequential. It must be acknowledged, however, that regional programs of healthcare provision have been modified to suit governmental suggestions since 1999. In particular, calculations of volumes and expenditures have become more explicit and detailed.

According to Health Insurance Law 1991 (articles 24, 15) the rationing of the financial resources is to be decided upon jointly by local or regional authority, recognized self-regulating organizations of the medical profession and regional Mandatory Health Insurance Fund. Private insurers have the right to provide their recommendations. This joint decision determines tariffs, prices for procurement purposes,[28] for certain items of care, such as bed-day, various manipulations, operations, use of equipment, etc. In 1993 a regulation issued by the Federal Fund of Mandatory Health Insurance (see Federal Fund of Mandatory Health Insurance 1993) made this framework regulation more precise by establishing a procedure of reaching agreement on tariffs.

So far, volumes of care supplied under Mandatory Health Insurance arrangements have been discussed. Some regions finance some items (for example, emergency care of and delivery of patients to hospitals, dental care) from regional budgets. Volumes of these items of care are determined by regional authority, and seem to be essentially limited by the size of financing, following certain common medical standards and federal volumes of free care guarantees. For example, the federal regulation provides the total volume of guaranteed emergency care in terms of number of calls attended.

[28] Hence the name '*tariffnoe soglashenie*', or 'tariff agreement', which is sometimes used in this context.

The actual performance of a regional system of emergency care is obviously affected by the actual rather than promised financing.

Similarly, the last component of free care, the items provided under various federal special programs, notably those related to high technology, cancer, etc, constitute the case of volumes directly determined by the (federal) financing accorded for the purpose. Moscow clinics are then obliged to receive a certain number of patients from regions for specialized operations.

To see how this works in practice, consider the following example. Suppose that there is a specialized centrally financed clinic in Moscow. This clinic is supposed to take patients from all over the country according to certain quotas. This principle generates various conflicts between regional and central health authorities regarding the number of patients from, for example, Yaroslavl' *oblast*, to be treated in this clinic. The example shows that rationing problems and procedures are not confined to patient-doctor relations. They pervade the system, even affecting relations between federal and regional authorities.

For a specific example, consider Table 2-1 below.

Table 2-1. Planned volumes of care per 1000 insured (practically equivalently, population), 2000

Category	**Volume per 1000**	**Cost of one unit, rubles**
Emergency visits	318	70.6
Policlinic visits	9198	34.2
Bed days	2812	200.3
Hospital admissions	198	N/a
Average stay at hospital, days	14.2	N/a
Bed occupancy, %	94	N/a

Source: Medical News, vol. 24 (139), 1999, quoted from Linnako 2002 (table 2.5)

These are all federal provisions, which go together with various recommendations, regarding medical, administrative and financial aspects. The document involves a breakdown of expenditures into labor, medication, and overhead (see Linnako 2002, tables 2.6, 2.7). Notice that all this mixes the three sources of financing: Mandatory Health Insurance, regional and local budget appropriation and federal budget appropriations.

Regional programs provide a more detailed assessment of these volumes. As an example, a Rostov *oblast*'s Program (Rostov 2001) specified volumes for 40 types of wards, classified according to the type of clinic and age group of patients (children or adults). The unit of volume is bed-day. Reading the document gives a clear picture of the planned use of all regional facilities, together with money rationing. There are certain patient groups (psychiatry, drug-addiction, tuberculosis, sexually transmitted diseases), which are financed only from budgetary resources. Others, apparently, have mixed financing.

Municipal and local programs are supposed to provide further detail for these planning arrangements. It is impossible to ascertain to what extent they do in practice. At this level of detail, variation in implementation of the plans shall be expected to turn rather material. Open documentary sources are not providing much information on this and the existing research is clearly insufficient to draw conclusions (see Ivánova et al.1999 on Republic of Chuvashia's healthcare governance as an example of the feasible level of detail).

The price of each item of care is defined according to the volume and overall financing available. In fact, of course, it could be that overall financing and price of each item of care are the primary data and volumes are results. But the difference does not matter here, even if in some sense it is real. What matters is that regional financing and MHI financing are mixed at the level of regional plans. It has been mentioned in the Introduction that regional budgets cover not only capital investment, overhead, and special programs, but also salaries and medication. Now one sees that the planning and management process explicitly mix the two sources of financing.

This notion of volumes does not put any limit on the individual entitlement for healthcare. Whatever the proclaimed role of the volumes (see Shishkin 2000, 306-307), their actual role appears to be that of planning. These volumes are of no interest to an individual patient nor to an individual professional. This is to be confirmed by looking at what exactly a patient is to expect from the public healthcare.

How do the aggregate volumes translate into individual rights to healthcare? The volume of curative, preventive or diagnostic measures for a particular patient is determined by the medical professional "according to medical standards or reasonable minimum" (a phrase typical for all relevant documents, also Government 1998). Anything in excess of that standard must be justified separately in writing and can be subject to expert inquiry.

Further rationing of healthcare is by queuing and urgency. Queuing is the major way of distributing the overall capacity and effort of the healthcare system among the patients, with corrections for emergency cases. Various regions ration healthcare with differing degree of detail. The waiting time for both out-patient check ups and tests and for hospitalization is limited in general. Maximal terms of waiting for hospitalization depend on the type of the medical institution and cannot exceed four months (Government 1998). Yaroslavl' *oblast* sets the limit for the latter at three months. A patient has the right to know the approximate date when he or she will be hospitalized (Yaroslavl' 1999, Art.4.1). Out-patient queuing is regulated by setting limits to waiting time (Yaroslavl' 1999, Art.1.2).

Various rules are imposed on doctors as to attending to a patient needs, first of all by setting time limits (for example, maximum 8 hours of waiting for a doctor visit).

Emergency cases are of course treated separately. For example, in the out-patient sector, the following categories of patients have the right to jump the queue: those with a fever, those with symptoms of life-threatening diseases, and those with certain privileges, such as war veterans. Separate queues must form for access to high technology and for admission to federal medical establishments.[29] Probably there is regional variation for all these arrangements.

In summary, urgency takes precedence over queue, while patients having urgent needs would form a queue among themselves. Hence the queues for expensive operations and access to specialized facilities. Regions may impose additional rules for queuing and other forms of rationing.

There seems to be no legal right for the provider to deny a patient a service on the grounds of the 'volumes' being exceeded. One reason for this is that patients should be treated equally. Another more technical reason is that volumes are not sufficiently detailed to enable patients to know exactly whether a given case falls within a 'volume of free healthcare' or not. Therefore, the overall volumes do not restrict the right to free healthcare from the perspective of a particular patient can put a claim to. This right is restricted only through the need to stand in queues whose length is regulated according to the aggregate volumes, limited capacity (which follows partly from financing limits), and the need of this particular patient as ascertained by a qualified professional.

In view of all this, aggregate volumes do not provide legal constraints on relations between a patient and clinician or provider. Though constantly referred to in legislative and policy documents, the volumes are not a source of legal definition for the right to free healthcare. They pertain to the relations between financing (and

[29] The semi-legalistic limitation of the query is in force here: the queues of the kinds must exist, but

procuring and controlling) authorities and the provider. Their exact role depends on the financing scheme. If a hospital is financed retrospectively, the volume of possible hospitalizations in a ward will be limited by the number of patients that can be treated over a given period of time. Given a standard length of hospitalization or nature of treatment paid for by the mandatory health insurance, the overall volume of free care will correspond to the financial constraint imposed by the financing agency (an MHI Fund or an insurance company) on the hospital.

If the hospital were financed by capitation or by annual budgeting, volumes would acquire the status of elements of planning. Volumes would influence, together with the number of staff hired and medical standards, the length of queues, possible investment decisions, and other elements of planning and management. Notice that budgetary (regional and federal) financing is required for capital investment and the level of this financing will necessarily be a further constraint on the capacity and therefore the way the aggregate volumes translate into access to care for individual patients. For example, the actual ability to perform operations will depend on availability of relevant diagnostic equipment. If the budgetary financing is low, there will be a longer queue for use of that equipment. People may be hospitalized in time, but access to the equipment and therefore operations may not be immediate. This is an example of the earlier mentioned mixture of financing sources affecting the execution of the right to free healthcare.

In conclusion, it seems that the right to free healthcare is determined by need as defined by symptoms and other relevant information; queuing weighted according to urgency, and the available financing and capacity, possibly defined through the volumes as a more or less strict target.

that does not mean that they always do.

It is easy to rationalize the lack of detailed exposition of how much a patient should expect from public healthcare. The very nature of medical care, the need for which arises unpredictably and where capacity and resources must be rationed, leads to reliance on the individual professional in deciding the eventual entitlement to free care. So, the contention of this Chapter is not that there should be more clarity in the definition of free care rights.

The problem is the conflict of interests that apparently emerges as long as the medical establishments are allowed to provide paid healthcare along with free healthcare. Inevitable imperfections of planning and standardizing will increasingly be of nuisance once such a conflict takes place.

2.3 Regulation of Paid Healthcare and Free Care Rights

Paid medical services were legalized as soon as market reforms started in 1991. From 1992 onwards, various provisions for paid medical services appeared across regions to answer the need to regulate more closely an emerging market. It is not clear how many regions came up with the early round of such regulations (Rostov and Leningrad *oblast*s are among them). A federal provision appeared in 1996, called *Rules of Delivery of Paid Medical Services* (Federal Rules 1996). According to the Rules, regulation of paid medical services is the responsibility of federal, regional, and local authorities. Henceforth, I will use the word Rules (federal or regional) to denote a by-law outlining separation between free and paid services and regulation of the latter.

A general definition of what is to be free of charge has been discussed in the previous section. Paid services are supposed to be additional to that entitlement. Such is the basic provision of the federal Rules. Methods of calculating the price of a paid service and accounting rules are suggested as well. The regional authorities have the right to provide a more detailed distinction between free and paid services. They can vary the list of free services towards enlargement, and sometimes tried to cut it. In this they met resistance from patients and the medical profession eventually supported by federal institutions.

During the second half of the 1990s, virtually all regions come up with Rules regarding paid services. These regional Rules determine the ways of measurement, more or less detailed, of what is considered to be 'in excess' of MHI program. Thereby, the Rules repeat and expand the 1996 federal provisions.[30]

Regulation of paid services consists of three components. First of all, conditions under which services can delivered for money are laid out. One specific provision within this component defines the services never to be covered by public funds. The need for such a provision could have arisen from the delegation of some details of the MHI coverage to local authorities. The second part of the regulation comprises pricing, accounting, and cash-flow. Services delivered by state providers are considered here, so all financial matters are naturally subject to tight regulation. The third component, the least developed and least explicit, concerns the ways a patient might complain about being forced to pay for a service free by law. All these components copy the pertinent federal rules and contain references to them. They enlarge federal provisions towards a more detailed regulation of the status of paid services.

I shall consider these three components in more detail now. The question I ask is not, what arguments could be voiced in a court of law based on the written provisions, but what a patient who has read the provisions should expect to encounter when applying for care. The provisions to be considered are by their nature bureaucratic rules and procedures and they remain such in virtual absence of applicable case law. Also, I shall not undertake to interpret of the provisions in light of the Constitution or the Civil Code or any other elements of the broader legal framework.[31]

2.2.1 Legal status of paid services delivered by providers of free services. The case of interest is provision of paid services by the same provider who delivers the free services in the same area, often to the same individuals. First, the service must be in excess of the guaranteed free healthcare volumes and conditions of service delivery. Second, the delivery of paid healthcare should not negatively affect delivery of free healthcare. In a sense, these two restrictions overlap in their scope. It must be clear however, that together they prohibit any decrease in delivery of free services due to provision of paid healthcare, if that decrease occurs without the concerned patient's consent.

The previous section showed the extent of the right to free healthcare, with the medical profession having a major role in rationing capacities, resources and efforts under the constraint of procedures and standards. Allowing the same professional to benefit from providing paid services to the same patient constitutes a conflict of interest, as long as the menus of free and paid services overlap.

The overall regulation of paid services is based on the following definition of when it is legal to charge a patient. A patient should be informed about what he or she is entitled to according to the free healthcare guarantees. The patient pays for everything that is in excess of volumes or in deviation from rationing rules for free

[30] In fact, there has been no total regulatory gap in a region even when there was no local provision, as the federal ones covered all the three requirement components.

[31] The Constitutional Court has confirmed that allowing paid services in no way represents violation of the right to free healthcare.

healthcare. Paid services should not prevent discharge of free services. These sums up the federal rules regarding paid medical services (Ministry of Health 1996).

This is the basic provision expanded on by regional authorities. Some regions explicitly allow all services to be paid, if patients want to pay and express their desire to do so. I shall call the respective clause in documents a ‘wish to pay’ clause. Other regions require written acknowledgement of having been informed about the right to free care. Yet other regions effectively permit all services to be paid for, provided there are any deviations from rationing rules. To be more specific, consider the regulation adopted in Yaroslavl’ *oblast*, where the bulk of interviews considered in the next Chapter were conducted. This *oblast*’s Rules are typical in the sense that all the elements constituting legal paid healthcare are mentioned, but also in that their phrasing leads to a number of uncertainties. The text closely followed here is the Yaroslavl’ *oblast* program of free care guarantees, which contains Rules for paid services as an appendix. The provisions have changed over years, with more details added. Here I summarize two documents, omitting details that are immaterial for this discussion: Yaroslavl’ *oblast*’s program of free care for 1999 (Yaroslavl’ 1999, Appendix 3), which lists cases when the provider may charge for its services. The interested Russian-reading reader may want to consult the web page of the *oblast* hospital in Yaroslavl’ *oblast* (*Yaroslavskaia Oblastnaia Klinicheskaya bolnitsa,* www.yrh.yar.ru) where the thin boundary between free and paid care is presented in a very straightforward format, repeating statutory provisions.

In general, the explicit intention of the regulator is to allow paid services that are in excess of free care entitlement. This intention is repeated in the Yaroslavl’ *oblast* rules with the following variation. Providers are to charge for all deviations from rationing rules. Here are the details. In the part concerning outpatient services, the following cases are legal: non-emergency consultations, diagnosis, and treatment outside of the queue, in excess of the "adopted volumes and conditions of delivery of [free] services". Also, as a special case, these services can be sold to the patient, if the latter does not have symptoms warranting the right to free care. This is in fact the first case, as there it was stated that paid healthcare is the one that can be delivered if the conditions of delivery of free services do not apply. Finally, there is a third case: diagnostic services (tests) which are delivered without reference from an internist, if the patient agrees to pay.

The first thing to notice is that free care volumes are mentioned here. I claim that this reference is not very consequential. As argued above, the volumes of guaranteed free care only define the rationing of capacities and resources among the elements of healthcare system and thus affect queues. They do not imply anything directly as to the obligation of providers to deliver free services. So, the only effective restriction is that the conditions for delivery of free healthcare must be absent. Among these, the most important is the absence of relevant symptoms. Jumping the queue is another case. But also, a simple lack of referral can lead to legal justification of charging for the service.

Out-patient curative measures can be paid for in a similar range of cases: jumping the queue, the absence of relevant symptoms, and the absence of referral. Finally, the in-patient case is regulated even more liberally. Lack of referral from a policlinic (alternatively: the assigned policlinic, depending on how one interprets the right to choose primary care supplier), symptoms, and jumping the queue make payment for healthcare services legal. If the individual declines the opportunity to be hospitalized for free, a hospitalization charge is legal. A patient may refuse to be treated in a standard way and require access to better equipment, which would imply payment.

Refusal to receive care with the common queue and on common grounds must be formalized in a written form.

There is virtually no difference among regions regarding all these rules and their potential effect on free service provision. Some, like Kemerovo *oblast*, are more explicit and provide more detail. In Kemerovo (Kemerovo 2000, para. 3.3), patients can be charged for hospitalization if they do not want free admission to hospital. Also, a patient can be offered a choice between a low quality and high quality hospital. The low quality hospital is free of charge, a paying patient goes to the high quality establishment. It is apparent that the quality difference will not be in terms of extra hotel services, but in terms of quality of treatment, which includes better staff, equipment, etc. Since all hospitals have limited capacity the rule effectively pushes non-paying patients into worse hospitals.

The clause allowing charging patients whenever they are wish to be charged is widespread, though whether presence or absence of the clause reflects any material differences in policies is not possible to ascertain. The stipulation of interest is found in Leningrad *oblast* (Leningrad *oblast* 1997, Appendix 4), Tula *oblast* (Tula 1998), Voronezh *oblast* (Voronezh 1996), Orel *oblast* (Orel 2000). The presence or absence of the stipulation is not likely to reflect variation in forms of control over paid services. It is not clear whether the absence of such a clause means that some services free by law cannot be delivered for money. With no case law to help at this point, I shall abstain from purely legal speculations.

I have been able to find only one case of explicit restriction on the scope of paid services. In Nizhni Novgorod *oblast*, infections and life-threatening diseases were excluded from the list of possible paid services. To summarize, there are three restrictions on delivery of paid services:

1. Paid services must be associated with some extra quality or other deviation from rationing rules, if a region does not explicitly permit all services to be on sale in the form of a ‘wish to pay’ clause.
2. The patient must be kept informed about whether a service can be delivered for free, and paperwork to this effect may be required;
3. Paid services should not cause obstruction for discharge of free services. Arrangements are normally to be maintained that would separate free and paid healthcare, as regards both distribution of workload and use of facilities.

As regional Rules sometimes put it, paid care cannot be delivered ‘instead of’ free care: crowding out of free care is not among the regulator’s intentions. The three restrictions embody this idea. But there are limits to them being material factors.

As to the second restriction, paperwork requirements probably differ across regions. Some hospitals would be satisfied with the patient making an entry in a ledger, thereby refusing to be hospitalized free of charge in the regular way, i.e. waiting in a queue or to a designated hospital. Moscow offered all patients a chance to acknowledge the fact of being informed of the status of the service (within the guaranteed free package or not) by signing a form (Moscow 1996). Written refusal is required only for hospitalization in Yaroslavl’ *oblast* (an entry in a ledger). Though no regulatory justification has been found, the practice is the same in Saint Petersburg.

Finally, the obstruction of free services through provision of paid services is clearly not allowed, but all deviations from rationing procedures are charged. At least in many cases, such deviations are necessarily associated with obstruction for free services. If a paying patient jumps the queue and gets hospitalized or operated ahead of a non-paying patient, it seems difficult to conclude that paid healthcare does not negatively affect free healthcare.

All the three restrictions appear to be inconsequential in restraining the actual ability of provider to charge. They do not restrict the scope of services for sale, only stipulate conditions for it. But these conditions require rather extensive supervision system for implementation and are subject to manipulation otherwise. The existing supervision system is considered below.

A hypothetical patient deciding on a rational strategy of using public healthcare system may come to the following conclusion. Paid care is legal, so everything on the list can be sold. This means that private payment for medical service is not anymore bribery, unlike the situation in the Soviet Union. There is nothing illegal about paying for what could by law be taken free of charge. Nor is it illegal to buy some medical service in excess of what is available free of charge. A patient shall not feel constrained to offer money.

The other side of the coin is that under the current regulation, any deviation from rationing rules means that provider has the right to impose a charge. The patient shall not be shocked if a doctor suggests that the patient pay for some extra service.

The conflict of interest that arises in connection with the right to charge the patient is only marginally tempered by regulation. First, the provider (or individual professional) is able to deny free delivery of a service, selling it or its analogue instead, and the patient will have hard time proving that that has actually happened. Secondly, the provider is the institution that defines what exactly is for free and what is not. In fact, it seems from the above consideration that the very legal definition of the right to free healthcare implies a very high role of the professional and institutional provider. A credible threat of low quality care will frequently be sufficient to extort payment.

It is difficult to see how rules for paid services could be interpreted in the course of a proper legal analysis and I shall not venture into hypotheses in this regard. In any case, given that the provider right to sell services is so unambiguous, a complaining patient would probably have to prove that a pressure to pay was exerted or free services denied where the patient was entitled to receive them. Additionally, the patient could try to demonstrate that free service delivery has been obstructed by discharge of paid services. All these things are difficult to prove. If rules of rationing of care and resources were not too ostensibly violated, only an inadvertent (and recorded) statement from a doctor forcing to pay for something free by law would constitute a legal ground for a successful appeal. Only foolishness on the part of the doctor could precipitate such an outcome, and even then, only if the usual protection from closed professional ranks fails.

Besides, it is not clear how much a regular patient would benefit from even a successful appeal. The initial incentive to complain may not be there in the first place. After all, if a patient expects better quality for money he or she does not have much motivation to complain afterwards, unless the quality of the service is too low. It is the non-paying patients who will suffer from any obstruction to free services caused by sale of paid care.

2.1.2 Pricing. Prices for paid healthcare are under control of regional authorities, which are supposed to follow the federal Ministry of Health's recommendations as to pricing. Prices include a profit margin, doctors' and nurses' wages, social taxes, and the cost of materials and equipment. Regions provide detailed (and varying) algorithms for price formation and for distribution of the proceeds.

For example, the Moscow Health Department established price rules based on multipliers applied to the MHI tariffs. For example, prices for in-patient services are 3.3 times the amount the provider receives from the MHI budget. For out-patient services, the multiplier is 2.9. There are customary concessions for pensioners and war veterans. Prices for standard out-patient check-ups and some tests are not linked to MHI tariffs (Moscow 2000).

Across regions, pricing regulations differ by the extent to which prices are uniform across clinics. In some places, such as Kemerovo price differentials between first-rate and second-rate providers exist. Price lists are published for certain services, mostly for the out-patient sector. Pricing for in-patient services is less clear. In Saint Petersburg, the official price for an in-patient service appears to be the MHI tariff.

The difference between pricing for out- and in-patient services appeared early on. Out-patient services enjoyed detailed price lists, while only hotel services and some manipulations were covered for the in-patient sector. Recently, rather detailed lists of prices for in-patient services have appeared with prices even for complex operations.

In October 2003, pricing regulation documents still suggested that policlinics use more detailed price lists than hospitals rendering in-patient services. If the in-patient services are indeed priced in a less detailed way than the out-patient ones, then we have two similar yet not identical regimes of forcing patients to pay for services free by law. The first is the case of in-patient care where patients are approached personally by a representative of the provider, potentially knowing about the patient's status and income. The patient may be forced to choose not between free and paid service, but between paid service and no service at all.

The second regime is that of published and enforced prices. Under this regime, the patient chooses between paid and free services as if in an open market. The provider's pressure is not personalized and is mediated via general quality differentiation: queues are shorter in the Paid Service department, doctors are more attentive and better equipment may be available. No one is denied free care, but patients are discouraged from applying for it.

This difference becomes important when the question of external control is raised. A patient cannot effectively complain against violation of patient's rights in the out-patient care, because she will find it difficult to prove that quality differentiation was present. In the case of in-patient care the patient cannot complain because she signed a contract with the provider, even though she was the weak party to the bargain. Though the effect is the same -- complaints are ineffective -- the underlying mechanisms are different.

As far as accounting is concerned, this is the most standard part of the paid medical service regulation. Proceeds and expenses are reported in a cash-based format. Approximately the same format is used for all commercial operations of state-owned legal entities whose chartered objective is some non-for-profit activity and which are allowed to earn extra funds through additional for profit activities.

Regarding distribution of proceeds, prices are supposed to cover cost, wages (paid in addition to the wages paid from the public funds and subject to the same social taxes), and sometimes yield a profit margin restricted to 30 or 50%, depending on the region. Managerial and professional staff can earn extra wages and bonuses if revenues are good. There are regional and sub-regional differences. For example, in Altai *krai* in 1997, personnel could receive bonuses up to 2% of gross revenue but not more than 40% of their regular salary (Altai 1997). In a survey conducted by the Institute for Social Policy in Yaroslavl'' *oblast* medical personnel reported receiving only 20% of proceeds from paid services as a net wage supplement (after social taxes, Bogatova et al. 2003, 69).

Price regulation can be seen as a check on the negative effects of the conflict of interest. Because the published prices are publicly known, patients can at least form clear expectation of the maximum they are to pay for certain items of care. But price regulation, even if it develops into an all-encompassing and consistently enforced price list, obviously does not eliminate the conflict of interest.

2.1.3 Enforcing the law. The last component of the regulatory framework to be considered is enforcement of the right to free healthcare. As has already been indicated, the distinction between free and paid services is vague enough to endow the medical profession with some power to deny free services to patients.

Two questions regard enforcement procedures regarding the right to free services:

1. Who exactly is responsible for the enforcement of the statutory border between free and paid services?
2. How effective are the procedures of supervision and enforcement?

Regional and local (municipal) Rules endorse the Health Committee of the respective administration (Governor's or Mayor's office) to control paid healthcare. Supervision of free services is vaguely shared between the regional health authority on the one hand and private insurance companies reporting to the regional MHI Funds or local branches of these Funds on the other. The idea of mandatory health insurance is that the MHI institutions carry out most supervising functions. The framework Healthcare Insurance Law 1991 (Art. 6) says that the patient addresses administrators of the care provider in case of violation of the right to healthcare.

If paid services mean denial of free services, the responsible authority is MHI institutions. Yet, if paid services are considered just as paid services, the local administration is such an authority.

As is possible to surmise on the basis of interviews with private insurance companies, the second point of view sometimes prevails as regards at least charges accompanied with regular paperwork (as of 2001, in Saint Petersburg). MHI institutions thus refuse to deal with cases of payments, unless violation of patient rights is too evident (see Appendix A.3 for information regarding relevant procedures in Saint Petersburg and Chapter Three, subsection 3.2.5 for other findings). As part of the pattern, insurance companies take the patient's wish to pay as proof of the legality of the transaction. MHI institutions do deal with cases where a free service is denied to a patient and no written consent to pay has been obtained from the patient. MHI supervisors have rejected responsibility for cases where the patient had not complained and paid for the service, and then possibly complained later.

To be sure, MHI does process complaints against charging providers, as the complaint statistics shows (see Tables 2-2 and 2-3 below). These will however be cases of charging a patient who is forced to pay and who has documented his or her right to free care in accordance with the conditions laid out above.

Neither MHI structures, nor Health Committees are obliged to act otherwise than on the basis of patients' complaints. If charging for healthcare is considered legal when a patient consents to pay in writing, there is a further barrier to investigate violations of free care rights in absence of complaints. Denial of free services and taking money from patients are not among the violations found in the course of either the quality of service or the so-called "medical economic audit" that concentrates on financial operations (see tables in Federal Fund of Mandatory Health Insurance 2000, 42; 2001,34). This underscores the complaint-based nature of the system.

As is clear by now, the patient may either complain to the head of ward or hospital, who reports to state institutions (local, regional or federal) or to MHI structures. MHI structures deal with direct complaints from patients. The difference between the two types of complaint may be material. In the first case, it is likely that the hospital will be the authority of last resort for the patient. In that case, the complaining patient invokes an internal investigation. An important factor here is that complaint is directed not to an independent entity, but to the superiors of alleged perpetrators.

The punishment system is not developed beyond regular provisions for taking ill-gotten profits away from medical providers, revoking licenses in extreme cases, and appeals to general tort and contract laws. Specific provisions regarding violation of free care entitlements are not found. Fines and administrative punishments can be imposed; personnel can be fired, for sure.

There is one component of regulation that is probably more or less enforced by local administrations: the pricing of services, notably of out-patient services. In this part, there is a check on the conflict of interest, or at least some of its consequences.

Turning now to the complaints to the MHI structures and court decision statistics, these appear to confirm the general intuition of rather limited restrictions on provider ability to charge. In general, the practice within MHI structures appears to indicate that all conflicts between a patient and a doctor or institutional provider are resolved in informal ways at the lowest possible level of the hierarchy. More along these lines is provided in Chapter Three (subsection 3.2.5) and in Appendix A.3. It may also be interesting to note that some court practice exists, but is rather insignificant as a method of resolving conflicts.[32]

The statistics on complaints filed in accordance with an MHI procedure exists, but there are none on procedures outside of MHI (through local administrations). The data are imperfect, as subsection 3.2.5 (and Appendix A.3-2) purports to demonstrate. Yet it is interesting to see which complaints tend to be officially processed. Table 2-2 summarizes the data on complaints related to payments for services free by law.

Table 2-2. Complaints related to actual or attempted charges for services free by law, when a patient applies for care within the region of residence at a provider contracted within MHI.

Year	**Complaints related to money charged, total number of cases, thou.**	**Complaints related to money charged, total number of cases satisfied, thou.**	**Complaints related to money claimed, total number of cases, thou. (total complaints satisfied)**
1997	8.3	6.5	12.3 (8.2)
1998	11.1	10.5	13.7 (6.6)
1999	13	8.7	
2000	15.5	10	21.2
2001	20	14	12

Source: MHI Yearbooks, 1997-2001

Complaints related to "asking for money in payment for medical services" in 2000 and 2001 were registered without indicating how many of them were deemed justified. This somewhat strange category may represent exactly the cases when a patient or the relatives informed an authority that doctors tried to collect money but did not succeed, possibly being thwarted by a call from an MHI fund or an insurance company.

Considering the figures in Table 2-2, one should keep in mind that the total number of paying patients in the hospital sector only can be estimated at about 2.5-2.8 million a year (see Table 3-2 in Chapter Three). So, even 20 thousand complaints are

[32] The other extreme would be to widen the avenues for suing doctors and institutional providers, which could put on hold paid healthcare at all, but could arguably lead to elimination of the 'added quality' of service currently 'bought' for private money. The lax regulation can be considered as a solution to the hold up problem in provider-patient relations.

less than one percent of the total number of paying patients. Fourteen thousand satisfied complaints is about 0.5 percent. So, the MHI system appears not to be exerting a significant pressure on providers in this regard.

All in all, the complaints related to allegedly unlawful charges are a fraction of total complaints processed by the system, and they are satisfied in a lesser proportion than, for example complaints regarding choice of provider. Table 2-3 summarizes the data.

Table 2-3. Three major reasons to complain within MHI and complaints against charges (filed complaints).[33]

Year	**Complaints related to issuance of MHI policies, % total (% total satisfied)**	**Complaints related to choice of provider, % total (% total satisfied)**	**Complaints related to medication provision, % total (% total satisfied**	**Complaints related to charging patient (within the region of residence, by a provider contracted within MHI), % total (% total satisfied)**
1999	46.6 (45.9)	15.4 (22.9)	17.5 (18.1)	2.7 (2.8)
2000	56.7 (55)	11.9 (19)	12 (15.1)	3.1 (3.0)
2001	61.6 (69)	10 (13)	9.2 (9.6)	3.1 (3.5)

Source: MHI Yearbooks, 1999-2001

Court practice is summarized in Table 2-4. Throughout MHI's existence, the total number of cases filed and considered has steadily increased. I exclude from consideration reimbursement claims related to medical care expenses due to unavailability of free care. These claims are filed by MHI funds of insurance companies and are not related to the provider-patient interaction.

[33] Complaints, regarding quality of care, choice of doctor, availability of medication, are all below 10% in the total number of complaints. At the same time, it may be interesting to note that complaints about the choice of provider are mostly considered justified (above 90% in all years, see Federal Fund of Mandatory Health Insurance 1999, 47; 2000, 38; 2001, 31).

Table 2-4. Court practice statistics.

Year	Claims filed	Claims considered / satisfied	Awards paid out (million rubles)	Pay-outs from pre-court settlements (million rubles)
1997	244 (150 by patients)	91/64	0.26	3.5
1998	343 (149 by patients)	156/105	0.5	4.9
1999	785 (577 by patients)	390/267 (199 by patients)	1.7	14.9
2000	834 (691 by patients)	391/252	3.06	-
2001	789 (619 by patients)	429/259 (206 by patients)	3.45	18.9

Source: MHI Yearbooks, 1997-2001

A majority of all court decisions is related to malpractice.[34] There are some cases of illegal charges for services free of charge by law. There is not enough information as to the substance of such cases to draw conclusions. According to the available information, in case of a successful claim, a patient is reimbursed for expenses. For a successful complaint, the payment must have been made to a cashier and acknowledged with a receipt and the patient must have a proof of being eligible for free care (Perm Regional Human Rights Center 2001, 129-130). The above consideration of the rather general conditions when charging a patient becomes legal therefore applies.

It is rather difficult to tell from these figures whether there are many or few complaints and settlements. This is true both of malpractice suits and even more so of cases concerning payments for services free by law.

Brennan 1992 (831 and 848) quotes the following figures for the U.S (more precisely, New York). About 3.7% of hospitalizations result in iatrogenic injury, 1% being related to malpractice. Half of iatrogenic deaths are due to substandard care. At the same time, out of total 2267 claims filed, 783 are related to cases of iatrogenic injury and of these 625 are related to cases of such an injury caused by substandard care. The total number of hospitalizations was approximately 2.5 million and the total number of iatrogenic injuries around 71.4 thousand. Americans use courts more often than Russians, but not on every occasion, possibly using internal complaint systems. Besides, there is little correlation between the fact of malpractice and filing a suit for malpractice. This means that substance of complaints must be looked at before drawing conclusions from such statistics. Similar reservations must be held in regard to complaints and settlements on the issues of free service.[35]

I have chosen to look at the regulation of paid care assuming that providers have large powers over the patient and that a partially informal internal system of complaints dominates. I favor the institutional and disregard legal aspects. The fact that only a few patients file suits and only a tiny minority uses the formal channel of the internal complaint system justifies my choice.

I have considered the nature of relevant regulation and sector-specific hurdles that a complaining patient faces. Both suggest that the role of patient right defense either through internal complaints or through courts is to react to the most outrageous cases, when the patient is ready to complain against a doctor and obtains the necessary support from supervising institutions.

2.4 Combined Paid and Free Services as a Public Policy

Legalization of paid services is officially argued for as a way to improve the quality of healthcare in the environment of budgetary constraints on free healthcare. Or, put simply, the officially proclaimed policy at all level is to enhance the volumes of paid healthcare in order to provide a supplementary financing for healthcare. Ministry of Health (Ministry of Health 2001) has recently issued a document encouraging every effort in this direction. The providers who made every effort to ‘earn money’, were

[34] Tort (Consumer Rights Protection Law 1992) and contract law apply.

praised for rational use of the resources under their control. At the regional level, authorities responsible for development of healthcare are expected to increase volumes of paid services.

Among many documents in which the intention to make medical provider 'earn money' is expressed there are some with extra policy measures, often quite original. For example, in Volgograd *oblast*, in 1999, a fund for financing extra purchases of medication was created. One of the statutory sources of finances was supposed to be "voluntary donations from the medical organizations providing paid services". (Volgograd 1999). As the notion of 'voluntary' is applied to a tightly regulated industry, one can assume that a certain mechanism of cross subsidization was created, to channel the money earned by the providers into the system of public financing of healthcare.

It can be argued that the goal of larger volumes of paid care adds to the conflict of interests described above. The providers are given a further incentive to extort money, while their supervisors are pressed for less aggressive policy. There is of course no logical inconsistency between the desire to increase the volume of paid care and the desire to defend the right to free healthcare. The latter can be a constraint on the former. Yet, if the law is violated anyway and enforcement is lax, the constraint becomes 'soft' and paid care is increased at the expense of free care.

There is a countervailing movement to compensate for the providers' tendency to violate the right to free healthcare. These have not amounted so far to a significant overhaul of legislation, though this is sometimes proposed. Here is a short history of such efforts.

[35] Procedures of complaint processing must also be looked at when analysing the controversial evidence regarding the practice suing allegedly negligent doctors (Dauer and Marcus 1997).

Before 2001, the only document issued by a state authority which contained recognition of the problem was a letter by the Attorney General's Office (Attorney General 1996). The letter described the results of an investigation into the practice of illegal charges for free services. The percentage of patients getting free care at one of Krasnodar *krai* hospitals was put at 12%. The remaining 88% were paying. Payments were made for children and adults alike. No follow up on this investigation could be found.

From 2001 on, concerns were being voiced regarding providers of medical services making patients pay for what is supposed to be free of charge. Some politically important discussions happened at the regional level. A number of regional legislative bodies, namely those of Orel *oblast* and Khabarovsk and Stavropol *krais*, and possibly others launched investigations into alleged violations in the delivery of paid services. In Khabarovsk *krai*, the regional legislative body forced the Governor to launch an investigation into denial of free care. The legislators suggested that there a list of services, which can never be charged for. They suggested reconsidering the pricing for medical services, while a letter sent to the Constitutional Court asked whether the very delivery of paid services alongside with free ones is compatible with the right to free healthcare.[36] Not only was the existence of a "black market" of medical services acknowledged, but its negative impact on the delivery of free services was noticed (Khabarovsk 2001). Interestingly, accounting violations, though these must accompany a black market, were not mentioned. The discussion apparently concentrated exclusively on the issue of patients' rights to free healthcare.

In the light of the above discussed policy of making provider "earn money", weakness of the efforts to contain corruption in public healthcare can be demonstrated

[36] As has already been said, the Constitutional Court has confirmed legality of paid care.

as follows. In the same Khabarovsk region a typical contract with the head of a medical service unit contains a clause of interest, according to which the head of a medical unit must make every effort to increase volumes of paid healthcare (Khabarovsk 2001). This reflects the official notion of paid services as a supplement to publicly funded services. Though there is no logical inconsistency between the official role of paid services in improving the finances of healthcare and the right to free healthcare, in reality they may stand in conflict with each other.[37]

Other regions also expressed their concerns about the 'black market' of medical services, as well as accounting irregularities. Again, the notion of a 'black market' was invoked. Consequences of the hearings and decisions could not be traced.

In conclusion, it seems that at least at the regional level and only very recently some progress has been made towards a minimal recognition of the problem. Nevertheless, the effective balance remains skewed towards encouraging paid care without tempering the conflict of interest.

2.5 Conclusions

2.5.1 The regulatory gap. Focusing now on the big picture of all components of medical service regulation brought together, I suggest the following conclusion about the legal status of payments for services within the entitlement for free healthcare. Providers do not have the right to deny services within the volumes of the free service entitlement if rules of rationing are followed. But volumes do not imply an effective constraint. As to the rationing rules, if a patient expressed a wish to pay for a service, charges become fully legal. The only remaining restrictions is that patients be

[37] Certainly, an argument can be invented to support the opposite point of view. Creating a pressure on healthcare establishments' management to increase the (officially recorded) proceeds, one provides an incentive towards greater accountability of the provider at least in terms of reporting the earnings. Transparency can increase as a result, as can the control over individual professionals.

informed of their rights and discharge of free services not be obstructed by delivery of paid care.

Because it is providers who define the exact boundary of entitlement for free service, a conflict of interest emerges. This enhances the traditional power of medical service providers over their patrons.

I have also considered pricing and accounting rules. Virtually all paid services are delivered using the publicly provided and owned facilities and goods. These are paid for from the proceeds from the paid services. Following the accounting rules, providers incur extra expenses such as income and social taxes and are restricted in distribution of the proceeds. The pricing regulation can also be assumed as further constraining the eventual profits of the provider. Pricing restrictions may also mitigate some of the consequences of the conflict of interest.

Enforcement occurs through an internal complaint system, either within the medical establishment or directly to a state authority, or else to MHI structures. Statistics about complaints appears to support the idea that patients are reluctant to use any of these channels. The role of courts is very limited.

Finally, health policies include incentives for development of paid services as complementary financing source in public healthcare. Attempts to balance these with restrictions on provider ability to charge are reduced to general declarations and sporadic investigations. So far, these attempts are inconsequential.

The upshot of the discussion is as follows. It is legal to deliver a paid service; it is illegal to force payments by denying free services. This framework provision is clear and indisputable. However, its immediate consequence is an unmitigated conflict of interest, where the provider has enormous power to elicit payments.

I shall call this situation a regulatory gap. This does not to be read as a negative evaluation of the provisions. Absence of regulation or control may or may not be socially useful. Legalizing paid care has perhaps increased transparency of transactions that would have taken place anyway. Now a patient can buy healthcare whenever he or she is dissatisfied with the free version.

2.5.2 Balance of incentives for providers and medical professionals. Public healthcare is supposed to be rationed by need or luck. At the same time, another arrangement allows the very same professionals and providers to benefit from skewing the distribution of effort and resources in favor of the paying patients. Though this skewing is theoretically prohibited, it is permitted in practice through lax enforcement and very broad conditions under which transactions can be considered legal.

Suppose that an (individual or corporate) provider is capable of forcing a patient to pay for certain services. Such a provider has two options: either a legal contract is signed and the payment is delivered 'over the counter', or the whole transactions proceeds in an 'informal', 'straight into the pocket' way.

On the one hand providers have an incentive to make the patient confirm willingness to pay in writing, as this provides a legal 'smoke screen' for an illegal action. On the other hand, applying accounting regulations leads to redistribution of profits from the individual to the corporate provider, or from the corporate provider to the state. Also, application of pricing regulation restricts the provider's freedom to define prices and, in particular to engage in price discrimination. These two latter considerations may force the provider to turn to informal ways.

A third ('middle') solution could be optimal for the provider: obtain a written confirmation of refusal of free care or provide an obvious deviation from rationing rules, but flout accounting regulations by failing to report the fact or details of the transaction.

One further element must be added to this balance of incentives under the current regulation: If legal paid healthcare is used as vehicle for successful violation of patient rights that would anyway be violated, double payment for a service is avoided.[38] Without the legal form, both government and patients pay for the same services. With it, only patients do. Of course, this can happen only if paid services are suitably recorded, regulations are followed and, as a consequence, taxes paid. Another condition necessary for the effect to take place is that MHI reimbursement for the services should be fee for service or per bed day or illness case payment (see Chapter One, section 1.2 for details on the current reimbursement arrangements). Probably, insurance companies indeed impose fines on those providers who attempt to finance their services both from public and private sources and record transactions (Perm Regional Human Rights Center 2001, 129).

2.5.3 An appraisal of the regulatory effort. As part of general deregulation of the economy, public healthcare in Russia can now provide paid services. This means conflict of interest for providers. The conflict of interest is far from being tempered by a vague and incomplete normative and institutional framework, which vests significant power in medical profession and low-level state institutions.

Assuming that patients had to pay for services free by law before the advent of market economy, the rules and procedures governing delivery of paid services in public healthcare have an interesting political and economic status. They legalized what had been considered illegal before without changing the nature of previously

[38] The supervisor of this thesis, Iván Csaba takes credit for this point.

existing practice. That per se is the market reform. But a side effect was that a legal front was provided for practices that are still illegal. Finally, they excluded an important part of health policy (the patients' rights defense in the part related to paid services) from the responsibility of MHI, making it prey to regulatory capture and unelaborated legislation.

Overall, the effect of legalization of the paid services has both positive and negative aspects. Legalization of paid services over the 1990s has filled a legal void and supplied the state with adequate instruments of control. Yet, the provider's power has amplified. It has become safer to force payments for the reason that selling the services had become legal unlike the situation in the Soviet Union. At the same time, there has been no political will to defend patients' rights in this respect, either at the stage of law creation or at the stage of law enforcement.

Political and economic interests may be reflected in the controversial regulations, though one certainly must not draw sweeping conclusions at this point. It appears that the following balance of facts currently prevails. All regulatory gaps that have been discussed here could in general be just inevitable imperfections of human law applied to a complex area. The inadequacy of rules and procedures regarding patient right defense stems from the nature of healthcare and that of the transition economy with its own priorities. The problem comes not from an evil intention, but the very complexity of the issue has naturally, though deplorably, led to the problem.

At the same time, the regulatory gap ostensibly serves certain interests of both state institutions and professions. Hence the possibility that the form and content of regulation reflects the unwillingness of state institutions to defend the right to free healthcare. Chapter Four discusses these hypotheses, while Chapter Five provides some formal arguments as to rationality of such policy. Chapter Three examines whether and how payments for services free by law result from interactions between patients and medical professionals.

Chapter Three. Empirical Data on Payments for Services Free by Law: In-depth interviews and statistics.

3.1 Introduction

This Chapter is an attempt to create, from the sparse evidence available, a hypothetical picture of a certain very peculiar institutional set-up that underlies the payments for healthcare services free by law. A quantitative assessment of household expenditures on healthcare, interpreted in light of qualitative evidence from semi-structured interviews, enables a general characterization of such payments.

The main question is how patients and healthcare professionals (more rarely institutions) interact during these transactions. In Introduction (Chapter One) this was called the "micro-level" analysis. I endeavor to show when and how patients pay and how medical professionals structure their work activity in response to rewards from private pockets.

Quantitative evidence is considered in order to give a broad 'cash-flow' assessment of the role of private money in the public healthcare. Quantitative evidence does not single out exactly the payments made for services delivered 'in excess' of the state guarantees of free care. The reason is that the numbers do not render visible the structure of interaction between the provider and the patient. Qualitative evidence sheds light on the nature of the transactions and allows identification of components corresponding to payments for services free by law.

I will look closely at both motivations behind the transactions of interest and the ensuing distribution of costs and benefits. I eventually outline a precarious balance created by many factors, incentives and motivations, which explain the apparent stability of the institution. The Chapter is organized as follows. First, the available statistical evidence is presented and discussed (Second 3.1). Secondly, interviews with the representatives of the medical profession, insurance companies and local administration are described (Section 3.2). The analytical part of the Chapter (Section 3.3) assesses the likely motivations behind the system of payments for services that must be delivered free of charge at the point of need. Tentative conceptualizations are suggested, put to further speculative use in Chapter 4.

The main focus will be on the qualitative data, namely in-depth interviews conducted in two regions in 2000 and 2002. The Independent Institute for Social Research, Moscow conducted most of the interviews with the medical professionals in Yaroslavl' *oblast* in 2002. The author conducted three additional interviews with two doctors and a nurse from Saint Petersburg (see Appendix A.3-1). The Yaroslavl' *oblast* interviews are the main source of observations and insights. The interviews conducted in Saint Petersburg will supply some additional information as to the variety of scenarios of illegal payments. The author also conducted some interviews at insurance companies in Saint Petersburg in 2000 (see Appendix A.3-2). These interviews are secondary in importance and the detailed discussion of them is therefore relegated to an appendix, but some particularly interesting insights from them are quoted in this Chapter.

Yaroslavl' *oblast* was selected for the bulk of the interviews because of the regional administration's support necessary to conduct such sensitive research. Yet to dispel the legitimate concerns, I demonstrate that the *oblast* is in many ways a typical Russian region in table 3-1.

Table 3-1. Characteristics of Yaroslavl' *oblast*

Variable	**Yaroslavl' oblast, 2001**	**Russia, 2001**
Public healthcare spending per capita, yearly, rubles	1340.7	1606.8
Doctors per 10,000	54.5	47.3
Nurses per 10,000	102	108
Hospital beds per 10,000	124	115
Policlinic capacity, visits per day, per 10,000	253	248
Hospital admissions per capita, bed-days	3.37	3.33
Out-patient visits per capita	6.56	8.65
Emergency calls per capita	0.35	0.34

Sources: Federal Fund of Mandatory Health Insurance 2002; *Goskomstat* 2003a (table 9.1); *Goskomstat* 2003b (124).

3.2 Statistical Evidence

This section presents and discusses quantitative data on payments made by patients (or relatives on their behalf) for medical services and medication in public healthcare establishments.

3.2.1 Description of the approach used. Two quantitative measures reflect payments in public healthcare. One is the frequency, that is to say, the percentage of patients or households paying for care, in total and by various categories. The other is the amount of payments. Both measures reflect the burden on the patient and complement each other.

Alternatively, one could measure the income of doctors and nurses accruing from private payments, both in absolute and relative terms. Surveys in Poland and Hungary take this second approach (Kornai 2000), showing that payments by patients constitute a substantial portion of personal incomes of medical professionals. In the Russian case, researchers used only the first approach.

The best way of assessing the amount and frequency of private payments would be to identify a group of households whose members were hospitalized or underwent an out-patient treatment and then record payments through a regular household survey. Complementing this information with hospital or policlinic records (diagnosis, course of treatment, doctor or nurse assigned) would give an extremely reliable and complete picture of who pays how much for what.

Unfortunately, such a survey has never been conducted. Only general household expenditure surveys are available. These will be described immediately, with emphasis on frequency and amount of payments with breakdown into several categories. Because diagnosis and course of treatment could not be controlled, the usefulness of both these measures is rather limited.

The main technical limitation is however the fact that most of the patients do not pay, and of those who do, very few pay substantial amounts. Household survey formats are a bad approach to track such rare events, especially given that patients constitute a minority of respondents. For assessment of the amounts, this is a problem of reliability for the most basic estimates (averages, totals). To estimate the frequency of payments, this problem at least means that the sub-samples to be used in building an explanatory model are too small for statistical analysis in absence of strong patterns of correlation. For example, if there were 10 thousand individuals in the total sample, only five percent, or 500 had been hospitalized over the quarter preceding the survey. Even if 20% of these paid for treatment, one is left with one hundred individuals whose paying behavior is supposed to be explained. No significant correlation of explanatory use could be found in these data. This may be either because there are none in reality or because the sample is too small. A bigger sample could reveal some significant links among measured variables.

For these reasons, I have to limit the ambit of the quantitative research to the basics: estimation of means and variances of size and frequency of payments, divided into formal and informal payments, whenever such a difference was recorded.

3.2.2 The data. The idea that free care in Russia is increasingly a fiction has been around for a while. Public awareness of the problem should not however be mistaken for the proof that the problem itself exists: measurement is in order. Quite a number of attempts have been made to assess to what extent people pay for healthcare. All these attempts taken together give at least the range of values for the size and frequency of private payments in public healthcare.

Three sources can be considered as comprehensive: The University of Boston survey (UB, see Appendix A.1 for details), The Russia Longitudinal Monitoring Study (RLMS, see Appendix A.2 for details), and data from the Russian State Statistical Bureau (*Goskomstat*). Table 3-2 gives a comparative summary of features of these three sources.

Table 3-2. Comparison of three surveys.

Parameter	**RLMS**	***Goskomstat***	**UB**
Division formal (to a cashier)/informal ('into the pocket')	Partially	No	Yes
Subcategories	Partially	No	Yes
Identification of health-related expenditures	Yes	No	No
Exclusion of private clinics	Yes	No	Yes
Measurement of frequency	Percent of patients	N/a	Percent of households
Periodicity	Yearly	Yearly	1997 and 1998

Notes:

1.RLMS did not specify in their questionnaires whether expenditures were made formally or informally until Round IX (year 2000).

2. RLMS divided payments into those for out-patient treatment, tests, in-patient treatment and check-ups (for a driver's license, for example). Only Rounds IX and X contained specific question about expenditures on medication as part of payment for in-patient treatment, both formally and informally.

RLMS appears to be the most appropriate of all three for general assessment of private expenditures. Though it is less detailed and less specialized than UB, it covers more years and, possibly most importantly, it was conducted over several fall months, while UB was conducted in December, a month of particularly high demand for health services and therefore for paid services as well. Tables 3-2, 3, 4 summarize the available statistics.

The RLMS data quoted here do not contain expenditures (formal or informal) on medication at hospitals (see Appendix A.2). There are two reasons for this. First, such a separation of medication was not performed before the year 2000 and the previous rounds appear to measure only expenditures on services. Excluding medication therefore gives a more consistent picture.

Secondly, the RLMS survey is used here as a conservative estimate of payments for services free by law. Expenditure on medication unavailable free of charge outside of interaction with medical professionals is not of interest. However, what patients believe to be spending on medication may in fact be payment for services (personal remuneration of the professionals), as the following section suggests. The Boston University surveys could serve as the upper limit, as they include expenditure on medication.

Table 3-2. Frequency of payments, %

Survey/variable	1994	1995	1996	1997	1998	1999	2000	2001
1. Out-patient payments, RLMS	4	5	5		7		10	12
2. Informal outpatient payments, RLMS							5.2	6.5
3. Laboratory, additional treatment payments, RLMS	9	7	8		17		17	25
4. Informal laboratory and additional treatment, RLMS							8.3	7.5
5. In-patient payments, RLMS	12	14	22		44		12	17
6. Informal in-patient payments, RLMS							6.8	8.33
7. In-patient payments, UB (average for two years)				34				
8. Informal in-patient payments, UB (average for two years)				14				
9. Out-patient payments, UB				13				
10. Informal out-patient payments, UB (average for two years)				3				

Table 3-3. Amounts of payments, billion 2001 rubles.

Survey/variable	1994	1995	1996	1997	1998	1999	2000	2001
1. Total yearly expenditures, RLMS	52.0	28.2	26.2		39.6		24.7	21.0
2. Total in-patient expenditures in the last quater, UB				19.1	18.3			
3. Of which informal				5.2	5.3			
4. Total out-patient expenditures in December, UB				5.5	10.9			
5. Of which informal				2.1	2.1			
6. Total yearly private expenditures, *Goskomstat*	19.9	27.3	22.8	31.1	34.4	32.2	33.0	37.9
7. Total regional healthcare budget, less payments to MHI for non-working population			173.4	189.4	149.8	131.2	133.2	140
8. Income of medical service providers from MHI budget	103.8	109.3	73.8	76.9	78.1	65.1	71.8	80
9. Total expenditures on wages (before tax) from MHI budget	41.5	49.0	38.3	38.8	37.8		32.4	40.3

Table 3.4 Amounts of payments, % GDP

Variable/Survey	1994	1995	1996	1997	1998	1999	2000	2001
1. Total expenditures, RLMS	0.28	0.2	0.3		0.5		0.3	0.23
2. Total private in-patient expenditures, UB				0.22	0.22			
3. Total private out-patient expenditures, UB				0.06	0.13			
4. Total private expenditures, Goskomstat	0.10	0.19	0.26	0.37	0.41	0.41	0.38	0.42
5. Total regional healthcare budget without payments to MHI for non-working			2.0	2.3	1.8	1.7	1.5	1.5
6. Income of providers from MHI	0.5	0.75	0.84	0.93	0.9	0.8	0.8	0.88
7. MHI expenditure on wages (before tax)	0.2	0.3	0.4	0.46	0.46		0.37	0.45

Sources: RLMS 1994-2001; Goskomstat 2003a; Boikov et al. 1998, 2000a, b, Federal Fund of Mandatory Health Insurance 1995-2002.

Notes:

1. Shishkin (2000, 142-143) uses these seasonal figures to obtain yearly estimates. He takes into account the fact that utilization in the fall and winter months is above average. The high figures obtained (27 billion for 1997 and 32 billion for 1998) may reflect the fact that paid care concentrates in fall and winter months to a higher extent that service utilization does.
2. Payments for all services related to medicine, in all establishments, including private.
3. Social taxes account for about 1/3 of the total payroll. Person income tax is 12% (13% in 2001).
4. For RLMS and University of Boston surveys: the author's estimations.

The relative weight of private payments may be better seen from the proportion private payments have in the national healthcare expenditures. Consider, for example, the ratios of variably measured private expenditures to public spending on healthcare payroll.

Table 3-5. Some proportions

Ratio	1994	1995	1996	1997	1998	1999	2000	2001
1. Total private expenditures (RLMS) to total MHI expenditures on wages	1.4	0.65	0.8		1.0		0.8	0.5
2. Total private expenditures (official statistics) to total MHI expenditures on wages	0.5	0.63	0.65	0.8	0.89		1.03	1.08

Source: Tables 3.3; 3.4

Differences in survey design and the already mentioned limited reliability of the estimates prevent a deeper comparative analysis of the sources of data. Yet a clear common pattern emerges. Patients (or their families) are paying for healthcare both formally (to a cashier) and informally ('into the pocket'). Formal and informal payments co-exist, with the former being dominant by amount (see Table 3-3 lines 2,4,6 and Table 3-3, lines 3,5). By any measure, only a minority pays; the in-patient care exhibits the higher frequency of payment.

As the University of Boston surveys show, patients are most willing to pay for a course of treatment and medication (see Appendix A.1, tables A.1-3 and A.1-4). Payments for laboratory tests and personal tips to doctors and nurses are much smaller. The literature (Sheiman 1998, Boikov et al. 2000a, b) concludes both from this and the dominance of formal payments that the function of private money is to buy service rather than provide a voluntary reward (a 'tip') to a doctor. At least, it seems fair to say that patients think that they are paying for services and medication.

The amounts paid privately are consistently below those paid from public sources (Tables 3, 4), but approximate official wages (Table 5). The private payments must be a non-trivial incentive, assuming that the personnel benefit from them. If however a payment goes through all official channels and is properly recorded, its incentive effect on medical professionals is to be diluted hugely (see subsection 2.5.2 of Chapter Two). One alternative scenario is that what seems 'over the table' to a patient, is not exactly a 'formal', properly recorded transaction. The service-rendering doctor or nurse is then the major beneficiary of the payment and the incentive effect remains.

Looking at the tables for RLMS surveys in Appendix A.2, one will conclude that private payments cannot be considered as very large when compared to the personal incomes of the payers. In year 2001 the average formal in-patient payment was 2047 rubles, or USD 70. The maximal payment was 20,000 rubles, or USD 700: see table A.2-2, line 2001-18.

It is probable that the surveys exclude minor payments for minor services of nurses and orderlies. They also exclude in-kind rewards. A car-mechanic may ‘pay’ a doctor by fixing his car, and this transaction will be not reflected in the surveys.

The INDEM Foundation asserts, on the basis of a statistical survey of petty corruption among 2017 respondents, in 1999-2001, that the total size of informal charges in the public healthcare is 600 million US dollars (about 18 billion rubles) yearly. Among all petty corruption expenditures, bribes in policlinics and for hospital admission rank second and third, right after bribes for university admission (Satarov 2002).

It is hard to interpret this figure, because it comes from a survey about informal payments, which are deemed to be in the same category as common bribes. But the figures being within the range provided by the RLSM and Boston University surveys, the INDEM research seems to confirm the above estimates in the sense that private money constitutes an important source of financing of the public healthcare sector, although lower in amounts than public funding. Frequencies are reported to be 22.2 percent for policlinics, 22.8 for hospital admission, 27.3 for in-patient treatment (Satarov 2002, table 3). Ahead of the discussion of quantitative evidence, it is worthwhile noting that such frequencies per se do not support Satarov’s conclusion that charge for healthcare are extremely pervasive.

Among less reliable results, the following can be mentioned. Details on these surveys were not available, so I quote them with caution. This is one reason to suspend judgment on them. Another is that many local survey report quote very high figures for frequency of payments (see below), which is a reason to think of a design bias. In the city of Taganrog in 1998, 22% of patients in policlinics paid, while only 8% received completely free treatment in in-patient care. 28% said that the payments were made informally (Bogatova et al. 2002, 19). In the city of Kemerovo, in 1998, a similar survey gave the following results: 51% of in-patients paid for treatment, 45% for diagnostics; 18% of patients in policlinics paid for treatment by general practitioners and 38% in specialized out-patient wards (Morozova and Kulibakin 1998). But then these are local surveys and extrapolation does not seem warranted. Besides, 1998 was a year of economic crisis, which makes any extrapolation even more doubtful. There were several other surveys (see Bogatova et al. 2002, 19-20) with a small sample size, which means that their results are rather unreliable.

In the Yaroslavl' interviews that are discussed in the next section respondents gave the following assessments. In different clinics different share of, for example, operations are made 'for' money: from 20 to 70 percent. At the same time, non-invasive treatment is paid for in 10% of cases. This appears in line with other data: a sizable minority pays. The dependence on the type of clinic, though non-quantifiable at this point, also stands to reason.

3.2.3 General comments on the data.

As has already been said, the purpose of the statistical enquiry is limited to total amounts of private payments and their frequencies. Three items seem to be firmly on the list of statistical facts about payments for healthcare:

1. only a minority pays;
2. total amounts are comparable with state expenditures on wages, with the latter possibly growing faster than the former over the last few years (RLMS);
3. formal payments are the more significant in terms of amounts in both the RLMS and Boston University surveys.

The less reliable conclusions are as follows (see Appendix A.1 for details):

1. frequency of payment is not strongly associated with ability to pay (income group), as witnessed by tables A.1-5 and A.1-6 in Appendix A.1;
2. payments are made mostly for a course of treatment or medication. Payments for separate services appear to be rare (see Tables A.1-3 and A.1-4 in Appendix A.1);
3. frequency of payments for children is extremely small as compared to that for pensioners and working age patients (see Tables A.1-3 and A.1-4 in Appendix A.1).

The literature (see Introduction, subsection 1.3.3) asserts that in the Russian case, such payments mean that private money is largely coming in to replace the insufficient public financing. One extreme scenario of this replacement is the refusal to serve patients unless they pay. The apparently predominant role of formal payments has been widely discussed in the literature, though this fact's actual significance is still unclear.

It appears that the data demonstrate co-existence of free and paid care, rather than disruption of free services and a catastrophic transition to paid care. Free and paid services co-exist everywhere and it appears very difficult to trace any reliable correlation with location or type of settlement or even income group. Co-existence of the two modes of financing of healthcare may still mean that the phenomenon is characteristic of certain settlements or locations, for statistics are clearly insufficient to draw strong conclusions. Alternatively, it may mean that separation into paying and non-paying patients happens at the level of an individual provider, ward, or professional. Obviously, it is not possible to trace such separation in the patient pool.

If paid and free care co-exist while only a minority pays, without a clearly visible pattern of division into paying and non-paying patients, the hypothesis of widespread denial of free care and the advent of paid medicine in Russia is not confirmed. Certainly, even the strongest version of the hypothesis is not refuted completely. It may still be the case that the paying minority receives quality care, while the rest is accorded only a semblance of it.

In terms of amounts, the bulk of spending on healthcare comes from public funds. Private money may however be a vital source of financing, while being minor in terms of amounts. The reason is that even the lowest estimates (either from official statistics or the RLMS surveys) suggest that private payments should play an important role in rewarding the medical personnel. Comparing official wages and private payments shows that private and public sources of personal income of professionals are nearly equal.

I propose the following conclusion. The population appears to pay for services free by law and the more traditional vision of such payment as informal, 'straight into pocket' forms may not be even prevalent, at least in terms of amounts. At the same time, neither frequencies of payments, nor the overall amounts support (or disprove) the vision of free care in Russia as completely or nearly completely replaced by paid care. Free and paid services are both present, with private payments hypothetically playing an important incentive role and thereby tilting the distribution of effort and resources in favor of paying patients. Similarly, formal and informal payments co-exist. Their co-existence reflects general and local constraints and incentives to institutional providers, professionals and patients (see Chapter Two).

This rather abstract and vague notion of co-existence of free and paid care and formal and informal payments is the first step towards understanding of the phenomenon of interest. Co-existence of paid and free care is an alternative ‘metaphor’ to replacement of free care with paid care happening once a certain financing threshold is crossed.

One can hardly expect a more detailed picture regarding the size of the phenomenon of payments for services and goods free by law unless very specialized research is undertaken. Rather than asking how much Russia spends on healthcare privately, I propose to ask why a patient would pay any money at all. In contrast to the previous research, I will shift the focus from the size to the nature of the phenomenon. Understanding the circumstances under which a patient may want to pay will reveal more details of the co-existence of free and paid care.

3.3 Interview-based Evidence

This section draws on the evidence from the interviews that were conducted in the Yaroslavl’ *oblast* in February-April, 2002. 35 interviews were made with representatives of the external controlling and financing institutions (city administration, MHI fund, insurance companies) and 135 with medical professionals, part of whom had administrative responsibilities. Three interviews conducted by the author in Saint Petersburg (see Appendix A.3 for details) are used as a supplementary source of information.

3.3.1 Methods and overview. This section describes payments for medication and healthcare services, which, according to the respondents themselves, must be delivered for free. Questionnaires with open-ended questions were used both in the Yaroslavl’ and Saint Petersburg interviews. Besides, in a short questionnaire given to the staff of insurance companies in Saint Petersburg, the respondent had to make a choice from 3 or 4 alternatives. Appendix A.3 contains details on the interviews that I conducted in Saint Petersburg. Momentarily, I will describe the method of extracting information from the Yaroslavl' interviews in detail.

Each Yaroslavl' *oblast* interview was structured around a general plan, consisting of several topics and subtopics on which respondents were to comment. Interviewers would ask additional questions, adopting the structure of the interview to elicit maximum information. Respondents were encouraged to provide their own opinion and even speculation on functioning of healthcare. The following topics were covered (see Bogatova et al. 2002, 194-196 for a detailed account):

1. development of paid services in light of recent changes in the law;
2. informal payments for medication;
3. informal payments for services.

This survey in Yaroslavl' *oblast* itself was not part of this research. This means that the structure of the interviews does not reflect many of my own research preferences. To give an example, in my opinion, the Yaroslavl' survey does not provide enough information on the actual step-by-step procedures of interaction between patients and medical professionals. Instead, in my view, it puts too much stress on respondent opinion exactly where such opinion is suspect of bias. Moral justification of payments is an example.

Extracting usable information from such semi-structured interviews is a problem. It is worthwhile reciting benefits and shortcomings of qualitative data collection. Among its many benefits, one is able to monitor such important things as the degree of flexibility in arranging a payment as well as the ways a doctor would justify the practice. Peculiar methods of quality differentiation and forcing to pay could be discovered, which may not be possible through quantitative statistical analysis. There is a price to pay for this important information. The latter is not quantifiable, and is often mixed with personal impressions. People in our society are trained moralists and often cannot distinguished between fact and value. This means that a fair degree of skepticism is in order.

The literature on qualitative data collection contains a number of techniques, which I borrow and tailor to the needs of my research. Particularly, I used Carney 1972 and Lindlof 1995. I primarily extract reliable information by comparing different sources. First of all I identify opinions that are shared by a maximal number of respondents. Then I delete from consideration those, which bear obvious marks of self-justification, whether I would grant the points to the respondent or not. Subsequent comparison with quantitative data allows selection of the most reliable components of the emerging picture.

The strategy of presenting information from the Yaroslavl' and Saint Petersburg interviews is as follows:

1. a broad issue is identified; for example: scenarios of payments classified according to the pressure exerted on the patient. This largely repeats the structure of questionnaires;
2. one or more most typical (consensual or near-consensual) scenarios are presented;
3. explanations for the typical scenarios are quoted;
4. analytical responses are quoted, as well as interesting particular cases that appear to express more general patterns or shed more light on the typical scenarios;
5. a picture that consistently incorporates the identified elements is constructed.

With regard to steps 2 and 4, the idea is to single out those scenarios that dominate, but also quote the most important general observations that doctors made and which are consistent with the overall picture. I consciously refrain from frequency analysis, even when it is possible. Even if, for example, 70% of health professionals and administrators clearly claim that extortion of payment rarely happens, it is difficult to judge how rarely extortion happens in reality. These 70% may either report in bad faith or may not have the opportunity to charge more aggressively. There are indeed reasons to believe that inequality among professionals may favor certain minorities, as will be discussed shortly.

The structure of the section is as follows. First, the problem of formal versus informal payments is discussed. Secondly, the evidence on payments for medication in Yaroslavl' *oblast* is given. Thirdly, in the major part of the section, the evidence on payments to doctors and nurses for their services is reported and discussed. In the last fourth part of the section institutional policies are discussed.

A terminological remark is in order. The discussion of payments in healthcare often makes a distinction between payments for medication and payments for services. Doctors and administrators were asked to separate these two components of paid care. I report the two as different, both above in the quantitative evidence and below when describing the semi-structured interviews. However, there is no compelling evidence to the effect that such a distinction is tenable. Payments for medication may conceal payments to doctors or management. This means that the incentive roles of the two types of payments are mixed. The degree of this mixture is impossible to identify on the basis of the interviews. Blurring categories of payments here is in line with the dubious legal status of all such payments, as has been emphasized in Chapter Two.

3.3.2 The role of formal (as per contract) payments. The first topic on which Yaroslavl' interviews demonstrate a total consensus is the payments for services free by law made to the cashier and properly accounted for. For lack of a better term, such payments will be called formal. Respondents do not consider these payments for services as significant, either by amount or as a part of their additional personal income. This is a consensus view. Formal payments for medication are however more significant and do not generate such a consensus, and these will be accounted for in the next subsection.

Respondents suggested two reasons why patients do not go the official way. First of all, the medical professionals are not very much interested in formal payments, as they enjoy only ten or twenty percent of the gross proceeds as wage supplements. This conforms to the published rules for distribution of proceeds in paid care. An average doctor will also not control the distribution of proceeds that constitute the profit (after accrual of wage supplements to the staff). The other reason is that patients are not likely to use these services as prices are too high for the given quality standards and personal incomes. They prefer direct transactions with a doctor or nurse.

Both reasons are almost trivial. People avoid exposing their transactions in order not to pay taxes, in healthcare as anywhere else, if punishments for such a practice are insignificant. The degree of protection that could have been afforded to both patient and doctor by contractual relations is evidently not enough to work as a counter-balance to tax evasion incentive. Paid care as a separate component of public healthcare does not emerge, because there are other, less expensive channels for doctors and patients to reach agreements.

The insignificance of formal payments creates only one problem: it is ostensibly inconsistent with other data, presented above. There are however important differences between the two sets of data. First of all, the quantitative data were from a great number of regions. Secondly, and more importantly, comparison of two sets of data can be hampered by the possibility of the following scenario. It may be the case that there is a certain degree of formality involved, for example, a contract with a patient is signed. Proper accounting does not follow, and taxes are not paid (and price regulation and other controls do not apply, either). The patient may believe that the payment was made under enforceable legal contract, and report it as such (see Appendix A.3-2 on contracts as a smoke screen for extortion). One interviewed deputy head doctor of a central city hospital acknowledged a practice that may amount to such a scheme. All paying patients were made to acknowledge in writing what they were paying for: the choice of doctor, a particular manipulation, type of anesthesia, etc. (Bogatova et al. 2002, 133).[39]

Another explanation refers to the concentration of payments in certain parts of public healthcare, as will be discussed below. If a few doctors in a few major clinics controlled significant formal payments, the majority of the interviewed will not know about such practices as they themselves observe only informal payments.

Both accounting aspects and regional or time differences may together explain the discrepancy between the two sets of data. It does not seem possible at this point to track the borderline between formal and informal payments any further than this general suggestion. The conclusions of this chapter will remain largely invariant to changes in the formal status of payments.

[39] This is consistent with the respective observation about the rules for paid services (Chapter Two). The paperwork required to legalize paid care that is delivered instead of free care is minimal (entry in a hospital ledger, signature on a receipt copy containing the necessary stipulation of voluntary refusal of free care).

The bulk of cases reported in the Yaroslav *oblast* interviews are formal payments for medication and informal payments for services. These two categories are discussed below. The Saint Petersburg interviews with medical professionals also concerned informal payments only. The roles of formal payments were touched upon in the interviews at insurance companies.

3.3.3 Payments for medication.

Quantitative evidence shows that payments for medication are second in size to payments for services (see Appendix A.1, tables A.1-3, 4 lines 6). Doctors are most ready to admit that patients have to pay for medication (see also Appendix A.3, interview with an *Ask-Med* representative). In the Yaroslavl' interviews, doctors were asked to assess the proportion of patients paying for medication. All respondents put the figure in the region 70-80%, with exception of one central district hospital, where it was 10-20% and several doctors who claimed that in several wards the percentage was as high as 90.

A closer look at the particular scenarios respondents provided shows that these figures can be interpreted in different ways. No consensual picture emerges. If some scenarios get mentioned more frequently than others do they are also the more suspicious in presenting medical professionals as not benefiting personally from selling medication.

One more or less typical explanation why patients pay for medication that must be in hospitals for public money is lack of public funding. One head of a ward comments:

> All our patients are sharply divided into categories. Some are well-off and can pay, but they are from the administration [*meaning that there are privileged ones who do not pay, because they have*

> *connections – M.R.*]. The second category are those who cannot buy even medication. [Bogatova et al. 2002, 76][40]

Those having connections in a city administration or among medical professionals receive privileged treatment and may even be cross-financed by the poorer fellow-patients. One respondent even reported that doctors and nurses buy medication for the well-connected patients, though probably this is not typical. Several others claimed that some patients buy medication and other materials (syringes, etc.) in excess of their individual need and doctors pass on the excess to other patients, especially emergence patients who cannot pay for themselves. Patients pay for medical services in-kind, with medical supplies, but it is other, non-paying patients who benefit from such 'payments', not doctors.

A concrete example from an interview shows how this can happen. There are, for example, two groups of patients. Planned hospitalizations comprise one group. Emergency patients comprise the other group. The former group acquires medication or other supplies in excess of its need. The difference between the acquired amounts and the immediate need of 'regular' patients is used for emergency cases. Patients my also be differentiated by income group or social status. Interestingly, this differentiation does not always mean that the rich subsidize the poor. In fact, it can be the other way round.

Several interviewed however reported schemes through which doctors personally benefited from selling medication. This information belongs to the category of specific scenarios, but there is no reason to believe that the under-supply of medication for public money is the main factor and these scenarios are secondary in importance. There are three versions of them. Doctors directly sell medication to

[40] This and subsequent quotations from interviews are translated by me.

patients, though it is available for free: they effectively sell what does not belong to them. Alternatively, doctors sell (or prescribe) expensive substitutes for equally effective free generics. A doctor can have an interest in a pharmacy associated with the hospital or receive a commission from a pharmaceutical company for promotion of their products. The third scheme is that doctors simply collect money from patients on account that their medication must be paid for.

Interviews in Saint Petersburg confirm that doctors may engage in profitable selling of medication to patients. The named patterns are followed, except that cooperation with pharmaceuticals was not mentioned. Cheating a patient into buying something rather expensive and useless was emphasized (see Appendix A.3-1, Topic 3).

One hospital in Yaroslavl' *oblast* developed an ingenious scheme which apparently solved some of the legal problems arising in connection with selling medication to patients. The gist of the scheme was that patients paid a certain amount of money to a private insurance company and received a policy from that company which covered their need for medication during their stay in that particular hospital. The insurance company then transferred the money to the hospital, which bought the medication.

The reasons for patients to be happy with such a scheme were that they would always receive whatever they needed and that the substances were bought on wholesale market, which made price lower for the patients. In fact, however, as the respondent who provided the information about the scheme indicated, the price of medication to the patients was eventually higher than the market retail price. Also, the scheme did not always work. It often took too long to order a drug, which made the scheme less popular with patients. It was decided at some point not to run the scheme anymore. It is unknown who exactly initiated termination of the practice and whether any sanctions followed.[41]

According to the comments provided in the Saint Petersburg interviews, payments for medication may in fact play the role of payments for service: a patient thinks that he or she is buying medication while in fact a service is being bought. For example, a reputed doctor agrees to take a patient for an operation, on the condition that the patient pays for medication. In this case, and consistent with the above schemes of reselling publicly funded medication to the patient, the patient pays not for medication but for the operation, without necessarily understanding this.

3.3.4 Payments to doctors and nurses (payments for services).

Topic 1. Pattern of concentration. What are patients paying for? The list of mentioned 'items' that patients buy seems to comprise everything that can possibly be paid for (see Bogatova et al. 2002, 80). Some pay for change of bed linen, others, at the other end of the range, pay for a whole course of treatment. The opportunity to receive an operation, jump the queue, and other aspects of apparent quality differentiation are reported, both for out-patient and in-patient care. Both in the Yaroslavl' and Saint Petersburg interviews, the choice of doctor is mentioned as one of the most important items of differentiation.

One doctor interviewed in Saint Petersburg claimed that the choice can be presented to a patient as follows: "Either I will operate on you or my student will" (see Appendix A.3-1). This appears to be a concise formula applicable across many different situations, even if it is rarely used in such a blunt version. The patient buys privileged access to professional effort.

[41] In the Russian law such schemes are called "pseudo-insurance" and are explicitly prohibited

Choice of hospital is another item in the category. Admission to a central or large or reputed hospital is payable. In the Yaroslavl' set of interviews, the *oblastnoi* (region-level, as opposed to local) hospital is such an object of desire. Quite expectedly, access to better equipment is often quoted as an important type of quality differentiation, additional to the choice of doctor or hospital and apparently correlated with those.

Nurses are paid for timely injections, work in shifts for the particular patients, etc. (See table A.3-1 in Appendix A.3 for detailed account of some items payables in these schemes). In general, it is ”attention” that is bought from nurses, since attention easily translates into valuable extra quality of service. Effective contracts with clearly defined conditions of payments and service delivery may emerge between patients or their relatives and nurses, especially for post-operational care, where nurse effort is especially valuable.

This notion of everything for sale must, however, be balanced with some more facts. Only a minority of patients, 20-40%, in rare wards up to 70% pay for treatment involving surgery. If there is no surgery involved, especially for out-patient services, payments are rare, below 10% of all patients. In other words, though each possible item of care was at least once reported as possibly paid for, when asked to assess the frequency of payments, doctors agreed that payments are concentrated in certain wards, clinics and around certain reputed personalities. These figures appear in harmony with the frequency estimates considered above.

This suggests that notwithstanding an avalanche of various scenarios of quality differentiation, in the overall picture such scenarios concentrate around substantial and/or continuous medical involvement or services demanded by young and middle

(Ministry of Finance 1999).

aged people: surgery, notably neural surgery, venereal diseases, child delivery, chronic illnesses. These areas will be called concentration areas. This concentration has three dimensions:

1. Frequency: patients more often pay in some wards or situations than in other;
2. Size, by amounts of money paid;
3. Nature of payment: it appears that in concentration areas payments are more tightly correlated with quality differentiation between paying and non-paying patients, up to the point where free service is not available at all.

This pattern of concentration is the first and most important conclusion to be derived from the interview-based statistics. Various doctors contributed three mutually compatible explanations of the pattern. The first is that patients are willing to pay for particular types of services and not others. Namely, patients are ready to pay for surgery but not for non-invasive treatment. As one respondent comments, an operation is:

> a one-shot, immediate improvement of health. […] Cardiologist […] gives you pills to swallow, but you have no idea when you are going to get better.” (Bogatova, et al. 2002, 90]

The urgency is heightened by the possibility of death resulting from operation by a bad doctor (Bogatova et al., 2002, 84). The best of all for the patient is to know the doctor personally and use personal connections in conjunction with money, as follows from the use of the archaic-sounding expression “my man” (see Chapter One, section 1.3 for the Hungarian case).

Urology and treatment of sexually transmitted diseases are the areas that are, in the language of the respondents, totally “commercialized” as per custom. Child delivery, pre-natal, and post-natal care are also frequently paid for. Again, the ability to produce observable changes in a patient’s state of health is instrumental in convincing patients to pay for any of these fields. Also patients are more willing to pay if they are in constant contact with doctors and nurses, in other words, if they are chronic patients.

Respondents also give two sociological reasons. Respondents report that doctors working in policlinics, district hospitals and non-invasive wards tend to be females and are often secondary sources of income in their families. This is understood as explaining lower level of pressure on patients to pay from these doctors.

Also patient sociology is quoted as a factor. Brain surgery tends to beperformed on a richer pool of patients than non-invasive cardiology. Emergency cases do not pay because they simply are not in a position to, in contrast to chronic patients.

These explanations sit well with the quantitative picture. Together, they suggest a hypothetical notion of why patients are motivated to pay. People pay *in extremis*, for substantial medical treatments; but not if a classic emergency happens and there is no time for a transaction. People also pay when they are in constant contact with medical personnel.

This picture of concentrated payments partially interprets the observed frequency of payments in quantitative data and low dependence on personal income. If examined across institutional providers, payments are concentrated according to nature of service required and quality expected, rather than ability to pay. Further exploration is required for more detailed picture of how payments are distributed within those wards where they are concentrated.

All this being said, I do not imply that patients receive quality care free of charge in a pulmonary ward in a district hospital and quality care only for money in a urology ward in a central hospital. It may very well be the case that where patients are unwilling to pay, there is little to pay for. Or, alternatively, patients are too greedy to part with their money unless in emergency, so they do not receive quality care free of charge anywhere.

The case of policlinics was not discussed as separate in the Yaroslavl' interviews. There is extremely anecdotal evidence from the Saint Petersburg interviews, both with insurance companies (notably *Vesta*) and medical workers saying that there may be a system of extorting payment with transactions being formalized. The system is based on the legal separation of Paid Service Department (*khozraschetnyi otdel*). By law, service delivery at a Paid Service Department should not affect that of the rest of the institution (policlinic). However, the following strategies of the management and doctors were reported (and partly observed by the author). First of all, paying patients are served ahead of the general queue, which delays service to the non-paying patients. Also, doctors become officially hired part time within the free service, while serving paying patients the rest of the time. Though this latter strategy is not prohibited by law, it is on any account a way to make patients pay for something free by law.

The third strategy is less conspicuous. The non-paying patients are offered very dense scheduling of a test. For example, an ECG room is open for the non-paying patients only from nine to ten am, and all those who cannot fit within this hour are to wait until next day at least. A paying patient is coming at 10:10, or later, and is allowed to take as much time as is necessary. Normally, the room would not be very busy, as there are not so many paying patients, and an occasional non-paying patient can slip in and be served even after ten am. Quality of the service, measured here in terms of attention, seems to be skewed in favor of the paying patient, as well. All paying patients pay to a cashier and obtain appropriate receipts. This is a special scenario and it does not violate the notion of concentration of important payments in certain hospital wards. But its existence shows that concentration does not necessarily follow from policies or legal status of private payments, only from peculiar effective demand and possibly some sociological factors.

Topic 2. What does the patient eventually receive for money?

This topic does not enjoy any consensual view. I suggest to look more closely at both the pattern of concentration, identified above and also at a number of specific scenarios that help interpret the pattern of concentration.

Under the pattern of concentration, payments in general do not go for particular services on a piece-by-piece basis as much as for whole treatment or access to a facility or a doctor. This was also a result from quantitative studies (see Appendix A.1). But here a specific object of payments suggests itself. Payments are concentrated in areas where quality of professional effort is important and at the same time is hard for a patient to control directly. Surgery is the prime and the most important example of such a situation. So, private payments happen against the background of asymmetric information, whose general consequences in this context will be seen in Chapters Four and Five. Consider the following case, reported by one doctor. A patient paid for a coronary operation in Moscow formally, and yet also informally. The fashion in which he did that is important:

> There is the anesthetist passing by, I gave him 150 dollars, then the surgeon with the same intention. Paid him. Then also to that orderly who always forgets to change towels. (Bogatova et al. 2002, 116)

There was no negotiation, nor even indication that the payment was supposed to induce any extra service, and yet the payment was made in expectation of some extra service.

At this point, it is important to indicate yet another possible answer to the question of what exactly the doctor sells to the patient.

> Patients are usually thankful not for that the operation has been well done, but for attention. The doctor who held hand and listened to the patient would receive more gratitude that a professional who is just technically apt in operating. (Bogatova et al. 2002, 85)

Clearly, this is just the opinion of one person. But in conjunction with some claims in the literature, it seems indicative of certain important features of the transactions in question. The patient reacts to signals that are not necessarily those relevant to that patient's health. Whether this suggests some mistaken beliefs and behavior that follows from them, cannot be decided on the basis of so scant evidence. But whatever explains the patient's motivations, the doctor seems to play some marketing strategy, thus overstepping the basic role of a professional asking or even extorting money for his or her services.

Speaking about misperceptions, the following case warrants attention.

> One patient, who saw the ward's head doctor, decided that she was supposed to give him 1000 rubles *[about USD 35 – M.R.]*. So she did. He decided that this was because he was the head doctor and never visited her again. She thought, it was for the room.[42] Now she expects him to run errands for her. (Bogatova et al. 2002, 130)

[42] Patients sometimes pay for not being placed in the corridor, even when there are places in the rooms.

Though anecdotal, the small story may be representative if not of the way patients normally engage in informal transactions, then at least of the extent to which these transactions can be non-transparent even to their immediate purported beneficiaries.

Naturally, patients must keep alert to the possibility of simple cheating. Some doctors and nurses approach patients with claims for money without the desire or ability to provide these patients with extra quality of service. For example, they 'collect' contributions for services delivered by others, or just collect money without bothering to provide any good reason. Perhaps they simply approach patients who look gullible.

The very opportunity to collect money without providing any reasons for the patients to comply with the request appears to be in tune with the general situation of uncertainty for patients who cannot fully control what they receive for the payment.

So far, there are three elements defining the specifics of the situation of a patient under the pattern of concentration. These emerge in a number of particular scenarios:

1. apparent readiness of a patient to part with his or her money without pre-negotiating the object of payment;
2. lack of reliable communication channels between patient and doctor that would underpin quality assurance;
3. ability of some less conscientious doctors to collect money without providing anything in exchange.

These features are related to the pattern of concentration, as they are all reported in conjunction with circumstances of the latter. They also connect to the general notion of informational asymmetry affecting the position of a patient having to pay for healthcare.

Before proceeding to other insights, one important notion must be discussed. The ostensibly concentrated fashion of informal payments in Yaroslavl' *oblast* does not mean that payments for services outside of the main junctures of informal transactions are non-existent. Doctors say that these are mostly gifts of appreciation. And yet, gift-like payments in district clinics and policlinics may induce the same types of quality differentiation as the more obvious incentive-creating payments in surgery.

Doctors both from small and large hospitals and all wards tend to use the word 'gratitude' to refer to payments and patient motivation behind them. One intended meaning of this is that patients voluntarily pay, often after the service has been discharged and without prior negotiation. In this sense, use of the word 'gratitude' may reflect an important element of flexibility in arrangement of payments. But as far as patient motivation is concerned, no conclusion must be drawn from doctors reporting that someone 'expressed gratitude'.

There is a linguistic idiosyncrasy to the notion of gratitude. Gratitude as motivation to pay is mentioned very often, but more often than not it is not clear whether the word is used in its original meaning. In fact, in the current Russian, the verb *otblagodarit'*, meaning 'to express gratitude by words or by gift' can also mean, in way of a euphemism, paying for something in an informal way. Each time procedure of payment, conditioning of service on payment and other circumstances must be considered to decide whether one talks about gratitude or common bribery hidden behind elliptic language.

The pattern of concentration and lack of strong coupling between payment and service are the two features that appear to fit together very well. At this point the preliminary answer to the question what patients are paying for appears to be as follows:

1. Patients pay for complex treatments, where the doctor's personal skills (surgery) and general quality status of the facility (reputation, equipment, etc) matter;
2. They pay when personal relations with the doctor matter;
3. They pay when they are expected to pay where payment is generally expected (child delivery, venereal diseases).

These reasons to pay overlap. A patient having to undergo an operation will like to maintain good relations with the doctor, even if no definite return to surgery room is planned. This is what I have called the *pattern of concentration*. This pattern of concentration suggests a specific answer to the question why patients pay: they build relationships with a doctor. It seems appropriate to call the strategy as *ingratiation with money*. I shall explain the concept in more detail in section 3.4. Ingratiation with money appears to be common to many particular scenarios from interviews. More importantly, the element of ingratiation is natural for a situation of paying a doctor informally for a vaguely defined complex treatment. The nature of such relationships will to be explained in section 3.4.

Topic 3. Pricing, payment scheduling and pressure on the patient. What is the doctor strategy in view of such patient choice? Let us define flexibility as readiness on the part of a medical professional to treat patients differentially either by not pressing for a specific payment or differentiating price according to financial capability of the patient. All particular scenarios that were reported in both Yaroslavl' and Saint Petersburg sets of interviews involve various forms of such flexibility.

Respondents in Yaroslavl' *oblast* claim that patients are not directly asked to pay, except for some particular cases, when there are effective tariffs that must be paid for admission to a hospital or ward. These payments may even be formal. Patients are rarely presented with direct quality differentiation, doctors say. The size of payment is either decided by the patient, or subject to negotiation. Doctors will usually take into account the financial situation of a particular patient. Scheduling of payment is such that patients often pay after receiving service, and then they may cheat on the doctor if there was any promise of payment beforehand.

This near-consensual picture must be balanced against muted admissions of certain doctors not being flexible, or even, in parlance of some respondents, 'extorting' money from patients. If payments are indeed concentrated in certain wards, then it is seems to be most plausible that extortion happens exactly there. In any case, the following pattern may betray self-justifying confabulations. A respondent says that he or she personally never extorts money, but in another ward a doctor will not pay slightest attention to a patient if the latter does not pay:

> They who are in child delivery and surgery have certain norms *[to the effect of aggressive extortion – M.R.]*, and they do not change them, even for the benefit of a friend or colleague. (Bogatova et al. 2002, 118)

Another particular scenario that tempers the notion of a powerful patient is tariffs imposed at possibly most important junctures: access to the best facilities and best doctors. Here flexibility is compromised through certain corporate agreements among professionals themselves. Tariffs may be established at the level of hospital. These can be for admission to the hospital, availability of places in a particular room, or availability of certain operations.[43] Tariffs can also be used in the case of services of a

[43] Representatives of one insurance company in Saint Petersburg claimed that Hospital #2, one of the elite hospitals in the city, effectively imposed the following admission charge. They made the patient pay for a magnetic resonance imaging which would not necessarily be needed on medical grounds. The payment would be 1000 rubles.

particularly high ranking and highly reputed doctor.[44] Different doctors in the same institution may sell their services at different prices. (Bogatova et al. 2002, 108). Colleagues would not mind their fellow professionals quoting higher or lower price than themselves, as prices tend to reflect reputation and access to equipment.[45]

Flexibility in district hospitals and non-invasive wards can result from the general unwillingness of patients to pay. Gifts, often symbolic and thereby often offensive (a block of cigarettes, a bottle of vodka) infuriate more than reward some of the respondents. As a doctor from a child clinic acknowledges:

> Nobody offers us informal payments…We pay attention to everyone equally…Sometimes they give chocolate bars. But only sometimes. Very rarely. (Bogatova et al. 2002, 96)

In opposition to this situation, the following is a case of claimed professional benevolence (an obstetrician explains):

> I always do quality work, whether [the patient] has promised to pay or not. Whether he *[a patient – M.R.]* gives a thousand rubles or nothing at all. I cannot do worse [for a non-paying patient than for a paying one – M.R.] (Bogatova et al. 2002, 117)

Sometimes quality differentiation between paying and non-paying patients makes payment into an incentive for good work, as another obstetrician explains:

> In our field, nobody gives money in advance. We have to … earn this money…And when you have earned your money, then you can go and say: there has been a talk about a reward, I have earned it. (Bogatova et al. 2002, 133)

In this situation, a patient can easily cheat on the doctor and not pay the promised reward. This situation can also be seen as a case of flexibility, if rather adversely affecting the professional.

[44] This alleged correlation between reputation, rank, and ability to impose tariffs is only a hypothesis. Additional factors, such as the nature and status of the particular medical establishment must also be considered.

[45] Notably in surgery.

Among particular scenarios, whose frequency is impossible to assess, patients pay voluntarily without specifying services they would like to buy. Such voluntary payment may occur before or after treatment. Hustling money to the doctors has already been mentioned in this category. Paying a nurse is another example. In the interview with a nurse in Saint Petersburg, the respondent claimed that nurses mostly receive voluntary compensations, which are however given in expectation of better work (Appendix A.3-1). This does not exclude pre-negotiated (usually with relatives) extra attention and services, often delivered instead of regular duties.

Lavishly sponsoring the hospital after a successful operation is a common practice among executive management of large enterprises. Obviously, in such cases the issue of explicit pressure to pay becomes irrelevant.

In the Saint Petersburg interviews, an additional aspect of 'making the patient pay' emerged. The actions of a doctor are not confined to either directly or indirectly asking to pay or waiting for the money. Doctors may also select patients to be pressured for a payment. Those selected are likely to be weak bargainers with the ability to pay. Though apparently reasonable, this strategy was not reported in the Yaroslavl' interviews, so I consider it as marginal in the overall picture (see Appendix A.3-1 for details).

What can a critical reader of these scenarios make of the picture of rather significant flexibility of arrangements, punctured with stories of extortion? There are three alternative versions of flexibility:

1. flexibility is highest where there is less effective demand for services and lowest in the concentration areas identified above;

2. flexibility arises from professional benevolence and professional responsibility, where money are taken either only from those who can afford paying or for some significant extra effort;
3. flexibility is due to institutional constraints on professional behavior.

The first explanation seems to be confirmed by professionals from parts of the system not favored by paying patients: pediatric wards, non-invasive treatment, policlinics. This explanation is likely because it represents the cases where respondents claiming flexibility can easily be believed. It also sits well with the pattern of concentration.

The second explanation is more speculative and its consequences are discussed in Section 3.4 and in the following Chapter. Though respondents mention professional benevolence rather often, such explanations contain too much of self-justification to take at face value. At the same time, empirical studies on healthcare, referred to in the Introduction (section 1.3), suggest professional benevolence as a material factor in determining professional behavior.

The third explanation is unlikely for two reasons. One is that it does not correspond to anything in the interviews. The second is that, as the following subsections will try to demonstrate, institutional constraints are immaterial in general. This leaves the pattern of concentration and professional benevolence as two rival approaches to flexibility.

There are still many gaps in the description. To mention only what seems to be the most important outstanding issue, interviews do not render the interaction between patient and professionals visible. In particular, it is not clear whether any actual bargaining happens, at least for the more important cases, such as complex operations.

Topic 4. Differentiation among doctors: professional profile, rank, and personality. As mentioned above, surgeons earn more than other doctors and venereal diseases, child delivery and chronic illnesses are also most likely to belong to paid care. Patients also pay for state-of-art equipment. Such equipment is likely to be found only in large hospitals and be controlled by senior surgeons, which reinforces the mentioned pattern.

It is not clear whether the size or 'centrality' of a hospital figures in patient willingness to pay. As long as the size is associated with the professional status or reputation of doctors and nurses as well as the range of choice,[46] it could be reasonable for a patient to pay for admission to a large hospital. How much of this correlation is there in reality is not clear.

Some scenarios of differentiation within the profession reinforce the picture of concentration around certain professional profiles and reputation, while others show that other factors may matter. One consensual scenario from the first group, repeated in several interviews, is that senior doctors with a good reputation most easily overcome all barriers to making their duties a profitable business. "Stars are permitted a lot," -- comments one head of ward, meaning that those who have unique skills and technologies at their disposal are not constrained by management. This is because such professionals are in high demand, and hospitals cannot afford to part with them. Moreover, they are less constrained by the initiative of patients to pay as such doctors more often condition their unique services on payment.

A less frequently mentioned particular scenario is that junior doctors, the most business minded and least scrupulous about traditional ethical constraints, engage in

[46] I am indebted to Balázs Varadi for this suggestion.

extortionate practices. They overcome the hurdles such as lack of reputation and status within the profession if they "behave aggressively with patients." (Bogatova et al. 2002, 100).

Doctor personality matters in breaking the mould of traditional poverty among practitioners in out-patient clinics (policlinics). One internist who apparently broke with that tradition comments:

> My official salary […] is one thousand eight hundred rubles with all bonuses. In reality, I make up to ten thousand a month. Informal wage. A well-marketed[47] smart internist can win bread exactly by [charging] these informal payments.[…] when a patient sees that this doctor does his job well, that patient will bring his family, friends. […] I charge money for my qualification. (Bogatova et al. 2002, 97)

Aggressiveness together with making effort pays off in terms of both immediate profits and a reputation to ensure future income. Personality and attitude matter, providing a counter-balance to the concentration picture, which is based on the power of senior doctors, concentration of the best facilities in few hands, and patient willingness to pay for particular services, as well as the sociology of the medical profession in the country.

The personality factor of differentiation among doctors should not be underestimated, even though it is the most hidden. The notion of 'juniors' making medicine a proper business enterprise sits well with the older doctors' grumbles about their unacceptably low social status, caused by the low official pay.

The large powers concentrated in hands of 'star' doctors may not necessarily mean harm to the patient. In fact, being less constrained in charging patients, for example, in establishing their own tariffs they are equally less constrained in

[47] A jargon expression was used to this effect (*raskruchennyi*), which may show the willingness of a doctor to speak what is perceived as the language of business.

differentiating among patients. In particular, they can take more money from the rich patients, which may motivate them to operate free of charge the poorer ones.

Occasionally, the patient may benefit even from rigidity of tariffs and from what is called "extortion" on the part of high-ranking doctors. Rigidity may secure a uniformly good performance by the doctor. If the patient quotes the price after the operation or has a chance to bargain, the doctor, who is not so sure about the gain, may be unwilling to be at his professional best. This is a hold-up problem: the impossibility of enforcing a contract makes both sides to the contract worse-off, because they do not believe each other and do not make good on their promises. Differentiation based on seniority and administrative power can therefore be effectively supported by patient demand for services that are rendered by senior doctors in the best establishments.

3.2.5 Institutional components. With regard to institutional involvement, no consensual picture emerged. One finds instead several particular scenarios that putatively point in the direction of general irrelevance of official institutions and their aloofness with regard to doubtful practices of doctors and nurses. It also appears that effective permission to accept and even demand money is a result of a number of factors working to constraint choices of both administrators and doctors. That such an effective permission exists is clear from the fact that payments exist and nobody tries to conceal the fact. Why the permission exists is a question that remains largely unanswered. It is worthwhile nevertheless to consider the respondents' perceptions of the situation.

Topic 1. Healthcare institution's internal arrangements. Respondents provided very little information as to how hospital or policlinic administrators, senior doctors and peer professionals shape professional-patient interactions and there is no consensual picture. Before describing a few particular scenarios, I shall explain how administrative and professional hierarchies co-exist in provider institutions.

The head of ward is the most senior practitioner and also administrator at the ward level. The head of ward hires if there is a vacancy, dismisses, and executes primary control over distribution of resources. Above the ward level, professional and administrative hierarchies become different in the sense of clearer separation of tasks and personnel executing administrative tasks not necessarily being practitioners. But both the head doctor and his deputies, even if they do not perform operations or analyze lab results, are still part of the profession as former (or occasional) practitioners. Administrative and professional elements of control over doctors and nurses are impractical to differentiate in most cases, even if job descriptions for administrative and professional personnel are different.

To what extent do head doctors become part of business or political relations, external to hospital affairs? This question is yet unanswered. If they do, they may be less dependent on their medical colleagues. The head doctor will also be the mediator between hospital and local administration, having to resolve conflicts and maintain balanced relations. Further analysis of these interactions could be an interesting research in double loyalty and hard choices between professional allegiances and political exigencies.

The managerial hierarchy is involved in the processes leading to informal as well as formal payments. One innocent way of being involved is to procure extra resources from patients towards maintenance of the ward or hospital. Informal payments by patients in this case are intended not for enrichment of individual professionals, but either for medication and thus cross-subsidization of other patients directly or for acquisition of extra bed linen or other small items.

Participating in illicit profits and resolving conflict situations are less innocent examples of managerial involvement. Respondents identified a number of scenarios here. Bogatova et al. (2002) generalize various descriptions into three patterns of relationships between managers and medical professionals. Sometimes, the management is not involved in solving conflicts and does not participate in profits. In other cases, management occasionally engages in resolving conflicts and receives a share of profit. Finally, the management may be involved in the process of charging patients on a permanent basis, receiving their share of profit.

As far as professional relations are concerned, doctors and nurses may or may not have cooperative agreements. That such agreements exist was also mentioned in the Saint Petersburg interviews. Namely, nurses deliver information about ‘terms of trade’ to patients. A nurse may advertise the services of a particular doctor among the patients. The opposite pattern received attention in the interviews in Yaroslavl’ *oblast*. Doctors may actually act to ruin the reputation of nurses. For example, a doctor asks a patient to take care of the timing of injection as if he does not trust the nurse. This fosters ‘special relations’ with this particular patient and facilitates enforcement of payments.

Do the institutional hierarchies, managerial and professional, play a role in constraining profit-seeking medical professionals? It seems that they do, but in a limited way. As far as the Yaroslavl’ interviews show, reaction to scandals is the way to punish excesses. Both hierarchies discipline professionals who are imprudent or unlucky in their extortionate behavior. Among the latter, there are those delivering low quality services for money.

The main impression is however that loud patients protests are required for formal institutions to get involved, moreover to act on behalf of the patient at the expense of relations within the professional team. As far as the professional hierarchy is concerned, it would be natural to expect that senior doctors restrict the attempt of the junior ones to charge and overcharge patients. However, their power to do so is limited. One encounters contrasting attitudes. The head surgeon of a large hospital says: "If a junior [doctor] tries to operate and gets money for that, I will finish him […]" (Bogatova et al. 2002, 103). Another senior doctor is less decisive:

> Yes, the youth are more active. Because they are not sure that they will receive money according to the result of treatment. So, they prefer to extort the money in advance. (Bogatova et al. 2002, 103)

The second respondent further said that nothing could be done. Senior professionals get involved at times to restrain junior colleagues, but much depends on the personality and character of a particular person in charge.

One head nurse (Bogatova et al. 2002, 138) believed that informal payments exerted a demoralizing effect on nurses. The latter become less attentive to the needs of non-paying patients and start demanding money for everything. It is not clear whether such an effect is an observation or expectation. In any case, little can be done. There are reasons for administrative superiors as well as senior professionals to look the other way except for rare cases of scandal. Some reasons to avoid applying strict measures were mentioned in connection with the special privileges that the best doctors are endowed with. The scarce financing and traditional under-supply of nurses makes this category also quite privileged. One ward head admits that she must allow nurses to take money from patients, as otherwise the entire staff would have been lost (Bogatova et al. 2002, 138).

So, enforcing the law is undesirable: there would have probably be less medical care in the country without illegal payments. It seems however that preserving the system of healthcare is not the only factor. The medical team as a whole will not be likely to support interventions in the current practices of charging patients. Their response to such intervention is disloyalty. As part of showing such disloyalty, they would reciprocate with denying the management privileged services. Such disloyalty can have many forms. Some comments suggest sinister consequences for those trying to work to rule:

> This [tough measures to eliminate the payments] will lead to backlash from those who are interested in its [*the system's of payments*] preservation. (Bogatova et al., 2002, 159)

It would be natural to suggest that variation in institutional involvement in arranging payments increases according to the pattern of concentration, but nothing in the interviews seems to lead to such a conclusion. There probably exists some balance between involvement and non-involvement scenarios, but how the balance is struck and what factors determine the eventual arrangement in a hospital remains unknown.

Topic 2 External controllers. According to the law, external controllers are local administration and Mandatory Health Insurance institutions, such as insurance companies and local funds. The Yaroslavl' interviews included meetings with administrators and representatives of insurance companies. It seems that the range of opinion among them mainly reduces to the following items. First of all, they complain about lack of legal empowerment in dealing with the payments, as well as acknowledgement of the technical impossibility to eliminate the payments without substantially damaging the system of healthcare delivery. Secondly, the respondents tend to share the justifications of the payments given by the doctors: ”Informal payments provide incentives [for the medical professionals] to work” (Bogatova et al. 2002, 156). Thirdly, it is acknowledged that interests involved in maintaining the status quo are strong enough to resist any initiative to destroy the system of payments:

> I have a family, and am quite content with the position I have. And to organize a fight against informal payments is to risk your position and a lot in this life (Bogatova et al. 2002, 157)

In any case, the supervision of hospitals and policlinics is, in effect, delegated to medical professionals themselves. One respondent says that there is inevitable reliance on the doctor conscience, which may be the nicest possible euphemism of this delegation of responsibility. Two alternative explanations are possible. It may be the case that doctors are sufficiently powerful. They may refuse to treat an administrator’s relative well, if the administrator prohibits charging patients. They may even quit their jobs in that case, because official wages may be below their reservation wage. Another explanation is that interests of higher and lower levels of administrative hierarchy are more interwoven than any respondent would be ready to admit.

Interviews with the insurance companies in Saint Petersburg allowed me to see the application of this logic of tolerance in the operating procedures of insurance companies. Partly, this will confirm the hypotheses of the previous chapter as to the lack of authority of the formal institutions in defending the patient against illegal charges (see details in Appendix A.3-2).

The only operating procedure dealing with illegal charges is preventive telephone calls in reaction to complaint. This works, obviously, only if a patient complained before paying. Apparently the procedure focuses on direct and unambiguous denial of free care to those who cannot afford paying or whose case is not urgent.

If a patient pays, then, barring especially outrageous cases of extortion, the patient's case is lost. In particular, payments made under formal contracts are not considered as violations of the patient's right to healthcare. Complaint is essential for initiation of any procedure. The majority of complaints are never filed, as they are made in way of a preventive measure (see Appendix A.3-2).

Some respondents justify these informal ways of control, such as use of personal connections, and general tolerance. The head of quality control department at *Nevskaia-Med*, a Saint Petersburg insurance company, claimed that administrative constraints on illegal charges were minimized not out of corporate solidarity, but because fighting the charges would severely damage the healthcare establishments.

It is also important to notice that most respondents recognize the illegal and discriminatory nature of the charges in question. There is no consensus on whether better financing would solve the problem. The root of the problem is considered most often as either low patient right awareness or the difficulties associated with filing a complaint. Appendix A.3-2 provides further details and insights from the interviews in Saint Petersburg.

3.4 Analysis

Upon reviewing quantitative evidence, I suggested the following question: how can one interpret the co-existence of paid and free care with a minority paying, or at least paying more or less significantly? The question was answered by identification of transaction patterns and, wherever possible, motivations behind them.

Consistent with the quantitative data is a picture with concentration of especially large payments in certain wards and for chronic cases. A typical scenario under such pattern of concentration would involve material informational asymmetry that leads to lack of meaningful pre-negotiation and other guarantees of extra quality assurance for the patient. Enhanced with flexibility of arrangements, notably price differentiation, this concentration picture suggests that payments buy not only services, but also a special relationship with a doctor or nurse through an ingratiation effect.

The question of balance between free and paid care is therefore putatively answered by the notion of concentration of payments with the ingratiation component in the patient motivation to pay.

By ingratiation I understand an object of payment that is different from buying a specific service and is normally additional to the latter. Ingratiating means that by paying, a patient aims to provide incentive for good work, but lacks any ostensible means of enforcing an implicit or even explicit contract.

Ingratiating is thus distinguished from two alternative motivations:

- buying of a specific service or procurement of specific effort, for example, acquisition of a specific medication;
- expression of gratitude or benevolence towards a medical professional, for example, sponsoring a hospital where the sponsor is unlikely to be treated.

Respondents, doctors and administrators alike, do not name external control as a material restraining factor. Furthermore, at least some of the restraining functions are delegated to low-level management where it coalesces with professional hierarchy. This last point means that in fact, the effectiveness of control policies very much depends on reliability of the lowest ranks in the chain of command. This justifies concentration on professional-patient relations for explaining the balance of paid and free care.

In this analytical section I will link theoretical arguments about the nature of healthcare services and the empirical picture just described, in order to make precise the notion of patient motivation to pay as a balance between buying a service and buying special relationship. This section makes this idea precise, linking together three topics:

- Balance between buying service and ingratiating
- Professional benevolence
- Trust necessary to enforce otherwise unenforceable explicit or implicit contracts between doctors and patients

Before proceeding, I shall clarify terminology. In the Introduction, I discussed benevolence as limited profit seeking rather than the actual motivation. Professional ethics, professional pride, the fear of causing harm to another human being, an internalized fear of punishment and many other elements can potentially concur to yield the results that will appear as benevolent to the patient. I shall not discriminate among the many motivational elements. Instead, I use the word *benevolence* as a catchall for many internal or internalized constraints on profit seeking.

3.4.1 Buying service and buying a relationship. In this subsection I shall expand on the notion of special relationships emerging in response to informational asymmetry. Benevolence will then appear as a possible, though controversial, cementing agent in building special relationships between doctors and patients. But certainly, special relationships are only a version of a more general story of informational asymmetries in healthcare provision and should be seen as such.

Since Arrow 1963, it is a staple suggestion of healthcare economics that patients, however exacting they are in their claim for value for money, are not able to buy healthcare as such. They cannot find out what they consume before consuming and, more often than not, they cannot do so even after the consumption. Specialist knowledge is required to understand contemporary medicine and the patient's personal experience (i.e. pain) is a rather noisy signal even of changes in health status, leave alone performance of a doctor. Even when a patient pays a nurse for extra quality post-operational care, the patient only tries to ensure that the nurse does her best on the patient's behalf and does not – as a rule – prescribe her a course of action. Therefore, trust is required.

Trust can be established by many means. A second opinion is one way. Reliance on external, statutory or professional, auditing of medical practice is another. It seems that in the specific environment of the Russian statutory healthcare, trust is being established via payment. Patients are expecting that the paid doctor is the best doctor. Such trust I call *special relationships* through *ingratiation*.

Ingratiating and special relationships enter the picture in connection with the pattern of concentrations identified above. Particular stories that embody concentration of payments for cases of decisive and/or prolonged involvement of doctors suggest informational asymmetry as a material factor. People pay exactly when they have little control.

Respondents justify payments partly as inducing them to "work better" (notably, see Bogatova 2002, 118). Whether or not this is so is hard to see. However, this is what a regular patient is likely to believe, for this constitutes a strong motivation to pay even in circumstances when a more detailed quality differential is not suggested or not observable to the patient. Together with buying an extra service or quality differential, patients establish special relationships with doctors and nurses.

It seems that doctors themselves believe that paying is a rational strategy for patients (the second general fact above). As Russians (including one of respondents in the interviews) like to say: “Cheap healthcare is cheap *[meaning, worthless]* healthcare.” If you do not pay for your treatment, you might as well not go to see the doctor. In fact, this is apparently an exaggeration in light of what respondents tend to say about the way public healthcare functions. But the popular belief in the power of money to induce better care is also present. Patients pay to motivate the doctor and nurse, however fuzzy and uncertain such motivation can be.

This should not create an impression that the patient necessarily employs a reasonably successful strategy of purchasing a special relationship. A respondent has already been quoted as saying that the patient is likely to react more positively to a doctor ‘holding their hand’, rather than to one providing good service. This means that patients may be reacting to false quality signals. Mistakes are possible and misguided attempts to ingratiate oneself with a doctor should be expected, especially in a situation of physical suffering and emotional distress. A patient remains within the confines of impressions about a doctor’s performance or promises thereof, which the patient may not be able to ascertain or enforce.

An overall account of what patients are paying for should therefore balance two different aspects of the illegal transactions: that of buying relationship and that of buying service. The two are intertwined and may not be fully recognized by the patient (or the doctor or nurse, for that matter) as such. The patient may not fully appreciate the power of the doctor to vary quality of service, or may rely too much on the presumed incentive fostered by payment. In both cases, the patient will believe that a service is being bought while in fact the only thing that can be arranged through payment is relationship. This opens an avenue for more extreme violations of ethical norms by doctors and nurses who may exploit patient ignorance and misperceptions.

One question remains unanswered. Is there any possibility of rationalizing the strategy of ingratiating? Can this strategy be rather adequate given certain assumptions regarding medical professionals' behavior or preferences? In asking this question, I shall clearly differentiate between two alternative approaches to the problem. The first is to admit that much of the observed behavior is based on customs, which are the final explanation of patient choice. In this vein, what could have looked a mistake, a worthless investment is in fact an act of performing a ritual. Rituals cannot be subject to rationalizations, at least from the performer's point of view.

The other approach is to 'deconstruct' the custom and rationalize patient choice. Rational choice may involve mistakes, but not irrationality of a ritual performer. I shall follow this second approach, and formulate an empirically testable hypothesis regarding the balance between buying a relationship and buying a service. Doing this, I shall also make every effort to understand the ingratiation strategy as a reasonable choice of the proverbial 'economic man' finding himself in an exceptionally involved interaction with fellow human beings.

3.4.2 The notion of trust. The patient entrusts the doctor or nurse to perform certain operation whose nature and effects will largely be unobservable to the patient. The payment therefore becomes somehow linked to emerging trust. In healthcare the ingratiating payment is not for doing something, but for exerting professional effort to do the best. This is what the notion of 'special relationship' is supposed to capture.

This constitutes a difference from the otherwise analogous cases of petty corruption. As the notion of trust becomes central at this point, it seems worthwhile to examine its role in more detail.

Trust will be understood in a very specific sense that can be illustrated as follows. The patient expects that the medical professional will give this patient treatment in excess of what this patient can secure by appealing to law or authorities. Trust is secured by payment. The trusting patient believes that in exchange for a payment, the medical professional will provide this patient with services in excess of what would have been provided without the payment.

The literature contains an extensive discussion of trust. Trust is naturally related to informational asymmetries. A relevant example is Giddens 1990: consumption of complex professional services requires general faith in professional institutions. Here is his definition of trust[48]:

> Trust may be defined as confidence in the reliability of a person or system, regarding a given set of outcomes or events, where that confidence expresses a faith in the probity or love of another, or in the correctness of abstract principles (technical knowledge). (Giddens 1990, 34)

Fiduciary relations between doctors and patients were examined in the literature, too. For example, Parsons 1978 (26-27) establishes fiduciary relations in healthcare on the notion of life-long medical career and self-willed subjection to moral obligations on the part of the doctor.

The trust appearing in connection with special relationships is also a response to complexity and uncertainty. Its link to medical career will be examined below. Two peculiarities must be noted, rather curtly and without giving full justice to the underlying problems. Trusting a medical professional in the course of an illegal or informal transaction, where institutional involvement is minimal, is based on an understanding of professional constraints to unbridled or short-term profit-seeking. Healthcare is, in this sense, an inherently ethical enterprise, which is not the case with workers at a nuclear power plant whom we also trust.

Yet the most important peculiarity is the hypothesis that trust is maintained through payment. Corruption is often regarded in the literature as something opposed to trust, as it offers only illusionary safeguard against excessive complexity. Among many authors, Elster (1989, 266) and Piotr Sztompka (1999, 116) consider corruption as imperfect substitute for trust. Piotr Sztompka says:

> Spreading in a society, it [corruption] provides some misleading sense of orderliness and predictability, some feeling of control over a chaotic environment, some way to manipulate others into doing what we want to do. (Sztompka 1999, 116)

Not attempting to assess the overall validity of the claim, it is important to indicate here that ingratiating payments in healthcare may present a case of genuine trust being generated through what can still be called corruption. In many cases doctors exploit patient trust for personal gains. However, it seems that the system has achieved some overall stability because a mechanism or mechanisms exist that allow the payments to foster something more than just illusion of control on the part of the patient.

The concept of ingratiating payment can be viewed as an alternative to sharp and therefore dubious singling out of quid-pro-quo in opposition to tips and gratitude payments. Many attempts have been undertaken to differentiate payments that are part

[48] Fukuyama 1996 explores trust in its relation to traditional values. My notion of trust is much closer to Giddens' than to Fukuyama's.

of quid-pro-quo relations and those that are not. In the literature on small corruption there are a number of definitions of what sets tips or gifts apart from the bribery. Zelizer (as quoted in Rose-Ackerman 1999, 92) defines tips as “legally optional, informally bestowed, the amount unspecified, variable and arbitrary”. Rose-Ackerman suggests that tips and gifts can be differentiated from prices and bribes by that they are not paid in return for something (*quid pro quo*, Rose-Ackerman 1999, 92-93).

Ingratiation may render the whole category of tips somewhat ill defined empirically. When a payment establishes a special relationship, the payment has all the observable characteristics of what is called gift or tip. Yet, it is a part of what eventually amounts to an exchange of favors. More importantly even, it is hardly possible to separate different motivations. For example, a patient may feel gratitude for preferential treatment. By paying money one both expresses the gratitude and ingratiates oneself with the doctor.

As a matter of fact, studies of small corruption indicate that people tend to bring uninvited gifts and hustle money to low level officials in many situations.[49] These payments too are likely to play ingratiation role without being gifts of appreciation. One important fact is that gratitude as motivation is often cited but is hard to believe, as the word is likely to refer to *quid pro quo* arrangements.

Miller et al. (2001, especially 150-155) cite empirical evidence to the following effect. The distinction between tips and bribes disappears as regards petty corruption in the Czech Republic, Bulgaria and Ukraine. The authors suggested a distinction between the two even stronger[50] than the one by Rose-Ackerman: the money paid or promised in advance is a bribe, the money paid afterwards without a promise in

[49] Sometimes, this happens after the service was delivered, the same way as this happens in healthcare.

advance is a gift. Though some respondents quoted ‘gratitude’ as a motivation for gifts, a majority of 58% indicated that the motivation is either that the officials expect gifts or that further encounters with them are expected. In fact, this figure is an underestimation. The notion of ‘gratitude’ is used in the quoted answers as an elliptic version of the extortion story: “He is very sick, my son, he misses a lot of school. So purely from gratitude I give something to the teachers, so that they will compromise.” This is a clear case of bribery, or illegal exchange of favors. The same can probably be said about quoted references to custom or desire to be polite. These can easily be idiosyncratic (or euphemistic) names for the same old extortion and bribery. The conclusion the authors give is as follows: “Taking both simple extortion and this complex anticipatory variant of extortion together, extortion was only a little less likely to the motivation for gifts given afterwards than for gifts given beforehand” (Miller et al. 2001, 155).

3.4.4 Roles of benevolence. To take stock, both theoretical and empirical arguments exist to the effect that patients engage in more than buying a service. Patients aim to promote trust with the provider. In some sense, however, trust is bought. Gratitude as part of the psychological signature of these processes may be a factor, but there is no reason to think of it as an important alternative to the proposed special relationship story behind payments. But where is the source of assurance for the patient that engaging in a transaction leads to establishing the desirable special relationships?

A natural candidate for the missing link can be freely borrowed from the literature. It is the provider benevolence, understood here as an internalized constraint on profit making. Again, I average over the wealth of potential psychological

[50] In the sense that incentives for gifts and bribes were more likely to be distinguishable empirically.

signatures. Benevolence can play a number of roles. One of them sits well with the inclination of interviewed professionals to justify money taking by low salaries. Taking the statement at face value, one will have to believe that should the salaries have been somewhat larger, the payments would not have existed. Medical professionals are ready to show benevolence and not demand money in excess of a certain target level, lest they lose patient trust in them as benevolent professionals.

The second role of benevolence may show in flexibility of arrangements for taking money. There is for example price discrimination apparently beyond the degree consistent with income maximization by a monopolist.

The problem with these versions of benevolence is that both are reasons for patients not to pay. Trusting a doctor to provide extra quality of service for money is inconsistent with believing in professional benevolence in two senses. First of all, if the doctor is benevolent, there is less reason to pay. Secondly, as long as the provider is benevolent there is less reason to believe that payment will induce better quality, especially if the patient intends to skew distribution of scarce resources in his or her favor. But most importantly, benevolence as simply a constraint on seeking higher income does not yield the required link between payment and establishing trust or special relationships. Benevolence in this sense does not rationalize payment. In the following subsection I will take a different angle on benevolence as well as on payment.

3.4.5 Payment as solution to the problem of trust. The benevolence of providers could explain flexibility of arrangements, as well as account for patient trust in doctors. One consequence of this is that a patient should pay less to the doctor whom he or she trusts most. This does not appear to be the case. Patients are eager to remunerate doctors if reliance on those doctors' benevolence and professional attitudes is high. Such areas have been identified: surgery and chronic conditions, as well as famous doctors in famous hospitals.

Wherever benevolence is expected, it is rewarded by money. This link between expected benevolence and monetary reward must be explained. The second aspect remaining unclear from the previous subsection is the actual rationalization of ingratiating. Though it appears very likely that patients tend to ingratiate themselves with doctors, the reasoning behind this strategy remained only vaguely specified.

As a tentative answer to these two questions, I suggest to look at payment not only as a variable in doctor utility function, but also as a signal coming from the patient. In the construction to follow, payment triggers professional benevolence, which consists in providing good work for money even when the paying patient cannot condition payment on quality of service. Professional benevolence again refers to this outcome rather than the psychology behind it. This is the use of the word 'benevolence' advertised in the introductory remarks at the beginning of this section.

Suppose that a payment gives a patient access to a reputed doctor in a reputed clinic. Thereby, the payment opens access to the pool of professionals who are likely to be not only humanly benevolent, but also professionally benevolent, for possibly very egoistic reasons, such as pride. Also, these professionals have reputation and they therefore have a stake at maintaining it. Somewhat metaphorically, payment gives access to professional benevolence.

The alternative situation is when payment does not ensure any immediately observable extra service. In this case, payment still puts the patient in the 'paying patients' group. A doctor would be interested in increasing this group. An incentive emerges to keep the reputation of someone honoring an implicit promise of better quality. This regulates professional behavior in cases where quality of service is also imperfectly observed.

In general, payment signals to the medical professional that it is worthwhile to keep up the good work for this patient. The paying patient places him- or herself in a category different from the rest, from the non-paying patients. There is little reason for a doctor to maintain a reputation among those who cannot pay. Poor people do not have friends to whom they could pass favorable information about the doctor. Or, as Balázs Varadi suggests, non-paying patients fail to demonstrate their commitment to their own health. As a result, doctors will be less benevolent towards those who do not pay than towards those who do.

Finally, as a demonstration of respect and of ability to appreciate medical professional effort, payment can trigger doctor benevolence more directly. This last point rests on certain psychological assumptions, which are however rather plausible. Taking for the moment at face value doctors' indignation about their miserable salaries, I shall suppose that a benevolent doctor could view payment as a sign of a mutual understanding with a patient, that money should be reciprocated with quality service.

Payment is definitely a signal of ability and willingness to pay, which for a greedy doctor means that more money could be coming. For someone concerned about his or her reputation, payment means a due reward and also furtherance of a life-long career towards higher professional status. For a proud but poor doctor, payment means appreciation of professional effort. In all these scenarios, payment serves as a signal of the type of patient who is ready to pay. Payment is part of a game between a career medical professional and a patient who needs help but also needs reasons to trust the professional.

In the parlance of game theory, patients effect a separating equilibrium by paying and thereby distinguishing themselves from greedy, inconsiderate, or poor fellow patients. As with benevolence, I shall admit that there must be many psychological signatures of the link between payment and benevolence and trust. I use the word ‘ingratiating’ as a catchall for all of them. Ingratiating is therefore a purpose of payment separate from buying a service. It consists in inducing professional benevolence by signaling ability and/or willingness to pay, now and possibly in the future.

So far, patient motivation has been discussed. Patients expect medical professional to reciprocate payment with good work. But is actual benevolence on the part of the professional necessary for the whole system of ingratiating payments to work? Apparently, the answer is negative. A greedy doctor will find it to his or her advantage to demonstrate benevolence at least occasionally, in order to convince the patient that he or she is trustworthy.[51] One could suggest that the significant ostensible flexibility of many scenarios of payments and ostensible application of moral constraints is likewise an attempt by the medical profession to keep the patient assured that profit is not the only concern of doctors and nurses. Otherwise, the grounds for trust on the part of the patient would disappear.

Table 3.4.1 compares three alternative approaches to payments for services free by law: gratitude, remuneration and signaling. The gratitude story explains the payments as triggered by patient benevolence towards the doctor. This is the simplest and the least convincing story, included here for completeness. The remuneration story claims that patient payment is made in exchange for a service. In fact, as long as such payments defray costs of service, the remuneration story is part of what I have called the under-funding paradigm. The new story is that of signaling. Payment is designed to trigger professional benevolence. Payment plays the role of a signal in a 'meta-game' involving repeated encounters among many doctors and patients.

Gratuity is something cast in doubt by this and related research. Remuneration is made less relevant by assuming informational asymmetry (the patient does not know what to pay for) as well as by provider benevolence. Signaling is a story of payments for health services free by law that sits well with informational asymmetries and benevolence. Table 3.5 reflects relation between stylized empirical facts and the three approaches.

Table 3.5 Three approaches to payments of interest

Stylized fact	Payments are rationalized for the patient as		
	Buying a service	Tipping out of gratitude	Signaling willingness and ability to pay in order to trigger benevolence
Provider benevolence	-	N	+
Informational asymmetry	-	N	+
Pattern of concentration	-	N	+
Quality differentiation between paying and non-paying patients	+	N	+
Limits to differentiation between paying and non-paying patients	-	+	+
Institutional arrangements for enforcing payment	+	-	-
Limited institutional involvement enforcing payment	-	+	+

Note: "+" means that a given theory is justified by the fact, "-" means the opposite, while "N" means that there is no determinate relation between the two.

The signaling story appears to be a necessary, though by no means exclusive, component in the picture involving a patient paying for something free by law under

[51] Fudenberg and Tirole 1991, Chapter 9 is a relevant discussion of the behavior of rational agents

the informational and other healthcare-specific constraints. Signaling in no way substitutes the more basic "buying a service" interpretation of the payments. Signaling however explains how buying a service happens in a situation where mistrust and lack of accountability should have ruined opportunities for trade.

With this general construction in mind, I shall proceed to a short review of potential supporting evidence for my notion of ingratiation, or, more generally, for the signaling nature of payments.

The theory leads to the following predictions. If payments are concentrated in the group of chronic illnesses, where continuous relations between doctor and patient are normally found, the signaling role of payments is partially confirmed. Interviewed doctors directly and indirectly confirm that payments are indeed more habitual among chronic patients than among emergency cases.

Secondly, if patients are more prepared to pay for unobservable professional effort and other hard-to-control elements of medical care than for observable additions of quality or quantity, then the hypothesis of ingratiating and benevolence inducement becomes more plausible. So far, one has seen that patients are ready to pay for unobservable effort, reputation and other intangibles.

Finally, if quality differentiation in favor of paying patients remains even where patients cannot control what they buy, this confirms the idea of special relationship as alternative to buying service and as a rationalizable strategy for a patient. If benevolence is there, it must be induced by payment.

To recapitulate the proposed empirical test, the hypothesis of buying special relationship (ingratiating) as trigger of benevolence and trust would be supported if:

facing reputational effects of their choices.

1. Payments concentrated in areas where repeating encounters between patients and doctors are likely;
2. Doctors differentiate quality even when patients cannot control them;
3. Patients are ready to pay more for what they cannot control than for what they can control.

To perform the first test, one simply measures correlation of payment with diagnosis. The second and third would require more elaborate design, as more hidden information is to be retrieved. All three tests could be run on the basis of a household expenditure survey among users of hospital facilities with utilization of medical services being controlled for.

3.5 Conclusions

The co-existence of paid and free care is the main result of quantitative research. A minority pays and correlation with the personal income of patients is not strong. A large part of payments are formalized. Qualitative evidence has been used to examine the nature of this balance of free and paid care. My conclusions are as follows. Patients indeed pay for services free of charge by law. Payments appear to concentrate in areas involving surgery as well as chronic conditions. Child delivery and venereal diseases are also such areas of concentration. In general, arrangements of payment are endowed with all forms of flexibility, notably price discrimination.

Both the nature of healthcare demand-supply relations and the concentration picture with strong elements of flexibility suggest the following peculiarity of professional–patient relations. The patient buys not only service, but also a special relationship with the doctor. The latter plays the role of partially benevolent and trustable attorney, promising to do his or her best of the patient's behalf. The patient responds with a payment and a good deal of trust invested in the all-powerful professional.

In this picture, readiness to pay is rationalized by actual or expected quality differentiation. But because of peculiar circumstances of informational asymmetry, a patient has to rely on less direct mechanisms of ensuring extra quality of service. Then payment can be rationalized as triggering professional benevolence, with benevolence being a host of motivations constraining profit-maximization. A patient wants to be among the paying patients, not among the non-paying ones. The link between payment and benevolence is sought in such professional motives as reputation maintenance, desire to uphold explicit or implicit promise to exert good effort for a paying patient and veritably benevolent emotional disposition towards those who understand doctor need for better pay.

The concentration idea implies a very limited appeal of public under-funding as the cause of private payments. Under-funding is supposed to be uniform, but patients concentrate their resources in certain areas. This suggests that private payments do not supplement doctor wage to reservation level, but rather constitute a rent in excess of that level. If under-funding as explanation has a limited appeal, then better public funding as a remedying mechanism is also unappealing. It will not change patient willingness to pay at key junctures, surgery in particular. It will also not eliminate readiness to sell medication in crooked ways on the part of doctors, because apparently much of the proceeds from such sales constitute rent. Policy-relevant implications of this theory will be developed in the next Chapter.

Chapter Four. A Conceptual Framework of Private Payments in Public Healthcare

4.1.Introduction

This Chapter attempts a speculative modeling of payments for services free by law. The question this Chapter answers is as follows:

What is a plausible system of causal relations that would makes such payments a rational strategy for the parties involved?

I will use the word *rational* in the sense of a strategy being a reasonable response to certain needs by an actor facing certain constraints. This means that I suspend any judgment on the psychological signature of such rational response. I will be answering the question from the point of view of how payments affect the production and distribution of professional effort.

Such are the two sides of the coin: the payments as rational strategies and the payments as factors effecting distribution of costs and benefits and eventually decisions. The parties involved are of course patients, providers and a generalized regulatory and financing authority, referred to as the 'state'.

This Chapter finalizes the theoretical development alternative to the notion of replacement of free services with paid services as a result of public under-funding. Payments of interest will be understood primarily as various forms of informational rent earned by a medical professional and paid by a patient facing the need to motivate that professional to work better.

The alternative notion of payments is what has been called the under-funding paradigm (Chapter One): payments bridge a supposed gap between reservation wage and official salary of a medical professional. This function of payments will be reflected in the construction to follow. The arguments borrowed from Chapter Three will however suggest that this is a secondary function. That of informational rent is the primary one.

In this Chapter I will identify the most important and empirically confirmed causal links and then recreate, in a speculative way, a hypothetical causal net, centered on payments for services free by law. In doing this, I am intentionally silent on quite a number of less likely or less important causal links. Based on causal inference, I suggest a qualitative model in which rents are part of a rational equilibrium.

Such a rational equilibrium generates explanations regarding stability of the system and the role of health policies. Starting with micro-level interactions, I thus ascend to macro-level analysis and explain the regulatory gap (Chapter Two) and financing policies.

The Chapter will be structured as follows. Section 4.1 considers the aggregate level of provider-patient relations. There, I answer questions regarding distribution of costs and benefits aggregated over the healthcare system. Section 4.2 descends from this aggregate level to that of individual provider and individual patient relations, looking for more details.

Section 4.3 considers the effects that payments for services free by law have on public policies. The section discusses of the possible reasons for the state to condone payments for services free by law. Section 4.4 summarizes the findings.

4.1 Aggregate Level

I will start with a simplification. Suppose there is a range of services that are to be provided free of charge at least according to the law. I concentrate on the effects of private payments on supply of these services. Suppose that in absence of such payments, a given amount of the services are produced. Payments can either increase that amount, decrease it, or have no effect. These effects are defined here with respect to a counterfactual situation of everything in the system staying the same, except that patients do not pay anything. The effects of private payments will be called *' supply-increasing'* if they increase aggregate supply, *' supply-irrelevant'* if they do not. If the supply decreases, the payment can be called *'supply-decreasing'*. Each payment can make only one of these three effects.

The case of supply-decreasing payments is unlikely but not impossible. A monopolist first supplies a predetermined amount of services for a fixed amount of public funding and has incentives to maintain the level of production. Then, through some corruptive mechanism, the monopolist acquires an opportunity to charge the beneficiary of the services on a piece-by-piece basis. Then, the monopolist could prefer to decrease the amount of services in order to maximize profits, trading this new incentive against the pre-existing incentives coming from public financing and penalties for under-supply. In other words, bribery may worsen service.

It appears that the major effects the payments for medical services have are supply-increasing and supply-irrelevant. Looking more closely at both these effects, one can make the classification somewhat finer. Supply-increasing effects mean increased efforts, material resources and personal incomes of medical professionals. Increase of resources can first of all refer to more medication becoming available, but also any other material resources bought for private money. Effort is measured as time, attention, and application by a medical professional. Each total supply-increasing payment can effect increase in

1. effort and personal income of professionals, and/or

2. material resources

Supply-irrelevant payments only increase the personal incomes of professionals. Increases in personal income fall into two categories. (Here I subsume corporate incomes into personal incomes, assuming that the relevant components of payments paid to the cashier are extra wages and taxes). The notion of a reservation wage is important. If personal income does not amount to a certain minimum, called *reservation wage*, professionals leave their jobs. This means that increasing personal income to such a minimum is a supply-increasing effect. Yet there is no reason to suppose that supply-increasing effects stop here. Payments in excess of reservation wage can cause extra service, too. This is because providers have a superior informational position and can earn the rent a patient would be ready to pay so as to induce optimal effort. Chapter Five makes this important causal link formally justified.

By rent I will understand incomes in excess of reservation wage. More precisely, the following 'master' definition will be used:

> *Rents are payments by patients made in expectation to receive services or goods, when such services or good must be delivered free of charge and human and non-human resources necessary for the delivery are available (paid for from public funds).* [52]

In other words, rents are the part of personal income in excess of the reservation wage. In view of the overall picture, this rent can, with reluctance, be called illegal. Chapter Two showed in what ways transactions of interest may or may not be considered illegal. At this point, it may not be important whether the rent is illegal.

[52] In the language of Russian regulatory documents, this means that paid care is delivered 'instead of the guaranteed free services'.

Rent should not be understood as inherently contrary to moral values or social welfare. The notion of rent reflects the intuition that payments for services free by law are not a direct result of public under-funding. Instead, rents can be welfare-increasing solutions to sector-specific problems, as I hope to show in the rest of this Chapter and in Chapter Five.

Both supply-increasing and supply-irrelevant payments can be rents. But if supply-increasing payments may or may not be rent, supply-irrelevant payments are necessarily rent. The reason is that without these payments, the total amount of services, therefore material resources and efforts would have stayed the same. This means that supply-irrelevant payments necessarily come on the top of a doctor's or a nurse's reservation wage.

Supply irrelevance will be associated with various misguided attempts to ingratiate oneself with a doctor, mistaken beliefs in the power of money, but also absolutely voluntary contributions. Supply irrelevance obviously links to general flexibility of payment arrangements. But supply irrelevance is broader than this flexibility: extortion and fixed tariffs can easily be supply irrelevant, too. To summarize, three effects of private payments on aggregate supply of services have been identified:

1. Increase in material resources, which is executed only by supply-increasing payments;
2. Increase in effort, which results from supply-increasing payments; such payments can increase effort either by being up to the level of reservation wage and making a doctor stay on the job or by being a rent, which induces more effort as an incentive payment;

3. Increase in personal income, which can either be up to the level of reservation wage (non-rent, therefore supply-increasing payments) and rent (both supply-increasing and supply-irrelevant payments).

Yet this aggregate consideration does not distinguish among many intuitively different strategic situations. For example, a doctor might be ready to serve a poor or sympathetic patient free of charge. But if he does that, there will be less chance that others will go on paying. The doctor must then credibly threaten the patient with not supplying a service free of charge. The resulting effect of a payment is an increase in supply of the service, but only in the context of a certain game. If the patient had refused to pay, and others had followed the suit, the supply of services might have gone on without any decrease in quantity of quality. The effect is an increase of supply of service not against the benchmark of aggregate production in absence of any payments, but against that of the counterfactual supply of services to this particular patient in absence of the private payment.

There is also a more general question to be asked about private payments and their effects. In what sense does a payment for services free by law represent a rational strategy for a patient?

First let me dwell on the notion of rent. If there is no rent, payment simply makes more service possible. This makes payment rational as long as marginal valuation of money for the patient is less that the utility from extra service. On the contrary, it is not so clear why a patient would pay rent. Why not, for example, bargain to receive services free of charge, as long as there are enough material resources for their production and the doctor already has his or her reservation wage?

This consideration justifies going into more detail at the level of individual interaction. Supply-increasing and supply-irrelevant effects will have to be considered against a different benchmark, namely the provision of services to patients who do not pay or even refuse to pay.

4.2 Individual Level

Why would a patient pay for something that is free by law, when the patient knows it? Why not bargain for lesser price or even try and get a service free of charge? Private payments do increase total supply of services, in various ways, affecting either material or human capital. But this consideration does not amount to a good reason for an individual patient to pay money to an individual doctor.

Such a problem in fact mirrors the initial considerations in Chapter Three, towards the end of Section 3.1. Statistical data indicate that payments take place. The amounts are significant enough to affect quantity and quality of total services. The rationality of paying remains an unresolved problem, notably in light of concentration of payments within a minority of patients. Unless it is clear why patients pay, it is not clear whether effects of the payments are as significant as the numbers might indicate.

This section shows how strategies of paying and accepting payments with all kinds of effects can be an equilibrium of rational interaction. Discussion is kept free of formalism. Many things that are discussed here in qualitative format are given a formal treatment in Chapter Five.

4.2.1 Classification of effects. At the most general level, the effects of payments for an individual patient may increase the counterfactual supply of service, through increasing resources or effort and, as a consequence, personal income and possibly rent of the provider, institutional or individual. Such effects will be called *additive*. Resemblance to supply-increasing effects is obvious. It is important to notice that in difference from the aggregate level, I shall talk about effects rather than payments, because a payment can have different effects at the individual level.

Similar to the aggregate consideration, payments that do not change the amount of service in terms of either effort or material resources will be called supply irrelevant. But here comes the most important difference with aggregate level. The benchmark here is what happens to the patient if he or she does not pay. There are two possibilities. First, it may be that the service will still be provided, and the respective effect is then called *neutral*. To make any payment neutral, one must assume that if a patient refuses to pay, the service is still provided. Such an arrangement can be expected to hinge on the patient's ignorance of the situation, or on such motives as gratitude of the patient. This consideration implies that neutral effects constitute some extreme case. If a payment having neutral effect cannot be characterized as made out of benevolence on the part of patient, lack of effect on supply implies rather consequential informational asymmetry.

The second possibility is that in absence of the payment the distribution of effort or resources would benefit another, paying patient. In this case, there is no increase in the overall effort or resources. The effect is therefore supply-irrelevant at the aggregate level. But it is definitely not neutral at the individual level. It seems appropriate to consider such effects as a special class and call them *distributive* payments. They influence the distribution of scarce resources, such as the doctor's time, limited effort, available medication or utilization of equipment by patients, paying and non-paying.

There is a specific category of effects, which is related to subsidization of the non-paying, presumably poorer patients by the paying, presumably richer ones. Part of proceeds is used by a provider to increase the supply of services to those not paying, out of benevolence or specific medical professional considerations. The effect is additive in the sense that the service to a poor patient would not have been provided without them, but can be neutral if the service can be provided to a rich patient free of charge. In general, a payment can have both distributive and additive effects.

The importance of distributive effects transpires in the following argument. Distributive effects place patients into a state of effective interaction, or even a conflict over scarce resources. In absence of payments, such resources will be distributed according to urgency and place in a queue, as must be with services free of charge. Distributive effects are exactly when urgency and queue cease to be sole determinants of effort, medication, etc., distribution. As long as a payment effects only a redistribution of scarce resources, it is pure rent. Such a payment is obviously supply-irrelevant at the aggregate level.

If patient X decides to pay and thus obtains more scarce resources, another patient Y may also want to pay, at least in order to preserve equality of distribution. In this situation, a patient’s choice to make a distributive payment is a choice between two or more groups of patients, ranked according to their readiness to pay. If this ranking translates into quality of treatment varying across the groups, then a patient chooses to pay because others do.[53] While being an interesting and important example of rent payments this game does not transpire at the aggregate level.

To summarize, rent payments lead to two types of effects:

1. Those emerging from informational asymmetry (additive and neutral effects)
2. Those emerging due to strategic interaction among patients (distributive effects).

[53] Formal treatment of rat-race situations, when agents are locked in an inefficient equilibrium, can be found in Galasi and Kertesi 1989 for corruption and in Landers 1996 for competition among young professionals.

Informational asymmetry is necessary for rent payment to be rational strategy, unless effective competition among patients emerges.

4.2.2 Benevolence, trust and special relationships. The preceding description of effects at the individual level is the simplest version possible. There is one important element missing in the picture. Though informational asymmetry and/or interaction among patients are necessary, they are by no means sufficient reasons for a patient to pay rent. It is not quite clear why a patient would pay rents, either towards additive, distributive or neutral effects, instead of applying other forms of pressure on medical professionals in effort to receive what is free by law.

This section will employ the concepts developed in Chapter Three, section 3.4. There, it was suggested, in conformity with the empirical data, that the system of payments is based on patient desire to invoke special relations with medical professionals, triggering their professional benevolence by payment. Ingratiating payments are rent payments. Paying rent can be motivated by a desire to establish special relationships. The latter is a sufficient, though not necessary condition for rent payments to be part of a rational equilibrium.

Here is a more detailed consideration of this link. Both from general considerations and empirical facts, one knows that healthcare provision is fraught with informational asymmetries, whose consequences are mitigated by professional benevolence. The dire circumstances of a patient having to rely on hopefully benevolent doctors lead to peculiar motivations for private payments. As has been emphasized in Chapter Three, the system of payments for services free by law can only exist as a stable institution if patients believe in inducing better service by payments without being able to know even the nature of those services. Such patient-provider relations were called ‘buying a relationship’ as opposed to ‘buying a service’.

Doctors and even nurses are not exactly sellers on a market, but rather attorneys promising to do their best possible for their clients, that is to say, poor patients. I suggested in Chapter Three that the balance of motivations on the part of medical professionals will include benevolence, readiness to reciprocate the payment with better service, and at times willingness to renege on the implicit or explicit promise of better work. The patient then balances buying a service and buying a relationship. Such a balance of motivations is possible if professional pride, reputation, fear of conflict with a paying (as opposed to a non-paying) patient and natural benevolence somehow provide rational grounds for patient trust in his or her doctor.

This much has been said in Chapter Three. Now one can see that all these features are compatible with additive and distributive effects, as well as with neutral ones. First of all, neutral effects are a natural consequence of informational and often even cognitive superiority of the doctor over the patient. Neutral effects come in many forms, of varying degree of ‘innocence’ on the part of the doctor or nurse. In general, these are cases of reneging on an implicit or explicit promise of extra good work. To be sure, neutral effects can be driven by doctor desire of being professional and rendering service according to need and not pay.

Neutral effects are normal only if there is significant informational asymmetry in doctor-patient relations. This means that such effects correlate with attempted but failed ingratiating. As long as payment is intended to establish special relationships and create grounds for trust in doctors, neutral effects are cases of misguided trust, when money are not sufficient to create due incentive

Additive effects are different in that patients receive what they pay for, but doctors enjoy informational privileges in this case, as well. In general, an additive effect is achieved when a conscientious and/or benevolent doctor makes good on an explicit or implicit promise of better work. Additive effects may or may not be associated with rent payments: the payment need not be in excess of the reservation wage.

The notion of ingratiation is not necessary for additive effects. Chapter Five presents an informational structure under which paying rent is a rational strategy, because only partial assurance of quality service can be obtained.

Likewise, ingratiation is a suitable motivation for payments aimed at distributive effects. Paying rent can easily be a more effective means of tilting resource distribution in one’s own favor than scandal or imploring. Distributive payments are rationalized by patient desire not to stay in the non-paying group. Such payments are rent by definition. From patient perspective, such payments can only be justified by the fact that doctors for whatever motives favor paying patients. The more a patient pays *compared to others*, the more he or she receives in terms of professional effort and other scarce resources to be distributed. As with additive effects, ingratiating arises when informational asymmetry is sufficiently high. Chapter Five demonstrates that distributive effects result from payments, aimed at increasing supply of service, when there is a capacity constraint, but not enough informational asymmetry for ingratiating.

In conclusion, rent payments can be seen as a natural consequence of demand-supply relations in circumstances of informational asymmetries that trigger trust and special relationships. Paying a rent to a doctor is only natural, either towards additive, distributive or neutral effects on supply. The system is further balanced by the fact that doctors are endowed with ostensible benevolence. This benevolence here may or may not be an emotional equivalent of good-heartedness. It may only be a facade to placate public opinion, but for the patient it nevertheless means flexibility, price discrimination, and trust in medical professionals. Hence the following conclusion: Rents are related to informational asymmetries, that is to say, they are informational rents. Ingratiating and special relationships based on expected benevolence lead to a specific and rather extreme form of rents. Chapter Five investigates informational rents without ingratiation and with benevolence as a factor tempering rent-seeking on the part of medical professionals.

Rendering the system stable. The system of private payments for services free by law may contain seeds of self-destruction. It is based on trust in the force of money and in the benevolence of the doctor. Adverse selection in favor of greedy doctors and competition among patients for resources leading to neutral effects can destroy this trust. And yet, there appears to be some countervailing mechanism.

If relations of trust are important, doctors will aim at maintaining special relationships against the background of many moral choices they are obliged to make. Distributive effects imply a moral conflict for the medical professional, when a doctor faces a choice between making good on a promise to work more for money and helping someone who has not paid. Additive effects imply a conflict between benevolence and greed with respect to some patients, when a doctor sets the level of effort supplied free of charge.

Neutrality can also be part of a moral conflict. On the one hand, neutrality means distribution by need and not pay. On the other hand, neutrality means reneging on an implicit contract to supply better care for a paying patient. Doctors choose between obligations to non-paying patients and promises to the payers.

These moral dilemmas can be difficult as personal choices. Yet their existence indicates the presence of a stabilizing mechanism within the system of payments for services free by law. Doctors do not demonstrate extreme greediness that could jeopardize the relations of trust. Nor are they ostensibly ready always to favor paying patients over non-paying all the time, which could have been a trust-demolishing factor, as well. The moral choices they make will then contribute to stability of the system. The general stability of the system of payments for services free by law comes from the fact that patients are mostly satisfied with what they receive for their money.

What could further stabilize the system is the very same interaction among patients that leads to distributive effects. It was suggested that patients induce better work by paying money, because doctors frequently prefer to make good on their promise of better work. Now suppose that in one case there are only supply-increasing effects, while in another there are strong distributive effects. In the first case, the decision to pay or not to pay is made by a patient according to whether he or she trusts the doctor, believes in possibility of a better service and decides not to bargain any further or at all. In the second case, there are all these motivations, too. Yet there is another one: other patients may pay, thereby inflicting losses on this patient. Under distributive effects, a patient does not want to belong to a non-paying group. If however everyone pays an equal amount, there is no distributive effect. While intending distributive effects, a patient gets only additive ones.

If everyone pays with the intention to skew distribution of scarce resources, then a doctor receives a lot of rent and does not have to deal with many moral dilemmas. For example, good work for paying patients will not obstruct exercise of benevolence. Distribution will not have to be skewed, at least not too much. This consideration suggests that while payments for services free by law, and especially their rent component, may be a source of moral trouble for doctors, these moral problems become less significant, if more patients are willing to pay.

4.3 Rent Extraction and Health Policies

So far, the effects of payments for services free by law have been seen against two benchmarks: aggregate supply with no patient paying and supply to a non-paying patient, with others possibly paying. In both cases, public financing and control were kept fixed. I shall relax the latter assumption.

At this point I introduce the state as some generalized public institution that arranges public financing and control of healthcare. Patients want to get healthy while doctors and nurses want to exercise they professional duties, help patients and earn their living. What does a state want? I assume that the state balances minimization of public spending on healthcare, maximization of service supply and maintenance of care accessibility.

The state in this construction will have two ‘control variables’: the amount of public resources disbursed for healthcare and administrative policies. Below I consider three causal links between policies and private payments. One is between public and private financing; another involves informational asymmetries and impossibility of total control; the third is between lobbying and control and financing policies. I answer the following question, which is a specialization of this Chapter’s research question:

In what sense is allowing private payments for services free by law optimal for a state that aims to maintain healthcare provision, minimize spending, and maintain equitable access to care?

4.3.1 Private and public financing.

Imagining a counterfactual world without private payments in the previous two sections included the assumption that public financing would stay the same. If the state wishes to maintain a certain level of healthcare provision, it could increase public financing to increase service provision.

This means that private payments decrease public financing. This statement however requires further discussion. Prime facie, it may seem that not all private payments decrease public spending in this way, only those that have supply-increasing effects. Yet there are two reasons to think otherwise.

Much of private money inflow appears to be rent towards additive, distributive or neutral effects. Distributive effects may also decrease access to care for poorer patients. This is a detrimental effect from the state's point of view, as the state cares about equity of access. Maintenance of a certain level of provision can technically happen without some of the private payments, namely pure rents leading to distributive and neutral effects. This means that should the state be able to distinguish between supply-increasing and non-supply-increasing payments, it could prohibit the latter payments to placate the public opinion and maintain equity. It would not then be required to increase public spending to maintain healthcare production.

In reality, a state can either indiscriminately prohibit or equally indiscriminately permit private payments. Only a very limited amount of discrimination in this regard can be assumed. Such limited discrimination will not be policy relevant. The outcome of the situation is such that private payments indeed save public resources, even if these are rent payments.

This being the first effect of private payments on public policy, an immediate extension suggests itself.[54] Because the state aims to minimize public spending for each level of healthcare production, effective rent extraction is a rational strategy.

Rent extraction can be defined with the help of an example. A salaried waiter receives tips from patrons. The employer understands that and sets the salary below the reservation wage, because tips cover the difference.

In application to payments in healthcare, the state can set public spending so low as to make private payments sufficient only for supply-increasing effects. This is the effective rent extraction: the state takes advantage of patients paying more than a minimum for care. Non-rent, supply-increasing payments can be considered as extracted rent, at least in part.

Rents, including distributive and supply-irrelevant payments are ways to save public money. But the causal link here does not amount to replacement of one financing with another. Public capitation financing and private fee-for-service payments provide rather different incentives for providers and lead to different compositions and size of service output. They also have different rent components.

The most important feature of the policy switch from public financing to private financing is that we *do not need to assume* that charges of patients for services free by law are caused by decline in public financing. Financing policies can cause non-rent payments, but not rent payments. The latter result from the rather natural desire of patients to become healthy and providers to become wealthy in circumstances of informational asymmetry.

[54] I am indebted for suggestion to consider this ramification to Balázs Varadi.

The goals that the state pursues thus stand in inherent conflict. If the state is interested in equality of access to healthcare, a blanket permission of private payments trades off equity against spending containment. Extracting rent risks decrease in service supply. Thereby, quantity of health care is traded off against spending containment.

If there are rents to be extracted in the above fashion, there is additional flexibility for the state in deciding how much to spend on healthcare. Again, because the state cannot observe which payments are rent and which are not, it cannot fine-tune decrease in public financing so as not to compromise equity.

4.3.2 Administrative reasons to allow private payments. Previously, public financing and private financing was assumed to be similar as regards their incentive effects for providers of care and medical professionals. That is however not so in reality. Money flows are differently conditioned for the two sources of financing. Incentive effects are different. Informational asymmetries are different, too.

First of all, the state cannot observe provider effort and therefore must pay rent to maintain a desired level of that effort. This being a most general consequence of informational asymmetries, the burden of payment can be relocated on to patient shoulders. This train of thought is pursued in Chapter Five. There, a formal model is based on an incentive compatibility constraint linked to unobservable effort. The state can however 'relax' this constraint by permitting the provider to charge the patient. Then the patient pays supply-increasing rent payments (towards additive effects at the individual level). The amount of healthcare provision desired by the state is maintained informational asymmetries notwithstanding.

Control mechanisms in public healthcare management are imperfect and often subject to manipulation. From this perspective, private payments are a solution not only to the financial, but also the administrative constraints the state faces.

An example will help to see how this can happen. In Yaroslavl' *oblast*, frequency of post-operational deaths was at some point used to measure performance of public healthcare (see Antipova et al. 1999, 21). This measure aims to solve the problem of low quality post-operational care, which is very much a job for nurses. Certainly, this scheme can be subject to manipulation. Instead of making nurses work more, the system can decrease the number of useful, even critically important operations. In this situation, private payments supply a correcting incentive. Indeed, a prime example of payments for specific nurse services is exactly post-operational care, which can be no less important for patient well-being than the operation itself.

Some elements of doctor performance are not observable to external controllers, but are, at least as general impressions, to patients. For example, speeding up of procedures is often in hands of patients with money. In such circumstances, supply-increasing effects of private payments can be greater than supply-increasing effects due to public financing. This is an additional reason to allow the former. As the literature and this research suggest, professional pride, benevolence, fear of punishment for gross violation and other factors temper the negative consequences of this permissive policy.

4.3.3 Lobbying doctors and private payments. The last causal link that rationalizes permissive policies towards payments for services free by law is as follows. Rent has so far been considered as extracted from patients. Yet state institutions can also be subject to pressures, which will be somewhat imprecisely called lobbying. Lobbying doctors aim to extract additional financing from the state. Such additional financing can be considered as a supply-irrelevant payment from public coffers.

By allowing private payments, or, more precisely, their rent component, the state can reduce lobbying pressure and placate medical professions. This argument is in potential conflict with the earlier claim that the observed persistence of illegal rents is based on the difficulties of their elimination. If the state cannot eliminate rents otherwise than by distorting the system of healthcare supply and even that at large cost for public funds, then the medical profession cannot be prevented from lobbying for extra financing by allowing charging patients. If illegal rent is a factor preventing lobbying in conjunction with a threat to enforce the law if lobbying happens, then any difficulties with law enforcement impart credibility of the threat.

4.3.4 Benevolence and special relationships: their role for public policies. The state has two good reasons to allow private payments. First of all, this allows lower public spending, healthcare production being fixed. Secondly, private payments support an adequate production of healthcare, where the state could have failed due to administrative barriers. Yet the unrestricted right to charge would also mean inequitable access to care. Distributive and supply-irrelevant effects can have negative impact on access to care, making it inequitable. Supply-decreasing effects are also unwelcome from the state's point of view, because they both hurt the poor and potentially decrease aggregate supply of care. The state needs to balance the costs and benefits of the policy of permission. Ideally, some form of co-existence of paid and free care would be in order. In fact, this is what has been observed both in quantitative and qualitative data.

There remains a problem as to how exactly the state could reach this golden balance of two forms of healthcare financing. This co-existence cannot simply be allowed by the state, whatever the state gains from it. One way to temper these negative effects is to fine-tune permissive policies, so as to minimize distributive and supply-irrelevant effects. However, these permissive policies result from exactly the financial constraints and informational asymmetries that would also prevent the said fine-tuning. Fine-tuning is therefore of limited value.

Coarser policies of limiting tolerance could have involve securing a minimal level of free healthcare guaranteed for all. Alternatively, the state may tolerate payments only from certain diagnostic or social groups. Finally, the state may restrict the influence of payments on supply of services. It may, in particular, refuse to tolerate the most extreme cases of distributive effects, while being absolutely tolerant toward neutral effects.

While all these mechanisms to control excesses of private financing of care are technically feasible to an extent, the dominant option employed by the state is de facto delegation of partial control over the system of private payments to lower-level management and, most importantly, doctors themselves. The notion that the behavior of doctors is governed not only by greed but also benevolence and professionalism helps rationalize this delegation.

To see how this happens, notice that the actual limits of tolerance cannot be glimpsed from at least declassified documents. In fact, as far as such sources are concerned, tolerance seems to be unlimited, unless scandals trigger formal procedures. One potential explanation of this is that the limit to tolerance is determined locally by MHI funds, private insurers, and local administration, and, certainly, the medical profession itself. The latter is actively and even formally involved in formulation of healthcare policies.[55] Upon such a delegation, limits to tolerance become informal. Respective decisions are made on a case by case basis. Most importantly, the bulk of control over excessive charges appears to accumulate at the level of individual health institutions.

Professional moral constraints, mixed with a vague notion of the danger posed by the brazen pursuit of profit, help restraint and possibly prevent extortionate behavior with little or no help from the regulatory institutions. Ethical or prudential checks and balances comprise the actual constraining agents. The latter are created and implemented at the lowest level of the public healthcare hierarchy, the professional and managerial rank and file.

The state does not intervene because there is no need to intervene. Those who are supposed to be regulated and constrained implement the regulation and constraining themselves, and will do so as long as their motivation preserves. A varying degree of flexibility reflects varying motivations. Together with local variations in effective demand, it shapes local forms of co-existence of free and paid care. All the parties involved, including the state appear to be in favor of the status quo.

The long-run implications of the current regime are however impossible to glimpse. All the stabilizing mechanisms, both internal to the profession and external to it, are not about to disappear. The overall impression is that the system as it is now is sufficiently stable. Yet there are certainly countervailing factors, such as natural desire of patients to be less subject to vagaries of their current situation. One can only speculate what policy shifts are necessary to effect any significant changes. I am inclined to think that deregulating the whole sector to promote voluntary insurance and the financial autonomy of providers could at least enhance transparency.

[55] The “triad” of regional administration, regional MHI fund, and regional medical association together

Deregulation is further justified by the fact that even now patients and doctors somehow manage to maintain balance in their relationships.

4.4. Conclusions

The notion of payments for services free by law as a stable system can now be seen in light of respective rational strategies of patients, providers and the state. These strategies answer quite common needs, such as money and health, and are related to natural informational and financial constraints. Eventually, such rational strategies strike a balance between different extremes. The co-existence of free and paid care is the result.

Rent has been identified as income in excess of reservation wage of providers and medical professionals, given the amount of service produced. This rent is essential to account for the effects of informational asymmetries, but also of the ensuing complexities of motivations. Rent is a regular part of the equilibrium, being paid by patients, tolerated by the state and contained by the professional rank and file of healthcare institutions themselves.

The message of stability I tried to outline towards the end of Chapter Three is continued in the idea of a rational strategic equilibrium. It is possible that the stability as discussed here is a short-lived state. Some mechanisms jeopardize stability of the system, such as the self-selection of greedy doctors. But these will take time to have any effect. At the moment, for patients tucked in their beds and doctors thinking through their diagnoses and treatments, the decisions to pay and accept are sensible enough.

formulate and sign the healthcare budgets and determine the rules of the game. Private insurers are only consulted

The rent component of private payments cannot be caused by decline in public financing. Moreover, lower public financing may mean that patient will have to spend more in the form of non-rent, resource-increasing components. This negative correlation between rent from private sources and public financing leads to the last message.

The last message is of direct policy relevance, if in a negative sense. Private payments, that induce additive effects on supply and are not rent, can be considered as extracted rent. In other words, public financing is set so low because there are private payments that will bridge the gap, for example, between reservation wage and official salaries. However it would be wrong to say that low public financing causes patients to pay. Public policies change the economic nature of the payment, turning a rent payment into a non-rent (or extracted rent) payment. This subtle difference has the following implication.

Suppose private payments largely constitute a rent, which is partly extracted by the state. This means that a more generous financing policy will simply increase rents, because it will abolish rent extraction in the first place. Here I assume that the institutional and legal framework remains the same: prohibition of private payments does not occur. If extra public funding is administratively tied with higher volumes of production (more personnel, more equipment or medication, more quality control), then this additional public money will end up replacing non-rent supply-increasing payments currently paid by the patient, while private payments will be increasingly various forms of rent. If there is no link between extra money and extra effort or resources, then additional public funding itself becomes part of provider rent. In the circumstances, better public financing can mean over-funding of healthcare, rather than a sensible strategy to eliminated unwanted transactions.

Distributive payments violate patient right to healthcare, and are largely the most clearly illegal component of the rent. In the suggested simple framework, raising public funding cannot eliminate rent. Additional complex assumptions must be made to rationalize better financing as a method to reduce distributive payments. In other words, as long as the conflict of interest preserves, extra funding will not be useful as a policy mechanism. It will simply lead to more rent at the expense of private non-rent payments, possibly causing even more distortions. Not having to pay for medication, patients will pay for less work for other, poorer patients, as per rules of the rat-race.

Payments for services free by law have thus been conceptualized in terms of actors' strategies. Stability of the system also means that only improbably drastic administrative measures can stop money flows from patients to doctors. When a system is in equilibrium, it requires a large shock to change. The next section promotes some of the ideas outlined above to the level of a formal model. Thereby I check these ideas for consistency and also derive further hypothetical causal links between policy choices and private payments under various attending circumstances.

Chapter Five. A Model of State-Provider-Patient Relations

In this Chapter I will perform a partial consistency check on the ideas presented in the three previous section. To this end I will build, solve and discuss a formal economic model, intended to reflect some points made above.[56] This also means that I will have to simplify significantly the picture of payments for services free by law, if I want to make them a part of a formal model.

Balancing the desire to match the intended interpretation and the need for simplicity, I make the following choice of substantive elements to be formalized and checked for consistency.

First of all, the model will be used to explain the current limited enforcement of free healthcare entitlements by the state. It was suggested (Section 4.3) that the state, or its relevant institutions avoided enforcement of free healthcare entitlements for essentially two (generalized) reasons: cost-containment and lack of administrative capacity. Both reasons lead to allowing illegal rent in order not to pay legal one. Both reasons must be reflected in the model.

Section 4.3 suggested that one reason for the state to look the other way as regards illegal charges for health services could be the role the illegal charges play in solving a rent-effort trade-off in procurement of the services. If made into a consistent and intuitive formal model, this trade-off could be a clear-cut case of mixed 'administrative' and 'financial' reasons for allowing illegal rent.

Section 4.2 suggested that effects of a private payment on supply of services to an individual can be classified into distributive, additive, and neutral. This classification

[56] Explaining interaction between providers, patients, and state institutions has so far been conducted almost exclusively, if often implicitly in terms of utility maximization. This choice of maximization

will be preserved in the model, where payments will bring about the respective welfare effects.

Finally, in sections 4.2 and 4.3 provider benevolence played an important role as the source for stability of the system of payments. In the model, providers will be benevolent and consequences of this will be seen, but the meaning of benevolence will be more traditional. There will be no ingratiating and triggering benevolence by payment. Benevolence will only temper profit-seeking.

The model will be constructed as follows. A regulator and procurer of healthcare services, the State for short, faces the problem of financing healthcare in presence of uncertainty about effectiveness of services for eventual health status and professional effort exerted by a service Provider. A financing scheme should therefore have an incentive-inducing component, to take care of the said uncertainty. A choice between state financing and patient financing is to be made. This is called "regime choice".[57]

The regimes will be called:

- The State Pays regime and
- The *Laissez-faire* regime

Regime choice depends on the comparison of equilibrium health outcomes for both regimes. Laffont and Tirole 1993 and Laffont 2002 have been the source of methodological and substantive insights as to precise formulation and solution of this moral hazard problem.

The Chapter is constructed as follows. Section 5.1 succinctly relates the intuition behind the model. Section 5.2 defines the model formally and section 5.3 discusses the model's relation to my source of inspiration, Laffont and Tirole 1993. Section 5.4

paradigm over behavioral modeling justifies translation of the preceding discussion into a more rigorous language.

[57] Regime choice is a major tool of formalizing welfare over a set of institutional arrangements. Laffont 2002 contains examples of regime choice aplenty.

solves the model. Section 5.5 formally discusses implications of the model in terms of regime choice. Section 5.6 interprets and discusses the results. Section 5.7 discusses possible objections and alterations to the model.

Formal discussion is confined to sections 5.2, 4, and 5. Even in these parts, formalities are kept at minimum and simple intuitions are provided to keep track of the reality the model is designed to reflect. The reader who is not interested in formal details may wish to jump from section 5.1 to section 5.6.

5.1 An Intuitive Exposition of the Model

First component: rent-effort trade-off. Informational asymmetry is a prerequisite for almost all the essential effects of payments for services free by law, as I made every effort to emphasize in Chapter Three. Now I will show precisely how informational asymmetry produces private payments.

Section 4.3 suggested that the state cannot observe the actual effort exercised by medical professionals and, eventually, institutions. Effort is non-measurable or at least non-contractible. Hence effort's central role in models reflecting the roles of uncertainty in regulatory economics. Following the suit, I propose to look at the position of professional providers of medical services as one endowed with extremely important information regarding the patient and services rendered to the patient. Namely, neither the state nor patients can observe the relation between effort and health outcomes.

In the classical model of unobservable effort, the procurer of a service faces a trade-off between paying the agent for the service in excess of its cost (rent) and eliciting too little effort. Without reviewing the model in detail, consider the following intuitive explanation of the rent-effort trade-off.

Let a financing institution (the principal) observe the outcomes of an agent institution's activity imperfectly or periodically, so that a large margin of mistake can be expected. The agent has preferences different from those of the principal, and will therefore use its superior information in its own rather than principal's favor. Then the principal has a very simple means of making the agent do only what the principal likes. The principal will punish the agent extremely severely for every deviation from the principal's interest. This solves the problem.

This is not always feasible in a world of limited liability. And ours is fortunately one. Then there is another way to prompt the agent to act in the principal's best interest. It is to reward every sign of compliance excessively so that the agent's incentives to cheat on the principal are corrected in the right direction. An additional factor that would affect the make of such an incentive mechanism is risk-aversion of the parties involved. Here, I shall disregard this factor, assuming that every agent is risk-neutral.

In the model, effort is unobservable and the agent's 'uncorrected' incentive is to set effort below the principal's optimum. The principal observes the outcomes of effort, namely, health outcomes for patients. Health and effort are imperfectly correlated. If the principal gives a particularly good reward for each good health outcome in excess of a minimum, then the agent receives a rent. This rent prompts the agent – the Provider -- to increase its effort and thus corrects the agent's incentives.

This constitutes a rent-effort trade-off, familiar to the economic literature.[58] The solution to this trade-off will be presented in the form of regime choice, where one of

[58] Chapter One in Laffont and Tirole 1993 contains a history of the problem.

the regimes is exactly the observed Laissez-faire, or payments for services free by law.

Second component: Regime choice. The rent-effort trade-off is incorporated into the following game. State, Provider and patients interact in the model, maximizing their own utilities with respect to their ‘control variables’. More precisely, the State decides whether it is going to procure medical care for each patient, or let patients order requisite amounts of care and pay on their own. If State decides to pay for care (*State Pays* regime), then State decides on how much care it orders and how much it pays for it. Otherwise patients must. They decide the amount of care they want to receive and offer the Provider a contract, matching the effort and reward (*Laissez-faire* regime). The State wants to make patients as healthy as possible for as little money as possible. Patients want to make themselves as healthy as possible for as little money as possible.

The Provider maximizes the money it gets from either the State or patients, minimizes the effort exerted, and maximizes the health of patients. The latter element is called ‘provider benevolence’, or just ‘benevolence’ for short.

Patients differ in their physiological condition, which defines the amount of effort needed to bring a patient to a given level of health (health outcome). The respective measure of effectiveness of treatment will be called ‘effort-effectiveness’ or effectiveness of a patient.

The State and patients can observe the health outcome, but cannot observe the effort the Provider exercised to deliver it. Neither the State nor patients can observe effort-effectiveness. In other words, they cannot deduce effort from health outcomes for any number of patients.

In this set-up, patients and State solve essentially one problem: making the Provider exert an optimal level of effort, subject to the fact that they cannot observe effort directly. State is also deciding on whether patients will have to pay for healthcare. The basic dilemma for State is therefore between allowing patients to pay for themselves and paying for them.

Third component: effects on patient welfare. If patients pay (*Laissez-faire*), difference in their willingness to pay will cause difference in effort they receive in optimum. The model will reflect the main effects of payments on patient welfare: distributive, additive, and supply-irrelevant. To ensure all these effects, a constraint on total effort that Provider can exercise is imposed, together with Provider benevolence.

If patients pay, a patient's value of money influences effort. For some patients effort will be lower than in the case of State paying. For some other patients, effort will be higher as compared to the situation of State paying. This makes legitimate and non-trivial the question, how exactly inequality in ability/willingness to pay affects regime choice. This question is answered formally in section 5.4 and informally in subsection 5.5.

In conclusion, the central player in the model, State, faces a trade-off between rent to the provider and potentially poor expected health outcome in the Laissez-faire regime policy. This trade-off can be seen as an alternative to the more traditional health economics model of an oversupplying provider, who enjoys about the same informational advantage over the financing principal. It appears that the Russian reality is more amenable to a model, where Provider under-supplies, unless paid a rent. This is definitely true about the case of State being the principal, but also appears to be true about the case when patients are principals as sources of incentive payments.

In principle, it is possible that the state subsidizes patients directly, possibly compensating for income inequalities. Also, it is possible, and even probable, that in reality the third-party and direct financing are combined. Then, the provider receives extra incentive to provide more effort and poor patients are again effectively subsidized. These complications, however, do not enhance the illustrative and consistency-check role of the present model, and are omitted.

5.2 Notation and Definitions

The players in the game are: the Provider, the State, *n* patients (illness cases), generically denoted by *j*. Provider takes care of these *n* cases by according each of them some professional effort. Effort is unobservable to anyone except Provider. Effort spent on patient (case) *j* is denoted by e_j. Total effort $E \geq \Sigma e_j$ (summation over *j* from 1 to n) is the overall capacity for treating patients.

Professional effort results in the health status of a patient,

(1) $h_j = h(e_j, \beta_j)$

Here β_j is a measure of effectiveness of medical efforts for patient *j*, stochastically distributed between β_{min} and β_{max}; $0<\beta_{min} < \beta_{max}$ according to some distribution *F*. In other words, there are n parameters β_j identically and independently distributed according to *F*.

$$\begin{aligned} &\frac{\partial h_j}{\partial \beta_j} > 0; \\ (2)\, &\frac{\partial h_j}{\partial e_j} > 0; \\ &\frac{\partial^2 h_j}{\partial e_j^{\ 2}} < 0; \end{aligned}$$

For example:

(3) $h_j = \ln(1 + e_j \beta_j)$

This function will be used later on as a particular case of the model. The rest of the notation includes:

1. π is the policy variable. π=1 means that the State enforces the following law: it is illegal for the Provider to accept money from patients for treatment. This is a strengthened version of the actual notion of free care guarantees;
2. t_j is monetary transfer from the State to the Provider for taking care of patient *j*; μ_0 is value (to the State) of a unit of such transfer.
3. the Provider's utility from patient *j*'s good health is measured by Bh_j, where *B* is a positive constant; 0<*B*<1;
4. Patients may pay for their treatment if π=0, and then p_j is the amount paid by patient *j*. Value of a unit of payment for patient *j* is μ_j>1.

To summarize the informational assumptions:

1. State and patients observe health status and do not observe effort. Equivalently, State and patients do not observe effort and parameters β_j, which they believe are each drawn independently according to distribution *F* with density *f*.
2. Provider observes both effectiveness of treatment β_j and its own effort.[59]

The informational assumptions correspond to observable cost and unobservable cost-reducing effort in Laffont and Tirole 1993 (see Chapter 2 for their basic model). Box 5-1 below summarizes the differences between the model on which Laffont and Tirole 1993 is based and my model.

[59] It may seem that the idea of distributive payments, that is to say, payments that induce differentiation among patients as to the level of care accorded to them, could have been encapsulated in increasing marginal dis-utility of effort to Provider. It is constant in the suggested model. Making dis-utility nonlinear, and thus more realistic, would not undermine the basic results, only change the equilibrium rent-effort trade-off. Nor would it replace the role the effort constraint plays in reproducing distributive payments. If patient j does not make a distributive payment, there is more care provided to patient k. In other words, Provider maximization for one patient is coupled to maximization for another patient. Convexity of dis-utility of effort does not per se lead to such coupling. The basic conceptual difference to remember is of course that between patients (made interdependent through the total effort constraint) and types (of a patient).

This is the most general formulation of the model. To obtain all the interesting results, one does not need to consider a continuous range of β's, nor more than two patients. There will only be two values of β: $\beta^1 < \beta^2$. For clear notation, subscript will refer to patient, superscript will refer to type. β^1 happens with probability q.

A further and even less essential simplification is that the number of patients is also reduced to two. Patients are numbered by j, j=1,2; $1<\mu_1<\mu_2$.

e_j^i is effort exerted for type i of patient j;

h_j^i is health outcome for type i of patient j.

Now this simplified version will be formulated, solved, and discussed. The continuous version, which is a regular problem in optimal control, is presented in Appendix Six.

First of all, both exposition and solution will be helped by Revelation Principle. A few definitions will remind the reader of the idea:

- *Revelation mechanism:* A particular mechanism representing a game of incomplete information. The action of an agent whose type is unobservable to the principal only consists of a report about his type (private information).
- *Revealing (truthful) equilibrium* of a revelation mechanism occurs when each type reveals itself.

The following theorem holds:

Revelation principle for implementation in dominant strategies (Osborne and Rubinstein 1994 Lemma 181.4). N is set of players, C is set of outcomes, P is set of preference profiles over C, G is the set of all strategic game forms. Choice rule is a mapping from P to C. If a choice rule f is Dominant Strategy Implementable, then:

1. *f is truthfully implementable*

2. *There is a strategic game form G, where each player reports the true profile as part of a dominant strategy equilibrium.*

The game will have a dominant strategy equilibrium. This means that – in contrast to Nash implementation where some equilibria can be non-truthful – we will not have to bother about non-truthful equilibria.

The Provider's program. The Provider reports the type of patient *j* (which is β^j) so that State or patients then procure required equilibrium effort and pay respectively based on the report. This allows the following lightening of notation:

e_j^i is **equilibrium** effort exerted for type *i* of patient *j*;

h_j^i is **equilibrium** health outcome for type *i* of patient *j*;

t_j^i is **equilibrium** transfer (State pays regime) for type *i* of patient *j*;

p_j^i is **equilibrium** payment (Laissez-faire regime) for type *i* of patient *j*;

The Provider maximizes:

$$\begin{aligned} &\sum_{j=1\ldots n} Bh(e_j^i, \overline{\beta}_j, \beta^i) + t_j(\overline{\beta}_j) - e_j^i(\overline{\beta}_j) \\ (4)\, &wrt: \overline{\beta}_j; \\ &s.t.: \sum_{j=1\ldots n} e_j^i \leq E \end{aligned}$$

with respect to reported effort effectiveness parameters, $\overline{\beta}_j$'s.

For the Provider, it is profitable to claim that the actual $\beta_i = \beta_l$. Therefore, the condition under which truth-telling is forced upon the Provider's incentive compatibility constraints (j=1,2) is:

$$(5)\, Bh_j^1(\pi) + \pi t_j^1 + (1-\pi)p_j^1 - e_j(\pi, h_j^1, \beta^2) \leq Bh_j^2(\pi) + \pi t_j^2 + (1-\pi)p_j^2 - e_j(\pi, h_j^2, \beta^2)$$

where $h_j^1(\pi)$ and $h_j^2(\pi)$ are truthful equilibrium values of health outcomes and therefore they are shown as dependent only on regime choice. Here $e_j(\pi, h_j^1, \beta^2)$ is effort for patient j as a function of this patient's equilibrium health outcome and type. Mathematically, this is simply the inverse of health outcome function of effort with type as a parameter. Denoting:

(6) $\varepsilon_j(\pi) = e_j^1(\pi) - e_j(\pi, h_j^1, \beta^2)$

I rewrite this incentive compatibility constraint as follows:

(7) $Bh_j^1(\pi) + \pi t_j^1 + (1-\pi)p_j^1 - e_j^1(\pi) + \varepsilon_j(\pi) \le Bh_j^2(\pi) + \pi t_j^2 + (1-\pi)p_j^2 - e_j^i(\pi)$

All efforts and health status functions are equilibrium choices by State or patients, based on a report by Provider. ε_j designates the rent (in terms of effort) that Provider earns by mis-reporting effectiveness of its effort. Substantive interpretation of these epsilons is as follows. ε_j is how much less effort Provider spends on achieving the health status optimal for the ineffective type when the true type is effective, as compared to the optimal effort for the ineffective type.

Finally, it is reasonable that Provider not work unless the following is true:

(8) (Individual Rationality constraints) $Bh_j^i - e_j^i \ge \pi t_j^i + (1-\pi)p_j^i$;

The State's Program. The State's program is:

$$(9)\quad \begin{array}{l} Max W = q[h(e_1^1,\beta^1) + h(e_2^1,\beta^1) - \mu_0\pi\{t(h_1^1) + t(h_2^1)\}] + \\ (1-q)[h(e_1^2,\beta^2) + h(e_2^2,\beta^2) - \mu_0\pi\{t(h_1^2) + t(h_2^2)\}] \\ wrt: \{e_j^i\}, \pi \\ s.t.: \\ 1. \sum_{j=1}^{2} q e_j^1(\pi) + (1-q)e_j^2(\pi) \le E; \\ 2. (7); \\ 3. (8). \end{array}$$

The first constraint is the total effort constraint. The second constraint is incentive compatibility constraint. The third constraint is the individual rationality constraint.

Remark 1. Is *W* a social welfare function?[60] *W* is different from the regular social welfare of welfare economics. It does not include the full welfare of patients as well as that of doctors. But it is possible to prove that this does not diminish normative significance of the model. One counter-argument could be that omitting welfare of a small interest group (medical doctors) allows better to reflect political reality, that is, government preferences. Moreover, no information is lost through omitting some components of social welfare. The alternative 'full' social welfare function would be as follows:

(R1) $SW = W + \pi E(T + Bh(\pi = 1)) + (1-\pi)E(P + Bh(\pi = 0) - (1-\pi)EP$

where E is a shorthand for 'expected'; T is total transfer; and P denotes total payments by patients. $\pi E(T + Bh(\pi = 1)) + (1-\pi)E(P + Bh(\pi = 0) - (1-\pi)EP$ is the 'missing' part of total welfare. $\pi E(T + Bh(\pi = 1)) + (1-\pi)E(P + Bh(\pi = 0)$ is the expected welfare of doctors; $(1-\pi)EP$ is the expected payment by patients, which was not included in W.

The monetary part of the difference between SW and W is easy to deal with. Patient payments are lump sum transfers and thus social welfare-neutral. State transfers can be valued at the shadow price of public financing and then $\mu_0{}^{W}=\mu_0{}^{SW} - 1$. In SW, public transfer is valued as a loss to the budget and gain to the Provider; hence I can set $\mu_0{}^{SW}>0$, to ensure a net loss. The only difficult part is Bh in the Provider utility function. It is however possible to express B^W as a function of B^{SW} so that all changes in the endogenous variables due to the switch from *SW* to *W* are 'swept under' changes in exogenous parameters. Let a transfer for some patient be $t = t_0 - B^{SW}h$. The Provider benevolence decreases the transfer needed. The condition to be satisfied is:

(R2) $(t + B^{SW}h) - \mu_0{}^{SW}(t_0 - B^{SW}h) = -\mu_0{}^{W}(t_0 - B^{W}h)$ where

(R3) $\mu_0{}^{W} = \mu_0{}^{SW} - 1$

hence

[60] I am indebted for the suggestion to consider this problem and an insightful discussion of it to Balázs Varadi.

(R4) $B^{W} = B^{SW} \frac{(\mu_0^{\ W} + 1)}{\mu_0^{\ W}}$

Two models prescribe the same amount of optimal effort, if the respective benevolence parameters relate as in (R4).

Patients' program. Patients maximize the following (only for the Laissez-faire regime π=0)

$$
(10)\quad
\begin{aligned}
&Max: q(h_j^1 - \mu_j p_j^1) + (1-q)(h_j^2 - \mu_j p_j^2)\\
&wrt: \{e_j^i\}, \pi\\
&s.t.:\\
&1.\sum_{j=1}^{2} qe_j^1(\pi) + (1-q)e_j^2(\pi) \le E;\\
&2.(7);\\
&3.(8).
\end{aligned}
$$

The question is what regime is chosen under what conditions. The algorithm of choice is as follows. First one defines the equilibrium effort and health status for each patient for each regime and then compares the difference in total health status in two regimes with the amount of transfer due from the State in regime π=1. This is a backward induction, which simulates the State solving its maximization problem. Application of backward induction here is justified by the fact that there will be a single equilibrium. The fundamental justification of backward induction is that regime choice is made in advance of all moves by players other than the State and these players cannot strategically affect regime choice. They cannot for example blackmail the State into choosing a regime that the State would not otherwise have chosen.

In the following, the model will be solved for the case when the total effort constraint does not bind. The constrained case will be outlined. The choice of regime is explained in terms of the solution obtained. A comparative statics exercise will return conditions under which the State prefers to pay incentive payments and those under which the State delegates the task to patients. It bears repeating that very little of substance is lost due to considering the two-type, two-patient case, and the conclusions to follow must be read as quite a generic result.

Comparison with Laffont and Tirole 1993 model. The two models are similar in spirit, though they reflect two fundamentally different situations. Laffont and Tirole 1993 have a principal compensating expenses of an agent who works on a project and possesses superior intelligence about the possibilities of cost minimization. The principal wants to minimize the project cost. My model has a principal who buys an amount of service from a provider for a third party and does not know to what extent the service improves health of that third party.

Table 5-1. Comparison of two models.

Laffont-Tirole	This model
Observable variable	
Cost	Outcome of treatment
Unobservable variable	
Effort	Effort
'Master equation'	
Cost = cost effectiveness parameter – effort (linear relation)	Health status = f (Treatment effectiveness, effort); f – nonlinear (concave, increasing) function
Effort dis-utility	
Convex in effort	Linear

In words, the nonlinear component, which makes the incentive mechanism possible, is in the agent's aversion to hard work in the Laffont-Tirole model. In my model, though the agent dislikes work, non-linearity crops up in patient response to treatment. In my model, the State orders less health care than in the first best for the ineffective patient. In the Laffont-Tirole model, the principal agrees to less than the first best effort from the ineffective type and this affects only the amount of money needed to compensate the costs of the agent. Of course, this difference reflects the difference in the initial set-up. In the latter case, the principal minimizes the costs of a publicly useful project. In my model, the principal arranges supply of a service to a third party balancing cost and quantity of service.

Another substantive difference in the basic set up is that in my case, the principal is not a benevolent government, while in the Laffont-Tirole case, it is. I also make the agent benevolent in the sense that the agent's utility function partially coincides with that of the principal.

5.3 Solution

Unconstrained case (the total effort constraint does not bind)

State Pays regime π=1. First consider π=1, the State Pays Regime. Individual Rationality only binds the ineffective type β^1, while Incentive Compatibility only binds the effective type. This is because rent increases in type. Solving (9) I obtain the following first order conditions (q_2 =1-q_1):

$$(11)\ q(1+\mu_0 B)\frac{dh(e_j^1,\beta^1)}{de_j^1}-q\mu_0-(1-q)\mu_0[\frac{d\varepsilon}{de_j^1}+B\frac{dh(e_j^1,\beta^1)}{de_j^1}]=0$$

$$(12)\ (1+\mu_0 B)\frac{dh(e_j^2,\beta^2)}{de_j^2}-\mu_0=0$$

as well as the following expressions for optimal transfers and rent:

$$\begin{aligned}&t_j^1=e_j^1-Bh_j^1;\\(13)\ &t_j^1=e_j^1-R_j;\\&R_j=t_j^2-e_j^1=\varepsilon-B(h_j^2-h_j^1);\end{aligned}$$

where R_j is the rent received by Provider who serves a high effectiveness type patient j.

(11) and (12) can be obtained by setting π=1 in (9) and then differentiating (3) with respect to efforts, substituting transfer for the effort-effective type expressed through all other terms in the incentive compatibility constraint (7) and that for the effort-ineffective type expressed with help of the individual rationality constraint (8).

Consider the above expression for rent term by term:

R_j =a+b, where

a=$\varepsilon(\pi) = e_j^1(\pi) - e_j(\pi, h_j^1, \beta^2)$ is how much less effort is required to provide equilibrium health status of the ineffective type (h_j^1), if the true type is effective;

b= $B(h_j^1 - h_j^2) < 0$ is the difference in optimal health status for the two types weighed with the benevolence parameter. The sign follows from the fact that under truth-telling optimum effort is higher for the high effectiveness type and health status increases in effort.

The first order conditions are incomplete without a guarantee that Provider gets a non-negative utility from serving the ineffective type. This is implied by the individual rationality constraint. This problem will be taken care of immediately.

Now we notice that

$$(14) \frac{d\varepsilon}{de_j^1} = 1 - \frac{de(h_j^1, \beta^2)}{de_j^1}$$

The second order conditions are as follows:

$$(15)\ q(1+\mu_0 B)\frac{dh^2(e_j^1, \beta^1)}{de_j^{1^2}} + (1-q)\mu_0[\frac{d^2 e(h_j^1, \beta^2)}{de_j^{1^2}} - B\frac{d^2 h(e_j^1, \beta^1)}{de_j^{1^2}}] < 0;$$

$$(16)\ \frac{dh^2(e_j^2, \beta^2)}{de_j^{2^2}} < 0 \text{ (this is true by assumption).}$$

where e_j^1 and e_j^2 are the equilibrium efforts.

I assume from now on that

(17) $q(1+\mu_0 B)-(1-q)\mu_0 B>0;$

(18) $\dfrac{d^2 e(h_j^1,\beta^2)}{de_j^{1\,2}}\geq 0$

so that the second order conditions fulfill (this is a sufficient though not necessary condition).

Remark 2. Before plunging into further calculations, notice the intuition behind the first order conditions:

$(1-q)\mu_0[\dfrac{d\varepsilon}{de_j^1}+B\dfrac{dh(e_j^1,\beta^1)}{de_j^1}]$ in (11) is nothing but the derivative of expected rent with respect to effort for the effort-ineffective type. Changing amount of effort accorded to the ineffective type of patient changes the rent to the *effective* type. To wit, if *B*=0, (13) implies the rent does not change in the equilibrium effort of the *effective* type at all. This is because the utility from cheating depends exclusively on the effort required for the *ineffective* type, unless health status enters incentive compatibility constraints. The benevolence of Provider decreases the required rent. The size of this effect depends on the second derivative of *h(e)* in the relevant neighborhood.

Consider now a particularly simple functional form:

(19) $h_j^i=\ln(1+\beta^i e_j^i)$

Now I plug (19) into (12) and (13) and a simple calculation yields:

(20) $e_j^2=\dfrac{1+B\mu_0}{\mu_0}-\dfrac{1}{\beta^2}$

(21) $e_j^1=\dfrac{\beta^2 q(1+2B\mu_0-\frac{B\mu_0}{q})}{\mu_0(\beta^2-\beta^1+\beta^1 q)}-\dfrac{1}{\beta^1}$

(21) can be explained as follows. In the limit *q*=1, (21) becomes:

$$e_j^1 = \frac{1+B\mu_0}{\mu_0} - \frac{1}{\beta^1}$$

This is the first best effort, which obtains when no rent is due. And the equilibrium efforts conform to the basic result of contract theory, that is to say, the less effective type receives less effort in equilibrium.[61]

In conclusion, (11) and (21) imply that the optimal effort for the effort-ineffective type is less than the first best, because varying effort for this type affects the rent due to the effort-effective type, according to the Incentive Compatibility constraint.

Laissez-Faire Regime π=0. Patients solve essentially the same problem as State does: they have to induce Provider to work as much as optimal and have to pay it a rent for working more for the effort-effective type. The difference emerges from the fact that now each patient optimizes the effort for him/herself and therefore the difference in marginal willingness to pay, μ_j 's, is to show up in calculations.

If the total capacity constraint does not bind, the first order conditions are as follows:

$$(22)\ q(1+\mu_j B)\frac{dh(e_j^1,\beta^1)}{de_j^1} - q\mu_j - (1-q)\mu_j[\frac{d\varepsilon}{de_j^1} + B\frac{dh(e_j^1,\beta^1)}{de_j^1}] = 0$$

$$(23)\ (1+\mu_j B)\frac{dh(e_j^2,\beta^2)}{de_j^2} - \mu_j = 0$$

For the special functional form:

$$(24)\ h_j^i = \ln(1+\beta^i e_j^i)$$

a simple calculation yields:

[61] If q is close to zero, the ineffective type may even receive negative effort. As this would hardly be possible in reality, one can consider only values of q which are above a positive minimum, or impose a

$$(25)\ e_j^2 = \frac{1 + B\mu_j}{\mu_j} - \frac{1}{\beta^2}$$

$$(26)\ e_j^1 = \frac{\beta^2 q(1 + 2B\mu_j - \frac{B\mu_j}{\mu_j})}{\mu_j(\beta^2 - \beta^1 + \beta^1 q)} - \frac{1}{\beta^1}$$

Again, one can see that $e_j^2 > e_j^1$, e_j^2 is first best, while e_j^1 is lower than first best.

Constrained case.

I will not provide a final solution to the constrained case of the model, when the total effort constraint binds. The hypothesis to be checked with direct calculation is as follows.

The constrained case includes the following scenarios, three for the regime where State pays and four for the regime when patients pay:

- Patient j with type 1 meets patient k with type 1, with probability q^2
- Patient j with type 1 meets patient k with type 2, with probability q (1-q)
- Patient j with type 2 meets patient k with type 1, with probability q (1-q)
- Patient j with type 2 meets patient k with type 2, with probability $(1\text{-}q)^2$

Because patient willingness to pay matters only in the regime, when patients pay, for the State Pays regime, scenarios 2 and 3 are equivalent. The following is the hypothetical solution, which holds if the constraint does not affect incentive compatibility conditions.

State pays regime $\pi=1$. Intuitively, the distribution of effort shall happen according to effort effectiveness. The more effective type should receive more effort, while if both types are of same effectiveness, both receive same level of effort. From symmetry

suitable additional constraint on effort. However, this can happen only if benevolence is more than zero, so that State can use benevolence to force truth-telling.

considerations, it is evident that shall the constraint bind only in case 4, the optimal effort for the high effectiveness type would be *E/2*. If the constraint binds in case 1, the optimal effort for low effectiveness patients would be *E/2*, as well. The only non-trivial case is 2 (or, equivalently, 3). There, the distribution of effort is to be governed – we expect – by effort effectiveness, with other parameters affecting sensitivity of optimal effort to variation in effort effectiveness. To sum up:

- Effort for the ineffective type paired with an ineffective type = *E/2*
- Effort for the ineffective type paired with an effective type <*E/2*
- Effort for the effective type paired with an effective type =*E/2*
- Effort for the effective type paired with an ineffective type >*E/2*

Laissez-faire regime π=0. The patients are now in a situation of strategic interaction. The equilibrium specified for the unconstrained case is not longer possible: there is not enough effort for both patients: at least one of the patients must suffer. The patients effectively enter an auction for Provider's effort. The auction means that every unit of Provider's effort is delivered to the patient with highest valuation of it for the payment equal to the valuation of the second patient. Of course, this verbal description is only a hypothesis, and whether indeed the strategic situation described amounts to a second-price auction with a single pure-strategy equilibrium will be seen through calculation.

5.4 Regime Choice: Effects of Benevolence and Inequality

So far, the focus has been on optimal effort. This is just one of the State's two control variables. The other is the regime.

The different regimes imply two different solutions to the rent-effort trade-off, while sharing a number of properties, already derived in the previous subsection. Namely, effort increases in type; the higher effectiveness type receives the first-best while the lower effectiveness type a sub-optimal effort. The State's utility is:

(27) $W = 2q(h_j^1(\pi) - \mu_0 \pi t_j^1) + 2(1-q)(h_j^2(\pi) - \mu_0 \pi t_j^2) + 2qh_j^1(\pi) + 2(1-q)h_j^2(\pi)$

The utility from both regimes has been calculated as the first step in backward induction that the State makes. The second and last step is to maximize (27) with respect to regime. We shall start by comparing health outcomes in both regimes and then subtract the expected rent paid by State from their difference.

This subsection focuses on two things. First of all, the effects of benevolence and inequality are explored to see the consequences of the model in terms of its intended interpretation: healthcare provision. Secondly, these effects are interpreted in light of regime choice.

Benevolence, as is shown shortly and is rather self-evident from first assumptions, increases (27) in both regimes. But how does it affect regime choice? This is the first major question to be answered, for benevolence could potentially explain why certain types of policies persist.

The second important question is the role of the difference in patient willingness to pay for regime choice. Social inequality is behind this difference, which makes the question important to answer. In the following, regime choice will be explained formally with an eye on the role of benevolence and on that of inequality.

In general, benevolence decreases effort-compensating payments required for a given level of effort (and healthcare), but also decreases the rent. By the same token, it changes the optimality conditions for rent-effort trade-off, making decrease of effort for the ineffective type more profitable for the principal. This all can be glimpsed from the first order conditions above, and will now be demonstrated formally.

First of all, I rewrite the first order conditions so as to see the possible effects of benevolence:

$$(28)\ q(1+\mu_0 B)\frac{dh(e_j^1,\beta^1)}{de_j^1}-q\mu_0-(1-q)\mu_0[\frac{d\varepsilon}{de_j^1}+B\frac{dh(e_j^1,\beta^1)}{de_j^1}]=0$$

$$(29)\,(1+\mu_0 B)\frac{dh(e_j^2,\beta^2)}{de_j^2}-\mu_0=0$$

$$(30)\,q(1+\mu_j B)\frac{dh(e_j^1,\beta^1)}{de_j^1}-q\mu_j-(1-q)\mu_j[\frac{d\varepsilon}{de_j^1}+B\frac{dh(e_j^1,\beta^1)}{de_j^1}]=0$$

$$(31)\,(1+\mu_j B)\frac{dh(e_j^2,\beta^2)}{de_j^2}-\mu_j=0$$

(29) and (31) describe the plight of the effective type. B decreases the optimal slope of the health outcome function as compared to the situation when B=0. In other words, benevolence increases the optimal effort for the effective type. How does benevolence affect the equilibrium conditions. First I rewrite the F.O.C.'s:

$$(30)^*\frac{dh(e_j^2,\beta^2)}{de_j^2}=\frac{\mu_0}{(1+\mu_0 B)}$$ and

$$(32)^*\frac{dh(e_j^2,\beta^2)}{de_j^2}=\frac{\mu_j}{(1+\mu_j B)}$$

The Implicit Function Theorem states that

$-\frac{\partial F}{\partial x}/\frac{\partial F}{\partial y}=\frac{dy}{dx}$ for F a function of x and y;

Let $\frac{dh_j^2}{de_j^2}=y\,;\,B=x$. Then:

$$F = \frac{dh(e_j^2, \beta^2)}{de_j^2} - \frac{\mu_0}{(1+\mu_0 B)} = 0;$$

$$\frac{d\dfrac{dh(e_j^2, \beta^2)}{de_j^2}}{dB} = \frac{\partial\dfrac{\mu_0}{(1+\mu_0 B)}}{\partial B}$$

$$\frac{d^2\dfrac{dh(e_j^2, \beta^2)}{de_j^2}}{dBd\mu_0} = \frac{\partial\dfrac{\mu_0}{(1+\mu_0 B)}}{\partial B\partial\mu_0}, etc.$$

This means that if I want to understand how the first derivative of health with respect to effort changes in equilibrium, I have to take the partial derivative of the left hand sides of (30)* and (32)* with respect to the parameter of interest.

Now I differentiate $\mu_j/(1+\mu_jB)$ and $\mu_0/(1+\mu_0B)$ with respect to B and μ_j or μ_0, which yields:

$$(33)\quad \begin{array}{l} -\dfrac{2\mu_0}{(1+B\mu_0)^3}; \\ -\dfrac{2\mu_j}{(1+B\mu_j)^3}; \end{array}$$

The result is negative which means that B increases equilibrium effort, but the larger is μ_j or μ_0, the smaller is this effect. In other words, if benevolence increases, effort to the effective type increases more for the regime in which willingness to pay (measured as the negative of μ_0 or μ_j) is higher.

Claim 1. Benevolence increases equilibrium effort for the efficient type always. It does so more for patients and regimes with higher willingness to pay (lower μ's).

Incidentally, something else has been proved. Because the equilibrium conditions for the effective type are exactly those for the first best (full information case), willingness to pay has been shown positively associated with benevolence effects. For the rich, benevolence is more important than for the poor, at least for the first best and for the effective type.

As to the ineffective type (and a second best), the situation is more complicated. It should be, because effort for this type affects both the healthcare for the type and the rent for the effective type.

As above, I rewrite:

$$(34)\ \frac{dh(e_j^1,\beta^1)}{de_j^1}=\frac{q\mu_0+(1-q)\mu_0\dfrac{d\varepsilon}{de_j^1}}{q(1+\mu_0 B)-(1-q)\mu_0 B}$$

$$(35)\ \frac{dh(e_j^1,\beta^1)}{de_j^1}=\frac{q\mu_j+(1-q)\mu_j\dfrac{d\varepsilon}{de_j^1}}{q(1+\mu_j B)-(1-q)\mu_j B}$$

Differentiating the right hand sides of (34) and (35) with respect to B and μ_j or μ_0 yields:

$$(36)\ \begin{aligned}\frac{d(r.h.s(33))}{dB}&=-\frac{\mu_0{}^2(q+(1-q)\dfrac{d\varepsilon}{de_j})(2q-1)}{(q+2q\mu_0 B-\mu_0 B)^2}\\ \frac{d(r.h.s(33))}{dBd\mu_0}&=-\frac{2\mu_0 q(2q-1)(q+(1-q)\dfrac{d\varepsilon}{de_j^1})}{(q+2q\mu_0 B-\mu_0 B)^3}\end{aligned}$$

and a similar expression for the *Laissez-faire* regime. So if the likelihood of the ineffective type is higher than that of the effective type, benevolence increases healthcare for the ineffective type (that is to say, decreases the slope, according to the first equation of (36)).

Further, if and only if

$$(37)\ (q+\mu_0 B(2q-1))(2q-1)$$

is positive, the effect of benevolence decreases with marginal valuation of money. Taking into account (17), the following generalization holds:

Claim 2.1 If the ineffective type is more probable than the more effective type, benevolence increases and otherwise decreases the equilibrium effort for the ineffective type.

Claim 2.2 The said effects are negatively associated with marginal valuation of money if q>1/2.

The substantive significance of all claims will be discussed in the next subsection.

Increasing B increases effort and might thus increase payment, required to compensate the effort. Intuitively, increasing B makes extra effort cheaper, so the State orders more of it. Certainly, in both regimes, the utility to the State increases with B. But what regime benefits most from changes in B? So far, the analysis has been carried out solely with regard to the effects of benevolence on equilibrium efforts. This is not sufficient for understanding regime choice effects, for (27) depends also on the payments in the regime where the State pays, and does not depend on them – which is even more important – when patients pay.

With regard to the *State Pays* regime (π=1), one can apply the envelope theorem saying that the total derivative of a utility function at maximum with respect to a parameter is equal to the partial derivative. This means that in order to see the total effect of benevolence on the State's utility in the *State Pays* regime (π=1), one obtains the partial derivative of (28) with respect to B under the sole condition π=1:

$$(38)\quad \begin{aligned} &\frac{dW(\pi=1)}{dB} = (1-q)\mu_0 h_j^2 + (2q-1)\mu_0 h_j^1 > 0 \\ &\frac{dW(\pi=1)}{dBd\mu_0} = (1-q)[h_j^2 + \frac{\partial h_j^2}{\partial \mu_0}] + (2q-1)[h_j^1 + \frac{\partial h_j^1}{\partial \mu_0}] \end{aligned}$$

So, the State's welfare in the *State Pays* regime is positively associated with benevolence. However, the same approach does not apply to the *Laissez-faire* regime π=0, because in that regime, the maximized utility does not enter completely into State welfare (28). So, I will have to obtain the full derivative of State welfare for the *Laissez-faire* regime (π=0):

$$(39)\ \frac{dW(\pi=0)}{dB} = (1-q)\frac{dh_j^2}{dB} + q\frac{dh_j^1}{dB}$$

(39) is positive, but the sign of the derivative of (39) with respect to the marginal valuations of money by patients is not clear. Comparison between regimes on the basis of (38) and (39) and further derivatives is difficult, because one has to compare complex symbolic expressions. I shall analyze an interesting particular case instead, using a specific functional form. The functional form is the logarithmic (3), as above. The utility accruing to the State is calculated and then differentiated with respect to the parameters of interests for each regime separately and then results are compared.

If π=1 (the *State Pays* regime):

$$\begin{aligned} W(\pi=1) &= (1-q)(1+\mu_0 B)\ln[\frac{\beta^2(1+\mu_0 B)}{\mu_0}] - (1-q)\mu_0\{\frac{(1+\mu_0 B)}{\mu_0} - \frac{1}{\beta^2}\} \\ (40)\quad &+ (q + 2q\mu_0 B - \mu_0 B)\ln\frac{\beta^1\beta^2 q[1+2\mu_0 B - \frac{\mu_0 B}{q}]}{\mu_0(\beta^2 - \beta^1(1-q))} \\ &- (\frac{\beta^1}{\beta^2}(1-q_1) + q_1)\mu_0(\frac{\beta^2 q_1[1+2\mu_0 B - \frac{\mu_0 B}{q_1}]}{\mu_0(\beta^2 - \beta^1(1-q))} - \frac{1}{\beta^1}) \end{aligned}$$

The expected transfer increases in the amount of rent due for the high effectiveness type. Therefore, it increases – as can be seen directly from the formula – in the probability of the high effectiveness type (1- q). Rent also increases in the difference between effort effectivenesses of two types.

If π=0 (the *Laissez-faire* regime):

$$(41)\quad W(\pi=0)=(1-q)\ln[\frac{\beta^2(1+\mu_1 B)}{\mu_1}]+q_1\ln\frac{\beta^1\beta^2 q[1+2\mu_1 B-\frac{\mu_1 B}{q}]}{\mu_1(\beta^2-\beta^1(1-q))}+$$
$$(1-q)\ln[\frac{\beta^2(1+\mu_2 B)}{\mu_2}]+q\ln\frac{\beta^1\beta^2 q[1+2\mu_2 B-\frac{\mu_2 B}{q}]}{\mu_2(\beta^2-\beta^1(1-q))}$$

Now the State's utility does not depend on the amount of rent directly, but can depend on it indirectly, through the health statuses attained by both patients. If patients are poor, they may be willing to sacrifice a lot of health status when they are effort-ineffective in order to decrease the rent due when they are effort-effective. For example, a poor patient may offer a contract according to which no medical help is required unless the patient is in a very bad condition and effectiveness of effort is very high.

Equating (40) and (41) gives the set of parameter values for which the State is indifferent between two regimes. The first task is to identify the relation between marginal valuations of money by the State and patients that make the State indifferent between the two regimes. Yet (40) and (41) are still two complex symbolic expressions and a simplification is in order. Suppose that B=0. I shall examine the effects of benevolence at the zero benevolence point. Under this assumption (40)=(41) simplifies to:

$$2(1-q)\ln\frac{\beta^2}{\mu_0}-2(1-q)\mu_0\{\frac{1}{\mu_0}-\frac{1}{\beta^2}\}+2q\ln\frac{\beta^1\beta^2 q_1}{\mu_0(\beta^2-\beta^1(1-q))}$$
$$(42)\quad -2(\frac{\beta^1}{\beta^2}(1-q)+q)\mu_0(\frac{\beta^2 q_1}{\mu_0(\beta^2-\beta^1(1-q))}-\frac{1}{\beta^1})=$$
$$(1-q)\ln\frac{(\beta^2)^2}{\mu_1\mu_2}+q\ln\frac{(\beta^1\beta^2 q_1)^2}{\mu_1\mu_2(\beta^2-\beta^1(1-q))^2}$$

or:

$$(43)\ln\frac{\mu_1\mu_2}{\mu_0{}^2}=K$$

where K is a constant equal to the expected transfer the State pays.

When the geometric average of marginal value of money to patients is equal to that of State, the latter strongly prefers to delegate incentive and effort-compensating payments to patients, unless those payments are zero. Increase in the willingness to pay on the part of patients, that is to say, decrease in μ$_j$'s, so that their geometric average exceeds that of the State itself, means that the State prefers the *State Pays* regime, even though rent might be zero or even negative.

The full derivatives of State welfare with respect to B are:

$$(44)\quad \frac{dW(\pi=1)}{dB} = 2(1-q)\mu_0 \ln[\frac{\beta^2(1+\mu_0 B)}{\mu_0}] + 2\mu_0(2q-1)\ln\frac{\beta^1\beta^2 q(1+2\mu_0 B-\frac{\mu_0 B}{q})}{\mu_0(\beta^2-\beta^1(1-q))}$$

where the factor 2 comes from there being two patients;

and

$$(45\ \frac{dW(\pi=0)}{dB} = (1-q)\frac{\mu_1}{(1+\mu_1 B)} + q\frac{(2\mu_1-\frac{\mu_1}{q})}{[1+2\mu_1 B-\frac{\mu_1 B}{q}]} + (1-q)\frac{\mu_2}{(1+\mu_2 B)} + q\frac{(2\mu_2-\frac{\mu_2}{q})}{[1+2\mu_2 B-\frac{\mu_2 B}{q}]}$$

Claim 3. Suppose that the State is initially indifferent between the two regimes. Then B starts to increase. The following defines the resulting effect on regime choice:

$$\frac{dW(\pi=1)}{dB} > \frac{dW(\pi=0)}{dB} if$$

$$(46)\quad \ln[\frac{\beta^2(1+\mu_0 B)}{\mu_0}] > \frac{\mu_1}{2\mu_0(1+\mu_1 B)} + \frac{\mu_2}{2\mu_0(1+\mu_2 B)};$$

$$\frac{dW(\pi=1)}{dB} < \frac{dW(\pi=0)}{dB} if$$

$$\ln\frac{\beta^1\beta^2 q(1+2\mu_0 B-\frac{\mu_0 B}{q})}{\mu_0(\beta^2-\beta^1(1-q))} < \frac{(2q\mu_1-\mu_1)}{2\mu_0(2q-1)[1+2\mu_1 B-\frac{\mu_1 B}{q}]} + \frac{(2q\mu_2-\mu_2)}{2\mu_0(2q-1)[1+2\mu_2 B-\frac{\mu_{21} B}{q}]}$$

This result can be recast in words. Regime choice depends on benevolence through equilibrium health outcomes, willingness to pay being fixed as follows: if health outcomes are sufficiently good, the State considers the increase in benevolence as a reason to prefer to pay for care. If the equilibrium outcomes are not very good, because patients are not effort-effective enough, the State may think benevolence to be a reason to choose the *Laissez-faire* regime. Alternatively, low willingness to pay among patients may induce State prefer the *State Pays* regime.

As a specialization of this result, consider the case *B*=0. Differentiating thus modified (43) and (44) with respect to the marginal valuations of money, I obtain:

$$\text{(47)}\quad \begin{aligned} &\frac{dW(\pi=1)}{dB}\Big| B=0 = 2(1-q)\mu_0 \ln\frac{\beta^2}{\mu_0} + 2(2q\mu-\mu_0)\ln\frac{\beta^1\beta^2 q}{\mu_0(\beta^2-\beta^1(1-q))} \\ &\frac{dW(\pi=0)}{dB}\Big| B=0 = (1-q)\mu_1 + q(2\mu_1-\frac{\mu_1}{q}) + (1-q_1)\mu_2 + q(2\mu_2-\frac{\mu_2}{q}) \end{aligned}$$

A special case of Claim 3. Increasing benevolence B increases welfare of State faster in regime when State pays than in the regime where patients pay, if

$$\textit{(48)}\ \ln\frac{\beta^1\beta^2 q_1}{\mu_0(q\beta^2+\beta^1(1-q)))} > \frac{(\mu_1+\mu_2)}{2\mu_0};$$

and slower if

$$\textit{(49)}\ \ln\frac{\beta^2}{\mu_0} < \frac{(\mu_1+\mu_2)}{2\mu_0}$$

The model therefore implies a causal chain, which links the choice of private financing of care to professional benevolence. Though professional benevolence per se does not necessarily favor the *Laissez-faire* regime, it does so under high patient readiness to pay in certain diagnostic groups or in circumstances of low effort-effectiveness.

How does inequality affect regime choice? Inequality is a complex phenomenon that has many alternative quantitative representations. For the sake of simplicity, I suggest considering inequality as deviation of marginal valuation of money from an arithmetic average. I rewrite:

$\mu_1 = \mu_{average} + d$

$\mu_2 = \mu_{average} - d$

where $d>0$ is the inequality parameter.

For the general first order conditions this rewriting yields:

$$(50)\ \frac{dh(e_2^2, \beta^2)}{de_2^2} = \frac{\mu_{average} - d}{(1 + (\mu_{average} - d)B)};$$

$$(51)\ \frac{dh(e_1^2, \beta^2)}{de_1^2} = \frac{\mu_{average} + d}{(1 + (\mu_{average} + d)B)};$$

$$(52)\ \frac{dh(e_2^1, \beta^1)}{de_2^1} = \frac{q(\mu_{average} + d) + (1-q)(\mu_{average} + d)\frac{d\varepsilon}{de_2^1}}{q(1 + (\mu_{average} + d)B) - (1-q)(\mu_{average} + d)B)};$$

$$(53)\ \frac{dh(e_1^1, \beta^1)}{de_1^1} = \frac{q(\mu_{average} - d) + (1-q)(\mu_{average} - d)\frac{d\varepsilon}{de_1^1}}{q(1 + (\mu_{average} - d)B) - (1-q)(\mu_{average} - d)B};$$

Differentiating the right hand sides with respect to *d*, I obtain:

For the poor ineffective patient a positive derivative (slope increases, effort decreases with *d*):

$$(54)\ (q + (1-q)\frac{d\varepsilon}{de_j^1})\frac{q}{(q + 2qB(\mu + d) - B(\mu + d))^2};$$

For the rich ineffective patient a negative derivative (slope decreases, effort increases with *d*):

$$(55)\ -(q + (1-q)\frac{d\varepsilon}{de_j^1})\frac{q}{(q + 2qB(\mu - d) - B(\mu - d))^2};$$

For the poor effective patient a positive derivative (slope increases, effort decreases with *d*):

(56) $\frac{1}{(1+B(\mu+d))^2}$;

For the rich effective patient a negative derivative (slope decreases, effort increases with *d*):

(57) $-\frac{1}{(1+B(\mu-d))^2}$;

Notice that absolute values of these derivatives are ranked as follows (taking into account the assumptions made to fulfil the second order conditions):

(57)>(56)

(55)>(54)

under (17).

With the simple functional form (19), the State's welfare is rewritten as a function of *d*:

$$(58)\quad \begin{aligned} &W(\pi=0)=(1-q)\ln[\frac{\beta^2(1+(\mu+d)\ B)}{(\mu+d)}+\\ &q\ln\frac{\beta^1\beta^2 q_1[1+2(\mu+d)B-\frac{(\mu+d)B}{q}]}{(\mu+d)(\beta^2-\beta^1(1-q))}\}+\\ &(1-q)\ln[\frac{\beta^2(1+(\mu-d)B)}{(\mu-d)}]+\\ &q\ln\frac{\beta^1\beta^2 q[1+2(\mu-d)B-\frac{(\mu-d)B}{q}]}{(\mu-d)(\beta^2-\beta^1(1-q))} \end{aligned}$$

The components of (58) which correspond to the ineffective types have the following derivatives with respect to d:

(59) $\frac{q^2}{(q+2qB(\mu-d)-B(\mu-d))(m-d)}$

for the rich patient and

(60) $-\frac{q^2}{(q+2qB(\mu+d)-B(\mu+d))(m+d)}$

for the poor patient. The absolute value of (59) is large than the absolute value of (60) for q>1/2. This means that (58) increases with respect to d, in other words, the State's welfare in the *Laissez-faire* regime increases with inequality.

Claim 4. For q>1/2, if the arithmetic mean of marginal valuation of money stays the same, while difference among patients increases, the thus measured inequality increases equilibrium efforts for the rich patient faster than for the poor patient and therefore increases State preference for the Laissez-faire regime.

One can also further this analysis of inequality effects by checking what happens if the geometric mean of willingness to pay is fixed. Certainly, the additive inequality is immediately amenable to a utilitarian interpretation. This new multiplicative version is much less so, if at all. Yet a calculation with a fixed geometric mean would still demonstrate the role of the probability of the ineffective type in fixed the identified effect on the State's welfare.

Here is the calculation. I rewrite:

$\mu_1 = r\mu^*$

$\mu_2 = \mu^*/r$

where r is a parameter > 1 and μ^* is geometric average willingness to pay..

For the general first order conditions this rewriting yields:

$$(61)\ \frac{dh(e_2^2, \beta^2)}{de_2^2} = \frac{\frac{\mu^*}{r}}{(1 + \frac{\mu^*}{r}B)}$$

$$(62)\ \frac{dh(e_1^2, \beta^2)}{de_1^2} = \frac{r\mu^*}{(1 + r\mu^* B)}$$

(52) $$\frac{dh(e_1^1,\beta^1)}{de_1^1}=\frac{qr\mu^*+(1-q)r\mu^*\frac{d\varepsilon}{de_j^1}}{qr\mu^*B-(1-q)r\mu^*B}$$

(63) $$\frac{dh(e_2^1,\beta^1)}{de_2^1}=\frac{q\frac{\mu^*}{r}+(1-q)\frac{\mu^*}{r}\frac{d\varepsilon}{de_j^1}}{q(1+\frac{\mu^*}{r}B)-(1-q)\frac{\mu^*}{r}B}$$

Differentiating the right hand sides with respect to *r*, one obtains:

For the poor ineffective patient a positive derivative (slope increases, effort decreases with *r*):

(64) $$(q+(1-q)\frac{d\varepsilon}{de_j^1})\frac{q}{(q+2qBr\mu^*-Br\mu^*)^2}$$

For the rich ineffective patient a negative derivative (slope decreases, effort increases with d):

(65) $$-(q+(1-q)\frac{d\varepsilon}{de_j^1})\frac{q}{(rq+2qB\mu^*-B\mu^*)^2}$$

For the poor effective patient a positive derivative (slope increases, effort decreases with d):

(66) $$\frac{1}{(1+rB\mu^*)^2}$$

For the rich effective patient a negative derivative (slope decreases, effort increases with d):

(67) $$-\frac{1}{(r+B\mu^*)^2}$$

(66) and (67) are (65) and (64) in the limit *q*=1. Absolute values are ranked as follows:

(67)>(66) if

(68) $B\mu_{average}>1$

(65)>(64) if

$$(69)\ q > \frac{B\mu_{average}}{2B\mu_{average} - 1}$$

Condition (17) applies. If (68) and (69) fulfil, inequality increases the State's utility in the *Laissez-faire* regime. Two conditions are consistent: if q=1, (69) is implied by (68). It is worthwhile to see what these conditions mean.

Fixing geometric average and increasing inequality means increasing the arithmetic average. Conditions (68) and (69) ensure that the total equilibrium effort still increases with increasing inequality, diminished 'purchasing power' notwithstanding.

It remains to be checked whether the change of parameters as in Remark 1 preserves the results so far:

$$\mu_0{}^{W} = \mu_0{}^{SW} - 1$$

and

$$B^{W} = B^{SW}\frac{(\mu_0{}^{W} + 1)}{\mu_0{}^{W}}$$

I must prove that the results do not change if I assume that the State is benevolent (maximizes the utilitarian social welfare function) and two parameters change as above. Box 5-1 proved that the equilibrium conditions do not change if I make these two changes. Therefore, the first order conditions will be the same, only with new parameters.

$$(28)^{*}\ q(1+(\mu_0+1)B\frac{\mu_0}{\mu_0+1})\frac{dh(e_j^1,\beta^1)}{de_j^1} - q(\mu_0+) - (1-q)(\mu_0+1)[\frac{d\varepsilon}{de_j^1} + B\frac{\mu_0}{\mu_0+1}\frac{dh(e_j^1,\beta^1)}{de_j^1}] = 0$$

$$(29)^{*}\ (1+(\mu_0+1)B\frac{\mu_0}{\mu_0+1})\frac{dh(e_j^2,\beta^2)}{de_j^2} - (\mu_0+1) = 0$$

The most elementary algebra shows that (28)* and (29)* are equivalent to (28) and (29). In other words, if I substitute the new parameters in all formulae in this section, the same results emerge.

More generally, expression $\mu_0 B$ is invariant with respect to the simultaneous transformations:

$$\mu_0{}^{W} = \mu_0{}^{SW} - 1$$

and

$$B^{W} = B^{SW} \frac{(\mu_0{}^{W} + 1)}{\mu_0{}^{W}}$$

and benevolence parameter *B* is always accompanied by the parameter of marginal valuation of money, which means that the two transformations must always be applied together to any expression containing *B*.

5.5 Interpretations for Regime Choice

A model of the choice between two regimes has been constructed. To recap the main points:

- The *State Pays* regime means public funding of proper incentives for medical service provider under uncertainty. The State incurs expenses valued at monetary value of transfers to provider multiplied by some parameter. This parameter may mean the shadow cost of public financing or the opportunity cost of one unit of money spent on healthcare or possibly something else.
- In the *Laissez-faire* regime, the State does not pay incentive payments to the Provider. Instead, patients pay the rent according to their willingness to pay, given the expected effort effectiveness.

The model demonstrates the consistency of the idea that administrative and financial reasons may force state institutions to switch the burden of financing healthcare to patients. It does so in the context of effort-rent trade-off, which I claim to be extendable to a continuous setting (see Appendix B.1).

Secondly, a constraint on total effort by provider is introduced. If this constraint does not bind, the model implies only additive payments serving incentive correction. If the constraint does bind distributive payments emerge. Finally, the model attempts to reflect some effects of the Provider's benevolence. One consequence of benevolence is that one person pays for something another receives for free, and this is the case of supply-irrelevant payments.

In a wider context, the model formalizes the following causal link. A state (or a community) decides to minimize its efforts and expenses in creating a free-for-all public healthcare system, when such efforts and expenses do not pay out. Schematically, the conjunction of reasons for a state or a community to allow patients to arrange care for themselves is as follows:

1. The State is interested in health outcomes of patients;
2. Informational asymmetry makes optimal effort expensive;
3. The financial burden on patients does not enter the State's utility function.

When does the state decide to relieve itself of the burden of healthcare arrangement? This question was answered formally in previous subsection. The results will now be discussed in the order that I perceive to be that of importance.

Claim 4. Inequality increases State preference for the Laissez-faire regime.

The idea of increased inequality (actual or perceived) increasing the probability of the State's refusal to finance care has a certain intuitive appeal. Though the general situation with healthcare in Russia does not strongly confirm this claim, the latter sits well with the current policy tendencies in Russia. A state refuses to maintain a 'free for all' scheme that would enforce distribution of scarce resources by need and would include incentives for medical professionals. This could potentially be related to the inequality that is widely believed to have increased recently. Inequality makes all forms of cross-subsidization less acceptable to society (or, more precisely to its rich part), because there is both little solidarity and little economic incentive for the rich to subsidize the poor.

The model implies one peculiar channel for this effect. In this model, the State prefers the *Laissez-faire* regime when inequality is high, because, metaphorically speaking, the poor will pay under the duress of bad health, while the rich will pay because they can afford to. Rising inequality increases the average private willingness to pay: the poor still pay almost as much as before, while the rich pay substantially more.

In general, this claim is related to the assumption of a State concerned with provision of healthcare, but not with the burden it implies on private households, which are the reasons #1 and 3 above.

Claim 3. If medical efforts are sufficiently effective in terms of producing good health outcomes (per 'unit effort'), then the benevolence of doctors increases the State's willingness to arrange financing and pay the rent, that is to say, to prefer the State Pays regime. If medical efforts are sufficiently ineffective, then benevolence induces State to prefer the Laissez-faire regime.

Russian healthcare must be considered rather inefficient. There is overstaffing, equipment is obsolete, hospitals are too large and too many, at the expense of out-patient and preventive care. There are apparently interests vested in these sources of inefficiency. The model implies a link between low effectiveness of medical effort, whatever the underlying causes, and effects of professional benevolence. Benevolence increases health outcome and saves money. Which of the two effects prevails as far as regime choice is concerned depends on effectiveness of medical effort. The lower effectiveness, the less effect benevolence has in terms of saving money relative to that of improving health outcomes. In the *Laissez-faire* regime, the State cares only about health outcomes. This explains why low effectiveness implies that benevolence helps the *Laissez-faire* regime get chosen.

Claims 1-2. Part I. Benevolence increases equilibrium effort for the effective type and for a sufficiently likely ineffective type.

The fact that benevolence may decrease the equilibrium effort for the ineffective type is important, since rent is affected by benevolence. There are two effects of benevolence on effort for the ineffective type. Benevolence increases health status through increase in effort for the ineffective type. But also, benevolence decreases rent for the effective type, through increasing the difference in equilibrium efforts. If the probability of the ineffective type is large, then the first effect is significant. If the ineffective type is less probable than the effective type, the second effect is significant.

Claims 1-2, Part II. Marginal valuation of money decreases the effects of benevolence on the optimal efforts for a sufficiently probable ineffective type, for the effective type and in the first best.

In Chapter 3, benevolence played a multifarious role. It allowed patients to receive something for free or cheaply. Besides, it increased the amount of care provided to the paying patient, exactly because the patient paid. This might happen at the expense of another, non-paying patient. The second link between benevolence and payment was explained in terms of Provider honoring an implicit contract, based on trust. The second factor made benevolence effects more pronounced for those ready to pay.

No doubt, links between optimal effort and benevolence are highly model-dependent. In this model, an inverse relation between marginal valuation of money and strength of benevolence effects is usually negative. Second-best efforts entail a positive link between these two only when benevolence decreases the effort for the ineffective type. It appears that a situation when benevolence works in favor of the rich more than it does in favor of the poor is not so counter-intuitive as it may sound.

Suppose that there is a poor citizen coming for an appendectomy, which is urgent and necessary and this person can only choose between no operation and an operation by a doctor who is not particularly famous or skillful. The operation will happen anyway, whether the doctor is benevolent or not, because a patient will simply die otherwise. Benevolence does not have an effect.

Then there is a rich person requiring an appendectomy, choosing from a famous surgeon, a less famous surgeon, and a regular surgeon. Here, if the effort by a more famous surgeon is valued higher than that by a less famous one, the equilibrium effort will be higher than for the poor person and benevolence may have an effect. For example, if doctors in general are not benevolent, the person in question may choose the average doctor, while if doctors are benevolent, which lowers the price, the person may choose the best doctor. The outcome with benevolence is different from the outcome without benevolence. Such is the overall picture. All the claims have one underlying assumption: the State does not care about finances of the patients, only of its own. The eventual choice of the Laissez-faire regime happens whenever:

1. medical efforts are sufficiently ineffective;
2. patients exhibit sufficiently large inequality and there are enough rich patients;
3. the Provider is sufficiently benevolent.

All these three features appear to be present in today's Russian society. Inequality prevails according to standard statistical estimates. Health care provision demonstrates provider benevolence in various forms as a stable feature to be always taken into account. Finally, the healthcare system in general, as is acknowledged even officially, is inefficient. This makes the model realistic, if inevitably simplistic, as an explanation of why the Russian state, and eventually, society, has chosen to allow payments for services free by law.

5.6 Objections and Further Developments

First of all I propose to review some prospects for further development eventually based on a single abstract consideration. The model gives a lot of power to the principal and little to the agent. One distinct alternative is to recast the formalism so as to correct this counter-intuitive imbalance. The general way to do it is to switch from optimal contract theory to bargaining with two-sided incomplete information. Here are a few considerations that justify moving this direction.

The model was intended to explain when a state prefers to tolerate payments for services free by law, which constitute illegal rent of providers of care. In the model the illegal rent is a way to avoid paying a legal rent. There is an alternative, simpler model: the State simply fails to construct an incentive-correcting mechanism. One reason could be that patients can observe health status, while the State cannot. Illegal rent is the only way to increase effort through correction of adverse incentives. In this sense, the above model may be giving too much power to the state relative to that of providers of care. As an alternative Provider can be endowed with a larger strategy space within a bargaining game with two-sided incomplete information.

A related feature of the model is that State or patients successfully violate professional autonomy of providers by tightly linking performance and reward. The extant administrative capacity and the actual patient ability to bargain can be thought insufficient for such a mechanism to emerge. As one counter-objection, payments for services free by law in many instances do constitute such a high-powered incentive scheme. Furthermore, rewards for good performance may be embedded in career progress with increasing both pay and status with respective perks. A bargaining game with two-sided incomplete information may however better reflect the actual power of the profession.

Yet another feature of the model is that a profession less powerful than normally perceived. The model does not reflect the role of private payments for containing lobbying pressure from the medical profession, which was one speculative explanation of existence of rent in form of private payments (Sections 4.3). Further development of the model should necessarily take care of this aspect.

Instead of facing a one-shot contract menu from the principal, the profession can act as a bargaining partner, offering various levels of cooperation with governmental demands in return for varying rewards. This would be the desired game with two-sided incomplete information. The government does not know the effectiveness of a patient. The profession/provider does not know the actual valuation of money for the government.

One further possible alteration is as follows. The State imposes and enforces some minimal effort or health status, which is accorded to everyone free of charge. Then it would be interesting to see how variation of this guaranteed minimum might affect outcomes of the model. The discussion of how to define 'state guarantees' of free care' more precisely is ongoing. Thus altered, the model could demonstrate the welfare consequences of varying the guaranteed minimum.

Under both regimes, the Provider sells its services apiece. This is an essential assumption, which, however, may not always be realistic. A good ward with a professional and diligent team of nurses and doctors and state-of-art equipment could be a limited resource, consumed jointly or rationed among admitted patients according to their need rather than payment. But the access to such a resource is then charged for. Such charges will obviously have distributive effects. Appendix B.2 presents a formal exposition of such a situation.

Chapter 6. Conclusions

In Chapter One I promised to answer three questions:

1. a general question: What is the cause for emergence of a system of payments for services free by law?

and two specific questions:

2. What is the effect, if any, of state policies on this system of payments?
3. What is the nature of relations between medical professionals and patients whom the professionals manage to charge?

In this concluding chapter I review the received wisdom; the facts that the received wisdom does not conform to; my alternative theory; its policy relevant consequences; the outstanding issues, including empirical test of the proposed theory.

Received wisdom. The received wisdom is that payments for services free by law result from inadequate public financing of healthcare. Such payments complement public financing to the effect of maintaining the level of healthcare provision. This is the answer the literature gives to the first of the above questions. As to the second question, the literature takes relations between professionals and patients to be those between sellers and buyers of services. This notion sits well with the idea of under-funding: the official third-party payer system reneges on its obligations and direct compensation of providers through private payments steps in.

Regarding state policies, the absolute level of financing is the only dominant factor in shaping the system of payments for services free by law. The other elements of regulation are not considered as material factors.

I called these three ideas the ‘under-funding paradigm’ (subsection 1.3.3). The under-funding story is not internally inconsistent; yet it appears to neglect the issues of doctor agency that the literature on health economics and policy has sought to keep in focus. One important issue here is the specific principal-agent relations between the doctor and patient that are often considered to involve trust by the patient and professional benevolence by the doctor (subsection 1.3.2). Secondly, the under-funding paradigm does not conform to the known facts.

My main argument against the under-funding paradigm consists in providing an alternative conceptualization of the payments that better suits the known facts and also conforms to the received general theory, which has in turn been argued to reflect many observations about healthcare across countries and times.

Stylized facts against the under-funding paradigm. I have argued that, first of all, statistical data do not indicate that a major collapse of free care provision has occurred. A co-existence of free and paid care is more warranted by evidence. Only a minority pays, and amounts of payment, though sufficient to exercise significant incentive effects on medical personnel, stay significantly below public funding. This concentration of payments is the first sign that the level of public funding may not be an important factor determining private payments (Chapter 3, subsection 3.1.3).

Chapter Two also showed that whatever the level of public funding, the current laws and policies allow the provider to charge the patient for services free by law. The enforcement of free care entitlement is very unreliable, especially taking into account the specific nature of healthcare provision.

Interview-based evidence demonstrates that payments for medical services free by law concentrate in such areas as surgery and long-term care (Chapter Three, section

3.3). The primary example of why a patient may want to pay is choice of doctor or hospital, both being important for quality surgery and continuous care for a chronic patient. Even in these areas, only a minority pays. That is to say, the relevant transactions happen between certain patients and certain doctors or in certain wards and hospitals. Thus, payments concentrate in certain elements of public healthcare. I have called this a "pattern of concentration". There is also no uniformity as regards forcing patients to pay. At least in the reported cases, direct denial of free service is rare.

These stylized facts do not conform to the under-funding paradigm, as the latter predicts general transition of care delivery into a regime of private funding. Following the pattern of concentration, in a majority of transactions, patients do not cover cost of service In these circumstances, a payment is likely to constitute a rent in excess of reservation wage and official salary of a doctor.

Additional reasoning against the under-funding hypothesis is as follows. Healthcare economics has emphasized informational asymmetry between doctors and patients. The areas favored by payments under the pattern of concentration are exactly those where the informational asymmetry is particularly high. It has already been said that the under-funding paradigm ignores agency issues. The argument is only made stronger by considering the actual distribution of payments.

Finally, public health policies do not address the conflict of interest caused by allowing healthcare providers to sell service free by law to those patients who are willing to pay. Much of regulation in this regard is passed over to the lowest levels of administrative hierarchy and the profession itself. In view of large provider power over patients, this must be an important factor in shaping the institution of private payments in public healthcare, overlooked in the under-funding paradigm.

In continuation of this latter point, it seems interesting that the interviewed representatives of insurance companies in Saint Petersburg, having acknowledged the problem of payments for services free by law, have denied the role under-funding. They also do not think that increasing public funding would solve the problem (Appendix A-3).

Alternative conceptualization. Three elements constitute the alternative conceptualization I have proposed and defended:

- private payments are rents in excess of cost of service;
- private payments play signaling role and thereby provide ground for trust between patients and professionals in situations of extreme informational asymmetry;
- private payments for services free by law are effectively permitted by the current regulation, while low financing could hypothetically play the role of partial rent extraction.

In this scheme, payments create incentive for professionals to work more or better or else redistribute scarce resources and effort from the non-paying to the paying patients. These effects and their relation to the notion of rent are discussed in detail in Chapter Four. Formal validation of some arguments there is elaborated in Chapter Five.

Payments of interest concentrate in areas, where the patient has very little control over what extra quality or quantity of service is delivered for the payment. The theory says that in such circumstances a combination of external control and professional benevolence may provide rationalization of paying a doctor. In the particular case of Russian healthcare, the external control is rather limited (Chapter

Two). The trust is created through the specific role payments play (section 3.4.5). I repeat here a 'master' definition from section 3.4:

By ingratiation, I understand an object of payment that is different from buying a specific service and is normally additional to the latter. Ingratiation means that by paying, a patient aims to provide incentive for good work, but lacks any ostensible means of enforcing the implicit or even explicit contract.

I suggested two underlying mechanisms that rationalize emergence of such special relationships due to signaling role of payment:

1. reputation effects: a career doctor or nurse cares about his or her reputation exactly among those who are able to pay, because inducing trust in them creates a clientele through 'word of mouth';
2. true benevolence: a doctor or nurse shows professional benevolence more readily for those who in turn recognize professional effort and show respect by paying.

These scenarios have variously been outlined in many interviews. Their relative weight is impossible to assess. What is possible to assess, however, is the empirical validity of the pattern of concentration in connection with the special relationship construction (subsection 3.4.5). If willingness to pay correlates with asymmetry of information or other aspects of provider power over the patient, as the pattern of concentration implies, the signaling role of payments becomes at least indirectly confirmed. If there is no such correlation, the hypothesis of special relationships is falsified at least in the present form.

In view of the nature of such payments, the following hypothesis suggests itself: public financing not covering cost of service can be seen as caused by private payments and not causing them. The very existence of payments aimed at ensuring quality service does not depend on absolute level of public financing. But in presence of such payments, the state may engage in effective rent extraction by financing public healthcare at a level that per se insufficient to cover cost of service. State policies include ineffective regulation of paid service delivery and tolerance towards payments for services free by law. This observation sits well with the idea of rent extraction.

Chapter Five formalizes state policies in terms of regime choice. Under certain conditions, the state chooses a regime of allowing private payments and thus saves public money while maintaining a desirable level of healthcare provision. Implicit in the model is irrelevance of under-funding in causing private payments. The model also implies that inequality is a factor that enhances the utility to the state from allowing private payments.

Policy relevant insights. It is rather unlikely that changing the amount of public money available for salaries can eliminate private payments. It can only make those of them that are currently extracted rent into non-extracted rent as argued in section 4.3. Therefore a different approach is needed to directly address the conflict of interest that healthcare providers face. It is a question of regulating professional agency and governance rather than that of public financing schemes. The history of healthcare politics (see section 1.3) suggests in manifold ways that neglecting professional agency can invalidate efforts in public procurement of healthcare.

The insights of the research appear to support the idea that deregulating public healthcare towards more private choice and competition may be useful, if this leads to transparency and patient interest representation. Deregulation could be desirable for its potential in reducing the informational asymmetry or some of its effects. More provider autonomy, wider integration of private and mandatory insurance and more patient choice could be steps in this direction.

Outstanding research problems/hypothesis test. One outstanding issue is a test of the pattern of concentration. If positive it would provide support for the hypothesis of payments as rents. If extended with further measurements (those of perceptions, conscious motivations, circumstances attending a payment), such a test could be turned into a proper (in)validation of the hypothesis of special relationships and payments triggering trust.

Another outstanding issue is as follows. Existence of payments for services free by law and, more generally, co-existence of paid and free care greatly depends on professional freedom. By the latter I understand professional discretion to choose the means of treatment for a given patient. Some of such freedom is based on scientific uncertainty, which justifies variation in the service different professionals and institutional providers accord to patients with similar conditions. Yet the rest of professional freedom could be attributed to institutional arrangements and personal motivations. To understand further how free and paid care combines, it is important to measure both the total extent of professional freedom and its part due to pecuniary incentive differentially affecting agents.

Appendix A. Statistical data and semi-structured interviews

A1. Boston University Surveys

This appendix provides information on the household expenditure surveys conducted by the Institute for Social Studies, Moscow for the University of Boston. With the agreement of the University, the Institute let me use the primary data.

The following displays characteristics of the samples, as well as means and deviations for a number of variables, constructed from the primary data. All the estimations and calculation of weights were made independently by the author of this Dissertation and slightly deviate from the officially published results (Boikov et al. 1998, 2000a,b). The differences were due to the use of weights.

Table A.1-1. Household healthcare expenditure surveys of 1997 and 1998

#	Region	Sample 1997, h/holds	Sample 1998	Population, Thousand	Weights 1997	Weights 1998
1	Republic of Karelia	120	88	780	2407,407	3282,828
2	Saint Petersburg	98	72	4779	17478,6	23790,32
3	Leningrad *oblast*	82	62	1679	7697,598	10180,69
4	Moscow	175	129	8639	18016,68	24441,24
5	Tula *oblast*	215	158	1801	3209,481	4367,331
6	Orel *oblast*	216	158	910	1526,436	2086,773
7	Nizhnii Novgorod *oblast*	174	126	3711	7928,471	10948,84
8	Voronezh *oblast*	159	118	2499	5864,545	7902,226
9	Volgograd *oblast*	357	253	2703	2684,904	3788,58
10	Stavropol *oblast*	360	253	2674	2419,472	3442,726
11	Kurgan *oblast*	400	306	1107	1002,717	1310,742
12	Tumen *oblast*	304	227	3181	3442,045	4609,61
13	Krasnoyarsk *krai*	190	137	3095	5675,775	7871,513
14	Khabarovsk *krai*	150	113	1557	3530,612	4686,653

All payments were classified into formal and informal, and also into the following categories:

- “for services” – if payment is made for some service (diagnosis or treatment) or hospitalization as a whole;

- “for medication”;
- “to doctor”;
- “to nurse”;
- “for hotel services”;
- “for laboratory tests”
- “others”.

The category of “hotel services” was excluded from estimations here, for though this can include illegal charges and those for services that may be free by law, the charges will not be for medical services proper. The rest was included in the estimation of total amounts.

Table A.1-2. Variables

Code	Explanation
INTOTAL	Total amount paid by the household for in-patient care and medication in hospital (billions of current rubles)
INTOTUN	Total amount of informal payments made by the household for in-patient care and medication in hospital (billions of current rubles)
OUTTOTAL	Total amount paid by household for out-patient care and medication in hospital (billions of current rubles)
OUTTOTUN	Total amount of informal payments made by household for out-patient care and medication in policlinic (billions of current rubles)
INSER	Amount paid by households for service or hospitalization as a whole (bn rubles)
INMED	Amount paid for medication in hospital (billions of current rubles)
OUTSER	Amount paid for service in policlinic (billions of current rubles)
OUTMED	Amount paid for medication in policlinic (billions of current rubles)
INTOTAL1	Whether paid anything in hospital (0=No or 1=Yes)
INCHI1	Whether paid anything in hospital for a child (0=No or 1=Yes)
INWOR1	Whether paid anything in hospital for a working age person (0=No or 1=Yes)
INPEN1	Whether paid anything in hospital for a pensioner (0=No or 1=Yes)

Note: 1.Private clinics are excluded. 2. On January 1 1998 the new ruble was introduced that was in 1:1000 proportion to the old ruble. The new ruble is used in this Appendix.

Table A.1-3 Estimation of totals, 1997 survey

Variable	Total, bn rubles.	Std. Err., bn rubles	95% Conf. Interval		Deff
INTOTAL	5.286	1.07	3.19	7.38	1.562522
INTOTUN	1.540	3.88	7.79	2.30	1.109667
OUTTOTAL	2.131	2.07	1.72	2.54	1.217725
OUTTOTUN	5.937	1.13	3.72	8.15	1.146202
INSER	2.032	8.31	4.03	3.66	1.489157

INMED	1.330	2.06	9.26	1.73	1.258104
OUTSER	4.627	8.07	3.04	6.21	1.462821
OUTMED	8.536	1.12	6.35	1.07	1.300621

Table A.1-4 Estimation of totals, 1998 survey

Variable	Total, bn rubles	Std. Err.	95% Conf. Interval		Deff
INTOTAL	5.662	0.837	4.02	7.30	1.302136
INTOTUN	1.637	0.536	5.85	2.69	1.512652
OUTTOTAL	3.360	4.29	2.52	4.20	1.723756
OUTTOTUN	0.738	2.23	3.01	-1.18	3.037486
INSER	1.447	3.82	6.98	2.20	.9794745
INMED	1.828	2.82	1.28	2.38	.8933288
OUTSER	0.918	2.25	4.77	1.36	1.150372
OUTMED	0.977	1.80	6.24	1.33	1.237279

Note: 1. DEFF is STATA's measure of weighting effects.

Table A.1-5
Estimation of mean incidence of payment for in-patient care (INTOTAL1, INMED1, INSER1, etc.) by income quartiles, for years 1997 and 1998 together.

Variables	Estimate	Std Err	95% Conf Interval		Deff
Total sample					
NTOTAL1	.3352344	.0174705	.3009572	.3695115	1.592836
INCHI1	.0456883	.0064756	.0329832	.0583935	1.118522
INWOR1	.2238587	.0145844	.1952441	.2524734	1.423777
INPEN1	.4631935	.019711	.4245204	.5018666	1.817259
Lowest income quartile					
INTOTAL1	.3056206	.032403	.2418421	.3693991	1.414998
INCHI1	.0672497	.0155936	.0365569	.0979425	1.108677
INWOR1	.1922712	.0257969	.1414954	.243047	1.225518
INPEN1	.4194886	.0384121	.3438822	.4950949	1.732893
Second lowest income quartile					
INTOTAL1	.2724659	.0290893	.2152374	.3296943	1.378811
INCHI1	.0518745	.0131468	.0260102	.0777387	1.135079
INWOR1	.1639363	.0229423	.118801	.2090716	1.240406
INPEN1	.45673	.0376075	.3827435	.5307165	1.841092
Second highest income quartile					
INTOTAL1	.3387691	.0345567	.270764	.4067741	1.593961
INCHI1	.0317053	.0102706	.0114935	.051917	1.027356
INWOR1	.2102167	.0260185	.1590142	.2614192	1.219155
INPEN1	.4837653	.0388202	.4073699	.5601607	1.804282
Highest income quartile					
INTOTAL1	.443498	.0424314	.3599246	.5270714	1.801824
INCHI1	.0285688	.0114992	.0059199	.0512177	1.17686
INWOR1	.3479314	.0398438	.2694545	.4264084	1.72835
INPEN1	.4890876	.042911	.4045695	.5736057	1.820123

Table A.1-6. Estimation of mean incidence of payment in out-patient care by income quartiles (1997 and 1998).

	Estimate	Std. Err.	95% Conf. Interval		Deff.
All income groups					

OUTTOTAL1	.1370118	.0066081	.1240568	.1499668	1.78194
Lowest income quartile					
OUTTOTAL1	.1046341	.011528	.0820167	.1272516	1.683767
Second lowest income quartile					
OUTTOTAL1	.1309246	.0126846	.1060384	.1558108	1.715288
Second highest income quartile					
OUTTOTAL1	.146806	.0134767	.1203658	.1732462	1.763236
Highest income quartile					
OUTTOTAL1	.1671055	.0150272	.1376226	.1965885	1.924256

A2. Russian Longitudinal Monitoring Study (RLMS)

Here I present frequencies and amounts of payments as measured in RLMS. All details of the surveys are available at http://www.unc.edu/rlms. Rubles are in the 1998 denomination

Table A.2-1. Variables

Question	**Units/explanation**
Any health problems in last 30 days?	1-yes, 2-no
Treated by health worker or self	1-yes, 2-no
Paid for medical visit?	1-yes, 2-no
Paid cashier for medical visit?	1-yes, 2-no
Amt paid cashier medical visit	billions of current rubles
Paid personl for medical visit?	1-yes, 2-no
Amt paid personl med. Visit	billions of current rubles
Additional tests/procedures?	1-yes, 2-no
Paid for additional tests/procedures?	1-yes, 2-no
Paid cashier for medical tests?	1-yes, 2-no
Amt paid cashier medical tests	billions of current rubles
Paid personnel for medical tests?	1-yes, 2-no
Amount paid personnel for medical tests	billions of current rubles
Hospitalized in the last 3 months?	1-yes, 2-no
Paid for hospital medical care?	1-yes, 2-no
Paid cashier for hospital?	1-yes, 2-no
Amount paid cashier hospital	billions of current rubles
Paid personnel for hospital?	1-yes, 2-no
Amount paid personnel hospital	billions of current rubles
Paid for medicine whole/part	1-yes, 2-no
Paid cashier for medicine?	1-yes, 2-no
Amount paid cashier medicine	billions of current rubles
Paid personnel for medicine?	1-yes, 2-no
Amount paid personnel medicine	billions of current rubles
Total expenditures, yearly	billions of current rubles

extrapolation for the whole country	

Note: 1. Private clinics excluded.

Table A.2-2. Estimates.

1994	**Subsample, number of respondents**	**Maximum**	**Mean**	**Std. Deviation**	**Country total, (bn new rubles for amounts; persons for the rest)**
1. Health problems in last month?	11281	2	1.55	0.5	67500000
2. Treated by health worker or self	5064	2	1.55	0.5	30375000
3. Called doctor in or went to office?	2253	2	1.21	0.41	
4. Paid for medical visit ?	2259	2	1.96	0.2	1215000
5. Amount paid for medical visit	84	168000	29068.11	35144.37	35.3
6. Additional tests/procedures?	2253	2	1.59	0.49	12453750
7. Paid for additional tests/procedures?	923	2	1.91	0.29	1120837
8. Cost for additional tests/procedures	79	2000000	60742.64	234502.7	68.0
9. Hospitalized in last three months?	11278	2	1.94	0.23	9000000
10. Paid for hospitalization?	620	2	1.86	0.35	1260000
11. Cost of hospitalization	80	3500000	106736.8	380106.3	0.13
12. Total expenditures, for the year					1.8
1995	**Subsample, number of respondents**	**Maximum**	**Mean**	**Std. Deviation**	**Country total, (bn new rubles for amounts; persons for the rest)**
13. Health problems in last month?	10627	2	1.6	0.49	60000000
14. Treated by health worker or self	4210	2	1.56	0.5	26400000
15. Called doctor in or went to office?	1834	2	1.23	0.42	
16. Paid for medical visit ?	1830	2	1.95	0.21	1320000
17. Amount paid for medical visit	74	700000	83966.22	138832.6	0.11
18. Additional tests/procedures?	1831	2	1.61	0.49	10296000
19. Paid for additional tests/procedures?	710	2	1.93	0.25	720720
20. Cost for additional tests/procedures	46	600000	73134.78	109374	0.05

21. Hospitalized in last three months?	10625	2	1.95	0.21	7500000
22. Paid for hospitalization?	492	2	1.86	0.37	1050000
23. Cost of hospitalization	79	5000000	255271.41	701539.2	0.3
24. Total expenditures, for the year					3.0
1996	**Subsample, number of respondents**	**Maximum**	**Mean**	**Std. Deviation**	**Country total, (bn new rubles for amounts; persons for the rest)**
1. Any health problems in last 30 days?	10145	2	1.62	0.49	57000000
2. Treated by health worker or self	3892	2	1.56	0.5	25080000
3. Doctor to house or office visit	1692	2	1.22	0.41	
4. Paid for medical visit?	1693	2	1.95	0.21	1254000
5. Amount paid for medical visit	73	2000000	122407.3	243040.8	0.15
6. Additional tests/procedures?	1692	2	1.59	0.49	10282800
7. Paid for additional tests/procedures?	696	2	1.92	0.28	822624
8. Cost of additional tests/procedures	54	1000000	116663	178782.1	0.1
9. Hospitalized in the last three months?	10145	2	1.96	0.2	6000000
10. Paid for hospital stay/medical care?	442	2	1.78	0.42	1440000
11. Amount paid for hospital	98	5000000	378091	636836.8	0.54
12. Total expenditures, for the year					5.17
1998	**Subsample, number of respondents**	**Maximum**	**Mean**	**Std. Deviation**	**Country total, (bn new rubles for amounts; persons for the rest)**
1. Any health problems in last 30 days?	7798	2	1.6	0.49	60000000
2. Treated by health worker or self	3122	2	1.58	0.49	25200000
3. Doctor to house or office visit?	1297	2	1.21	0.4	
4. Paid for medical visit?	1300	2	1.93	0.26	1764000
5. Amount paid for medical visit	88	1400	121.9014	218.3841	0.22
6. Additional tests/procedures?	1300	2	1.54	0.5	11592000
7. Paid for additional tests/procedures?	597	2	1.83	0.38	1970640

8. Cost of additional tests/procedures	101	5000	130.8459	465.174	0.26
9. Hospitalized in the last three months?	7806	2	1.95	0.22	7500000
10. Paid for hospital stay/medical care?	387	2	1.56	0.5	3375000
11. Amount paid for hospital last three months	168	5000	552.04	769.68	1.9
12. Total expenditures, for the year					13.1
2000	**Subsample, number of respondents**	**Maximum**	**Mean**	**Std. Deviation**	**Country total, (bn new rubles for amounts; persons for the rest)**
1. Any health problems in last 30 days?	7400	2	1.58	0.49	63000000
2. Treated by health worker or self	3110	2	1.63	0.48	23310000
3. Doctor to house or office visit?	1153	2	1.2	0.4	
4. Paid for medical visit?	1154	2	1.9	0.31	2331000
5. Paid cashier for visit?	117	2	1.43	0.5	1328670
6. Amount paid cashier for visit	64	5000	322.69	749.85	0.43
7. Paid personnel for visit?	115	2	1.48	0.5	1212120
8. Amount paid personnel visit	54	15000	420.37	1817.38	0.51
9. Additional tests/procedures?	1155	2	1.58	0.5	9790200
10. Paid for additional tests/procedures?	519	2	1.83	0.38	1664334
11. Paid cashier for medical tests?	92	2	1.31	0.47	1148390.46
12. Amount paid cashier for medical tests	58	4700	290.76	718.91	0.33
13. Paid personnel for medical tests?	86	2	1.60	0.49	665733.6
14. Amount paid personnel medical tests	27	4500	395.35	812.12	0.26
15. Hospitalized in the last three months?	7405	2	1.95	0.22	7500000
16. Paid for hospital/medical care?	372	2	1.88	0.32	1050000
17. Paid cashier for hospital?	52	2	1.57	0.5	387000
18. Amount paid cashier hospital	24	1670	339.48	11579.03	1.3

19. Paid personnel for hospital?	52	2	1.43	0.5	513000
20. Amount paid personnel for hospital	24	5000	808.96	1062.64	0.42
21. Total expenditures, for the year					20.6
2001	**Subsample, number of respondents**	**Maximum**	**Mean**	**Std. Deviation**	**Country total, (bn new rubles for amounts; persons for the rest)**
1. Any health problems in last 30 days?	7751	2	1.56	0.5	66000000
2. Treated by health worker or self	4806	2	1.70	0.46	19140000
3. Paid for medical visit?	1407	2	1.88	0.32	2296800
4. Paid cashier for medical visit?	160	2	1.49	0.5	1211760
5. Amount paid cashier for medical visit	80	4000	358.1	591.05	0.43
6. Paid personnel for medical visit?	161	2	1.46	0.5	1283040
7. Amount paid personnel medical visit	82	3000	252.35	396.09	0.33
8. Additional tests/procedures?	1406	2	1.59	0.49	8118000
9. Paid for additional tests/procedures?	583	2	1.77	0.42	1867140
10. Paid cashier for additional tests/procedures?	143	2	1.26	0.44	1381683.6
11. Amount paid cashier additional tests/procedures	103	4500	226.35	492.7	0.31
12. Paid personnel for additional tests/procedures?	141	2	1.71	0.45	541470.6
13. Amount paid personnel additional tests/procedures	37	4000	407.7	757.2	0.22
14. Hospitalized in the last three months?	7761	2	1.95	0.22	7500000
15. Paid for hospital/medical care?	377	2	1.83	0.38	1275000
16. Paid cashier for hospital?	64	2	1.38	0.49	790500
17. Amount paid cashier hospital	39	4500	787.67	1119.9	0.62
18. Paid personnel for hospital?	64	2	1.48	0.5	663000

19. Amount paid personnel hospital	25	10000	1057.18	2014.9	0.7
20. Paid for medicine whole/part?	376	3	1.79	0.82	1575000
21. Paid cashier for medicine?	200	2	1.58	0.49	661500
22. Amount paid cashier for medication	71	4500	521.08	769.7	0.35
23. Paid personnel for medication?	200	2	1.91	0.26	141750
24. Amount paid personnel for medication	11	1000	415.16	509.9	0.06
25. Total expenditures (without expenditures on medication), for the year					21.0

A 3. Interviews with professionals and decision-makers.

In June-July 2000, the author conducted two interviews with doctors, one interview with a nurse and 16 interviews with experts and decision-makers from a number of medical insurance companies of the city of Saint Petersburg. The number of interviews with the medical staff was lower than desirable because it was extremely difficult to find those who could share sensitive information. The number of the interviews at the insurance companies was also limited by many refusals to participate (the staff being busy or, possibly, unwilling to share an opinion about the company's policies). The subject of the research was made explicit to the medical staff, while at the insurance companies, it was presented as "research into patient rights defense policies".

I use the interviews with doctors and nurses to find out not the opinions, actions, or perceptions of a particular person, but generalized facts about how payments of interest happen. The interviews with decision-makers and experts at the insurance companies administering MHI reveal the current policies of the insurance companies with respect to payments and also learn more facts concerning the phenomena of interest. I also asked their opinion regarding the causes and scope of transactions and the best ways of dealing with the problem.

There was no random selection of the interviewees, as it simply was not possible. Also, semi-structured interviews with many of the issues emerging in the course of the conversation appears to be the best way to approach the problem. The pros and cons of such an approach were reviewed in the methodological subsection of Chapter Three, section 3.2.

Interviews were based not on questions but on problems. For example, it was important to understand the nature of an insurance company's involvement in dealing with complaints about extortion of payments for things free by law. To this aim, the relevant questions somewhat varied in wording and order from interview to interview, depending on circumstances. In the course of the interview, questions were repeated or detailed in order to make sure the meaning of the answer is fully understood. The respondents were asked to fill out a questionnaire with basic questions, but that was only to provide background information on the respondent. This all was in line with the general strategy not to rely on statistical procedures, which are based on large number of observations and assumptions on distributions, untenable under the circumstances. Instead, I attempted to compensate for the small number of observations (compared with the number of questions to be asked) with an 'inquisitive' approach aiming to elicit as much information as possible and check for inconsistencies.

I assume that facts were truly reported during the interviews. To avoid simple lies, no questions were asked about what this particular doctor, nurse, or insurance company did, but what doctors, nurses and insurance companies usually do. This approach decreased sensitivity of information to be received.

A3-1. Interviews with doctors and nurse.

12-13.06.2000 Interview with two doctors, both surgeons (different hospitals, interviewed separately):

- Doctor A is in his early career
- Doctor B is a senior staff with extensive experience

15.06.2000 Interview with a nurse. The nurse had had two year's experience by the time of the interview, including that of assisting at operations.

Though preliminary questionnaires were distributed, the important information was elicited through unstructured conversation in order to make the respondent formulate his presentation of the situation as clearly as possible. Below, the exact wording of answers is given in parentheses.

I employed answers given by doctor A to formulate additional questions to doctor B, in form: "Would you agree that [the opinion of doctor A]". Obviously, due to the sequencing of interviews, a similar procedure of cross-checking could not be applied to doctor B's answers.

The interviews with the doctors were at their offices, while the interview with the nurse was at her home. Confidentiality was guaranteed. There was no financial motivation for participation in the interviews to the doctors; the nurse received ten US dollars for her cooperation.

Interviews with the doctors.

Problem 1: The process of negotiations.

Who takes the initiative in paying for a medical service? Both doctors admitted an active role of the doctor who approaches the patient. “The doctors start negotiation about the payment” and the process of negotiation involves “bargaining”. Yet, this role also implies a good deal of discretion, for those patients who are not likely to respond positively, are not asked.

Problem 2. Status of the difference between over the counter and under the counter payments: which payments can be called ‘forced’, ‘illegal’, and ‘related to quality differentiation’?

Doctor A differentiated between these two types of payment, namely considering only the under the counter payments as forced, illegal, and related to quality differentiation. He denied any quality differentiation induced by payments made via the *khozraschet* (paid service department, legal paid services) system, that is to say, over the counter. The other doctor (B) admitted that both types of payments are forced, contain an element of illegality, and are related to quality differentiation.

Problem 3. Willingness of patients to pay and their motivation.

Both doctors said that the patients are willing to pay under the counter because they believe that there is a significant quality differentiation. Doctor A denied the presence of quality differentiation for the over the counter payments but doctor B equalized both regimes of payments. They were reminded that they had mentioned doctors’ taking the initiative. I asked to make it more precise, whether patients may be expecting certain quality differentiation possible for payment before being approached with a particular proposal. Doctor B treated over the counter and under the counter payments as equal, and said that “everybody knows that one can receive better quality for money”, while doctor A seemed not to know much about the awareness of quality differentiation among patients. Yet, he said that if someone wants to get better treatment for money, it is essential to talk to the doctor directly, possibly after receiving some preliminary information by a telephone call.

Both doctors were asked to describe how patients were convinced to pay. Doctor A answered only about under the counter payments and did not consider himself knowledgeable about over the counter, ‘formal’ payments. The other doctor considered both regimes on a par: “There is no difference to the patient how to pay”. The most important and first mentioned way of convincing a patient to pay was elegantly summarized in the following slogan: “I will operate on you for money or my student will do it for free” (Doctor A). As this expression was subsequently reformulated for doctor B’s assessment, the latter said that he “basically agrees, though one does not necessarily say anything directly”. The choice of doctor here implies quality differentiation most clearly (both doctors agreed). As was subsequently confirmed by the nurses, the patients actively exchange information about doctors and the latter have reputations relevant for the patient to make decisions.

Way of treatment (for example, whether and when an operation is to be made) is a second important way of quality differentiation. Doctor A mentioned here the use of a more advanced equipment. Doctor B acknowledged that the equipment may be used for quality differentiation. For example, use of sophisticated techniques for appendectomy or any other operation is a way to differentiate quality. The doctor who considered only under the counter payments was specifically asked whether the patients paying over the counter have access to better equipment as well. He confirmed this.

In addition, as described by doctor B, the patient may jump the queue for an operation, which is indeed a quality differentiation, sometimes a life-saving one. Additionally, the patient may be required to pay for a very complex operation. This most often involved the choice of doctor, because not every doctor would be able to perform an operation of any complexity (the information was again given by doctor B).

The availability of medication is the last item mentioned. Two ways of selling medication and other items, such as implants of better quality, were mentioned. First of all, doctors may sell inexpensive medication that is supplied to the hospital and covered by public funds, to patients. The doctor promises to buy it from somewhere (a drug store, for example) though in fact this medication can be stored away by him or herself. The nurse confirmed the fact that doctors put away part of medication for later sale. Pain-killers are often supplied in this way, and are often oversupplied (mentioned specifically by doctor A, confirmed by the nurse).

Secondly, medication of extra quality, indeed unavailable for free, is supplied, often for overcharge and with exaggeration of its beneficial effects. In this respect A mentioned transplants. Doctor B mentioned payments for medication required for operations made for money.

When specifically asked, both doctors said that a doctor never buys medication at pharmacies due to the under-supply or under-funding from the public sources. This was the claim by one of the representatives of the insurance companies, interviewed later (see below).

Problem 4. How do patients get selected into paying and non-paying groups?

As has already been mentioned, the selection of patients to be asked for payment is based upon observation of their ability and readiness to pay (both doctors). The position of the doctor in the establishment, his or her reputation, and attending circumstances are also important factors. The severity of illness was not cited as an important factor. Yet, one can infer from the fact that complexity of operation is a factor decreasing the chances to get a free treatment (see above) that the severity of illness can at least indirectly influence the separation of patients into two groups.

Problem 5. The doctor's motivation.

The doctors both acknowledged that medicine is a business, thus a doctor is a natural profit-seeker. The corruptive behavior ensuing from this is caused and partly morally justified by the inadequate official wages. Then both doctors were asked whether they would expect payments (or at least under the counter payments) to fall either in amount or number of cases, if the pay were somewhat higher. They both denied that a simple increase of pay would change the situation. It should be combined either with a harsh system of punishment (doctor A) or strong connection between pay and performance (meaning here effectively fee for service arrangement for a professional and not only for the hospital – doctor B).

Doctor A also believed that the high number of applications to medical schools is connected with the expected high incomes, an expectation at odds with the low salaries of medical personnel. He also added that being medical professional also means good reputation and high social standing, not measurable in income. Thus, doctors would not quit their jobs in great numbers if devoid of the opportunity to earn extra money.

Problem 6. Attitude of management towards illegal charges

This is a question on which it was impossible to receive more information than agreement with the following statement. Management benefits from the payments in many ways, though it is disinterested in exceeding certain limits in scope and size of payments and will punish those who step out of these fuzzily defined limits.

Doctor A also said that if patients did not complaint about the quality of service, there would be no problem with the management of the hospital.

Interview with a nurse.

The interview with the nurse pursued two goals. In the first place, it was important to understand the transactions in which nurses were involved. Secondly, the nurse was also asked about transactions initiated by doctors to complement or confirm the information received from the doctors.

Problem 1. Who and on what occasion starts the negotiation. Whether the payments are tips?

Nurses rarely ask for payment, but it is common knowledge among the patients that nurses must be tipped. The tip-like payments for nurses comprise the first category of payments, which are not actually forced by conditioning a service on them. There is another category of payments, those for particular service. These payments are made in advance, in difference from the tips. The tips are usually paid in kind, not cash. Their amounts are usually very limited, up to equivalent of two US dollars. Cash payments are made for a particular service.

Problem 2. For what do patients pay?

Very roughly, patients pay for two things:

1. services that must be delivered for free, but will be delivered only for money;
2. services that need not be delivered free of charge.

The first category comprises the situations when a service (injection) implies a possibility of quality differentiation. For example, injection can be made on time or later, or missed altogether. Payment ensures the quality of the manipulation or service. The second category includes special attendance to the patient (permanent attendance in case of serious illness). The boundary between the two cases is very fuzzy. Because hospitals generally lack a sufficient nurse workforce, it is hard to define precisely what a nurse is supposed to do. Often doctors help nurses with their duties, which signals that nurses are already overloaded. Thus, if payment induces a nurse to work harder for someone, this can also be interpreted as doing what is not possible to do without the payment. Most of the cases of paid services, the nurse agreed, are related to doing what is done for other patients, only with higher quality.

Moreover, if a patient pays for something in excess of what is supposed to be delivered for free, then a nurse may spend time on working for the paying patients instead of working for non-paying patients. If a nurse stays overnight for extra payment, then she will not be able to put in good job over her ‘official’ shift. So, the system of paid nurse services is necessarily in conflict with the official obligations of a nurse, whether the service delivered for money is to be delivered for free or not.

The nurse was not aware of the over the counter payments to nurses. She admitted that sometimes patients require extra laboratory tests for which they pay but she could not comment on whether these payments are forced in any sense.

Problem 3. Interaction between doctors and nurses. Nurses and doctors often work in a tandem: nurse recommends a doctor if a patient is looking for better treatment, and then the doctor shares with the nurse the income that was received from the patient. Alternatively, doctors recommend nurses to the patient willing to pay. The professional reputation of both doctors and nurses plays an important role here. In general, good doctors and good nurses have better chances to earn extra income. The reputation is often based on consensus among the patients (for example, the chronically ill and elderly, who spend much time in the ward).

Problem 4. Different ways of payment.

Many payments are done in kind, both to nurses and to doctors. Doctors are interested in pre-selecting those patients who may render certain services to them in exchange for privileged treatment, for example fixing a car. The difference between in-kind and money remuneration of effort does not coincide with the difference between tips and payments for services or preferential treatment.

Problem 5. Are there groups of patients who can expect not to be approached for payment?

Emergency cases are more rarely approached for money than chronically ill or those subject to planned hospitalization. The reasons may be purely technical: there is less time for negotiation. Elderly people who look poor are not approached for money either. The nurse specified that the elderly may receive the same treatment as those who pay, so there may be no quality differentiation at all. There are moral reasons behind this lack of differentiation.

Problem 6. Relationship between staff and management on the issue of illegal, informal payments.

Hospital management is not interested in under the table payments made to nurses. More serious transactions, usually involving selling medication, can be a matter of internal investigation, but rarely so, and the nurse could not remember a case where an external agency made any intervention in these matters. The nurse agreed with doctor B that internal investigations are launched only upon receiving a complaint from the patient, which is always a complaint about the bad quality of treatment.

Problem 7 Policlinics.

The way payments for services free by law happen in policlinics is rather different from the case of hospitals. In policlinics, quality differentiation is not conveyed in words, but shown, since paid services crowd out free services. The mechanism is similar to the hypothetical competition for time and effort of professionals attributed above to payments in hospitals. Doctors are allowed to spend their extra hours with the paying patients. In fact, they become employed for serving the non-paying patients only part-time, and spend the rest of the time with the paying patients. As a result, the longer queues for the free services make those patients who have more money or less time pay. The interviews at some of the insurance companies confirm this information.

Discussion

Starting with a general impression I carried away from the interviews, it is a hesitant and guarded attitude of doctors and nurses involved in transactions of interest. They do not seem to act according to known rules. They rather play by ear, carefully making their way towards better income, better social status. They tread carefully across the minefield of moral dilemmas and administrative controls, seeking to compensate their professional effort without giving up on the poorer patient.

Additional to the scenarios reported in the Yaroslavl' interviews are the following. There is some selection of patients by those benefiting from the transactions. The selection is based on an estimation of the patient's social status and apparent readiness to pay: those are selected who are not likely to refuse to pay. One should not hasten to conclusions on the subject, for such information is scarce. Yet, the very fact of selection, however it proceeds, has important implications, as has been discussed in Chapter Three.

Payments in policlinics differ from those in hospitals. The most important and, perhaps, even fundamental difference is that there is no selection of patients in policlinics. Self-selection of patients may have two consequences. First of all, quality differentiation may become more hidden. Secondly, the self-selection would mean that there is a surer difference in treatment (and not necessarily in the actual quality of service) as compared to the hospital case.

Table A.3-1 below summarizes some cases of quality differentiation.

Table A3-1. Quality differentiation

Service/good	**Advantage for money**	**Justification of payment to the patient**	**Prices**
Operation	Payment allows choice of doctor and better treatment, but hardly doctor will work better for you than for a non-paying patient	"You can choose, either I will operate you, or a student will"	Generally varies and factors unknown. May separate into operation and medication components at least for patient
Cheap material for implantation, cheap medication, etc.	Expensive material for implantation, medication, etc. Often experimental medication or equipment	Indication (even exaggeration) of utility of the more expensive material/medication.	Medication costs USD 40, sold for USD 200
Transfer to another room	Less crowded room		USD 50
Magneto-resonance imaging	Condition for hospitalization in Hospital# 2 in 2001	Internal order in the hospital. Probably controlled by the authorities as salaries of doctors in	MHI tariff payable to hospital: 280 rubles Private payment: 750 rubles Hospital

		the hospital were recently down from seven to four thousand rubles	reimbursement under private insurance: 900 rubles

A3-2. Interviews at insurance companies.

An attempt to analyze laws and regulations pertaining to patient rights' defense offered a scenario by which the actual policy towards illegal charges (including efforts of investigating the alleged such charges) would partial abet this kind of petty corruption. Part of this abetting is engraved in the content of the law itself. Does the inadequacy of the law extend on to its implementation? Is there an effort made at the level of implementation of the regulation to compensate for the inadequacies in question, or rather the implementation is as inadequate as is the law?

To answer these questions, one should understand the actual procedures used by those who are in charge of defending the patients against the illegal charges. Insurance companies of Saint Petersburg, contracted under Mandatory Health Insurance arrangements, kindly provided some information in this regard. Some companies effectively agreed to engage in more in-depth loosely structured conversation, while the rest confined themselves to answering to the basic questionnaire (attached).

Ask-Med, Vesta, Nevskaya-Med, Kapital-Polis, Doverie, Medexpress provided extensive comments. *Spassk-Med, Rus'-Med, Med-Lux, Peterburgskaia Strakhovaia Kompaniia* [Insurance Company of Petersburg] offered only answers to the basic questionnaire. In addition, the director and chief financial officer of the Mandatory Insurance Fund of Leningrad *oblast*[62] answered the questionnaire.

Extensive interviews.

Problem 1. Procedures related to defense of patients' rights. Each company has a department of "quality control". By quality control, the following is meant. A patient

[62] Private companies are not contracted in Leningrad *oblast*.

has an opportunity to complain about his or her rights as an insured under MHI being violated. The spheres of complaints include denial of care or unlawful demands of payment, denial of specific rights, such as choice of provider, where applicable, violation of medical standards, etc.

The first important piece of information obtained invariably from all the respondents in the groups was that the relevant procedures were rather informal. The bulk of complaints are dealt with on the basis of phone calls to specific persons in medical establishments. Complaints are not filed, especially those related to demands of payment. This explains why there are not many such complaints in the official data.

The official statistics (see Chapter Two, section says that only 3 percent of all complaints (in 1998) were made on account of extorted payments. Based on this fact, the question was asked whether and for what reason patients tend not to complain when they are forced to pay. It transpired, however, that patients do complain in larger numbers. All the heads of the quality control departments said that there are very many complaints, but most of them are "oral" in the sense that their are made over the phone and never filed officially. At *Nevskaya-Med*, the number of complaints about forced payments officially filed was only 17, while up to 30% of all complaints, filed and not, are in some way related to private payments. At *Ask-Med* and *Vesta*, the large number of complaints was also confirmed, almost all of them being oral. At *Nevskaya-Med* the procedure of dealing with such complaints was described as follows. The interviewed refused to quote exact figures citing confidentiality.

The second description commonly supplied by the respondents dealt specifically with complaints about demands of payment. There are four major scenarios:

1. A payment was made under a formal contract, with all formalities attesting refusal of free care. It is impossible to investigate the case as one of unlawful/extorted payment.
2. A payment is asked for and the patient complains, before signing any papers. In this case, the company addresses the issue in an informal manner, often simply calling the relevant person in charge and asking to withdraw the demand for money.
3. A patient pays informally and complains only afterwards. There is nothing to be done.
4. When under obvious duress, a patient pays, informally or formally. Because of extreme circumstances, the insurance company may make extra effort in righting the wrong, even though normally, the company would not consider getting involved. The definition of 'extreme case' lies with the particular insurance company and the particular person in charge of the case.

This description of work of quality control departments is invariant with respect to the view a respondent would hold of the nature of such violations as denial of care and unlawful charges. This can be considered as a tentative warrant of adequacy of the information supplied.

Problem 2. Characterization of payments for services free by law.

The head of quality control department at *Vesta* and his counterpart at *RESO-Med* called the behavior of the provider towards patients 'extortion'. It is interesting that at *Vesta*, the payments were viewed as wide-spread, though amounts and frequencies could not be appreciated, while at *RESO-Med*, the amounts were considered as insignificant. Both respondents firmly asserted that the patients are paying for services, that is to say, for professional effort.

The opinion of the head of the quality control department at *Ask-Med* is the exact opposite to this. Agreeing that patients do contribute significant sums of money to the healthcare system, he believes that these are payments for medication, which is under-supplied to the hospitals. In his opinion, the medical professionals are forced to buy at drugstores the medication for operations and other use, and thus ask patients to contribute with their money. The doctors themselves confirm that some doctors sell medication to the patients from hospital stashes and never buy anything at pharmacies.

Doverie denied significance of illegal payments in terms of overall inflow of money in the healthcare system. Instead, the director of the company suggested that the burden on the patients is very high in terms of the portion of their income they are forced to pay for health.

Nevskaia-Med presented the issue somewhat differently. Though doctors do charge for their services unlawfully, they do so because of the dismally low pay they receive from the state. Payments are made for services and medication, so personal remuneration of doctors is acknowledged.

Kapital-Polis was of opinion that the overall distribution of forcing the patient to pay for something is in favor of certain 'top' institutions, which engage in both formal and informal ways of charging. In a way, the patient pays for reputation of a place or a doctor. The interview at *Kapital-Polis* gave some interesting information about pseudo-quality differentiation. At Hospital #2, which is deemed one of the most technologically advanced and otherwise reputed hospitals in the city, the admission of a patient under MHI policy happens only after a magnetic resonance screening (as of the time of the interview), for which patients pay 700 rubles. Given that magnetic resonance is not required for all patients, this is obviously a pseudo-quality differentiation, which is at the same time a way of concealing co-payment for admission to a prestigious hospital.

Problem 3. Ways of combating corruption and the apparent 'look the other way' official attitude.

The overall impression is that there is no official policy on restraining the illegal charges. This is rather solidly confirmed by all what the respondents said and by the very discrepancy of their reactions to questions. Moreover, Nevskaya-Med seemed to indicate that there is a policy of allowing illegal charges.

Nevskaya-Med justified 'look the other way' policy with the argument: "harsh measures are useless". One cannot maintain healthcare system without allowing the payments. *MedExpress* and *Vesta* acknowledged that there is no policy against the payments and that they are not empowered to invent one. *Vesta* admitted that the main insurer in the city, Municipal Medical Insurance Company does not care about the problem whatsoever. Both considered the problem of payments services free by law as that of extortion for personal enrichment. *Doverie* made a stress on refusal to complain due to dependence of patients on doctors, especially in the rural areas. This is the main reason why illegal charges persist. *Ask-Med* denied the very existence of the problem, but suggested that use of written contracts would be one reason for illegal charges to go unpunished.

Short interviews

Sixteen interviews were conducted at twelve insurance companies and Regional Fund of Mandatory Health Insurance for Leningrad *oblast*. The following presents answers to some questions concerning charges for services free by the law, or illegal charges as they were called in the questionnaire.

First of all it seems that there is some readiness to recognize that patients pay for services free by law, and not just for extra care. Seven out of 16 believed that illegal charges are a greater problem than quality of treatment in the public healthcare. Only two respondents believed that illegal charges could not be characterized as an 'acute' problem. These two surprisingly believed illegal charges to be the greatest problem among a number of others. Others either characterized it as a 'problem' or an 'acute problem'.

Seven respondents agreed that increased salaries or an increase overall funding are the most valuable tool of solving the problem. Others chose between administrative means and patient awareness of their rights.

A sizable majority of 12 respondents believed difficulties with proving violation of patient rights prevent patients from complaining about illegal charges. The alternative was that patients bought better quality and care and were thus satisfied. Three respondents chose this latter option.

Respondents were also somewhat reluctant to blame under-funding for the problem of illegal charges. Only half of respondents chose this option, less than one would expect. When assessing the size of payments, 12 respondents said that amounts paid for services free by law are significant. Only four said that the public healthcare would probably have survived without these payments. Three respondents characterized such payments as tips, though all these believed amounts to be significant. Two respondents thought them to be extorted, one believing the amounts to be insignificant. Eight respondents said that one could easily talk about a regular paid medicine behind the illegal charges.

A majority believed that it happens or may happen that a patient has to pay an amount ten times exceeding the patient's monthly income for an operation free by law. Eight respondents were sure such a situation was possible; four answered "probably, yes" to the question.

In conclusion, a majority of respondents believe that the problem exists. This fact must be weighed against the apparent sensitivity of the issue. Contrary to expectations, there was no consensus about a dominant role of under-funding. Also, many respondents believe payments to be forced rather than freely made by the patient in order to reward the doctor and receive better care. Payments are not considered as innocent tips or gifts. I expected more weight given to the tipping or any other innocent interpretations.

The market-type relations between patients and doctors to supplement insufficient public funding are not the dominant picture in the minds of the respondents. Institutions matter, and the institutions oppress rights rather than facilitate trade. The apparent lack of ready answers might indicate that the problem does not get much hearing and the respondents could not possibly be instructed or learn to answer my questions in a 'right' way.

Appendix B. Addenda to the Model of Chapter Five

B1. Effort-Rent Trade-off: Continuous Case

All notation and interpretation is as in Chapter Five. Subscript j=1,2,…n denotes patient. Type is now uniformly distributed variable β (between β_{min} and β_{max}). A (type, patient) pair is therefore β_j. The State's program is:

$$\max_{\pi;\{e_j\}} \int_{\beta_{\min}}^{\beta_{\max}} d\beta \sum_{j=1}^{n} \{ \pi[h_j(e_j) - \mu_0 t_j] + (1-\pi)h_j(e_j\}$$

subject first of all to a constraint, which follows from the notion of patient rationality:

(1) $(1\text{-}\pi)(dh_j / de_j - \mu_j dp_j / de_j) = 0$

that is to say, patient i maximizes the difference between health status gained and money paid for it, if the policy is π=0.

The remaining constraints are related to the Provider's choice: Incentive Compatibility and Individual Rationality constraints. These constraints are derived momentarily:

The Provider's utility function is:

(2) $Max\ \Sigma[\pi t_j + (1\text{-}\pi)\, p_j - e_j + Bh_j]\ wrt\ \{ e_j\}$

(3) $s.t.\ E \geq \Sigma e_j$

Monetary transfer by state (t_j) should compensate utility from 'cheating', that is to say, provider pretending serving a patient with a lower effort effectiveness $\beta_j' < \beta_j$, β_j being the true effort effectiveness. One takes into account that Provider enjoys a patient's good health to a degree. Denote rent to type β_i under truth-telling by

(4) $r_j(\beta_j) = \pi t_j(\beta_j) + (1\text{-}\pi)\, p_j(\beta_j) - e_j(\beta_j)$

The incentive compatibility constraint then is:

(5) $d\, r_j /d\beta_j \geq -\partial e_j /\partial\beta_j - B\partial h_j /\partial\beta_j$.

Provider's Individual rationality constraint is:

(6) $\pi t_j(h_j(e_j)) + (1-\pi)\, p_j(h_j(e_j)) - e_j + B(h_j(e_j) \geq 0$; or:

(7) $r_j\, (\beta) \geq -B\, h_j\, (\beta)$.

To avoid unpleasant and uninteresting technicalities, we need to assume that rent increases in type. This is called the single crossing property, and it is a more than sufficient condition for implementability of choice functions (see Fudenberg and Tirole 1991, 258-260):

(8) $(dr_i /d\beta \geq 0)$

A further constraint follows from State or a patient not willing to pay 'extra' (rent is costly as money is valued positively by both State and the patient):

(9) $r_j\, (\beta_{min}) = -B\, h_j\, (\beta_{min})$.

This partially determines the integration constant fixing the expected rent. To summarize, the State's program is:

$$(10)\ \max_{\pi;\{e_j\}} \int_{\beta_{\min}}^{\beta_{\max}} d\beta \sum_{j=1}^{n} \{\ \pi[h_j\ (e_j\ ) - \mu_0 t_j] + (1-\pi) h_j(e_j\}$$

subject to the following constraints, dubbed quite unimaginatively for easy reference:

1. Patient Rationality (PR):

(11) $(1-\pi)(dh_i / de_i - \mu_i dp_i /de_i) = 0$;

2. IC:

(12) $d\, r_i /d\beta \geq -\partial e_i /\partial\beta - B\partial h_i /\partial\beta$

(13) $(dr_i /d\beta \geq 0)$

3. IR:

(14) $r_i\, (\beta) \geq -B\, h_i\, (\beta)$

4. Costly Rent (CR):

(15) $r_i\ (\beta_{min})= -Bh_i\ (\beta_{min})$.

5. Total Effort (TE):

(16) $E \geq \Sigma\, e_i$

Assume for simplicity that TE does not bind. Then the model can be solved as follows. The choice of regime (π=0 or 1) is made by comparing two solutions, namely, levels of effort and the ensuing health status, for each regime. Here I shall give only the solution for π=1; the other regime is analogously solved and comparison is straightforward.

We have a regular optimal control problem with two control variables, *π* and e_i; state variable r_i; and parameter *β*. It can be turned into the canonical form by adding the costate variable and a Lagrange multiplier for TE. But there is a mathematically less precise, but intuitively more straightforward way to solve the problem.

Rent to type j can be expressed as follows:

(17) $r_j(\beta_j) = -\int_{\beta_j}^{\beta_{\max}} d\beta[\partial e / \partial \beta_j \ \ + B\partial h_j / \partial \beta_j] = e_{\min} - e(h_{\min}, \beta_j) - B(h_j - h_{\min})$

This rent increases in type, because of the uniform distribution, provided in equilibrium healthcare also does (the single crossing property). This rent does not depend on the equilibrium effort for type *j*, for this effort does not enter the incentive compatibility conditions. Alternatively, one can express rent as a function of effort for type *j*:

(18) $r_s(e_j) = [e_j - e(h_j, \beta_s) - B(h_s + h_j)]$

(19) $Er_s(e_j) = \int_{\beta_j}^{\beta_{\max}} d\beta_s [e_j - e(h_j, \beta_s) - B(h_s + h_j)]$

The last line is the expected rent, where expectation is taken over all types, whose rent depends on effort for type *j*. The State maximizes the difference between expected health status and this expected rent. The State's program turns into:

$$(20)\quad \max_{\{e_j\},\pi} \int_{\beta_{\min}}^{\beta_{\max}} d\beta_j \sum_{j=1,2\ldots n} \pi(h_j - e_j - r_j) + (1-\pi)h_j$$

and we take only the regime when the State pays. Rewriting:

$$(21)\ W(\pi = 1) = \int_{\beta_{\min}}^{\beta_{\max}} \beta_j [h(e_j) - \mu_0 e_j - Er_s(e_j)]$$

and differentiating with respect to the effort for type j, I obtain the first order condition:

$$(22)\ \frac{\partial h_j}{\partial e_j} = \mu_0 (1 + \frac{d\int_{\beta_j}^{\beta_{\max}} d\beta_s\ [e_j - e(h_j, \beta_s) - B(h_s - h_j)]}{de_j}) = 0$$

The second order conditions will be satisfied, if the expected rent to types more effective than j is convex in effort e_j and benevolence is sufficiently small.

From (17) and (22) the following classical results can be recovered:

1. Rent for the least effective type (when $\beta_s = \beta_{min}$) is $-Bh_{min}$, because then $j=s$ and (17) is zero up the constant defined by individual rationality;
2. Rent for the most effective type is maximal;
3. Effort for the most effective type is first-best (from (17), when $\beta_j = \beta_{\max,}$ the integral is equal zero and its derivative is equal zero);
4. The derivative of the state variable, rent, at the most effective type is zero (which is a restatement of point 2).

For example, consider the functional form:

h(e)=ln(1+βe)

Then:

$$(23)\ \frac{\partial h_j}{\partial e_j} = (1-\mu_0 B)(\beta_{\max} - \beta_j))\frac{\beta_j}{1+\beta_j e_j} = \mu_0 + \mu_0(\beta_{\max} - \beta_j - \beta_{\max}\ln\frac{\beta_{\max}}{\beta_j}) = 0$$

Simplifying:

$$(24)\, e_j{}^* = \frac{(1-\mu_0 B(\beta_{\max} - \beta_j))}{\mu_0 + \mu_0(\beta_{\max} - \beta_j - \beta_{\max}\ln\frac{\beta_{\max}}{\beta_j})} - \frac{1}{\beta_j}$$

B.2. Joint Consumption of Healthcare

The idea is to show distribution of costs and benefits, when Provider controls some scarce resource to be distributed among a fixed number of patients and when Provider can charge only a uniform price. This scarce resource will be called ‘healthcare’, but in reality could be access to a single high-tech device, or to a limited source of high quality medication, or to attention of some very reputed or very good professional. The basic assumptions are:

- Total capacity of provider is equivalent to amount 1 of healthcare.
- By a law, each patient must be entitled to at least amount q, $0 \leq q \leq 1$ of healthcare.
- There is a continuum of patients. n, $0 \leq n \leq 1$ is the share of those receiving free care;

Each patient’s utility from consuming healthcare is defined as a benefit function $b(r,i)$, where r is the amount of care consumed and i is the number, designating the place of the patient in the continuum. $\partial b(r,i)/\partial r>0; \partial^2 b(r,i)/\partial r^2 \leq 0; \partial b(r,i)/\partial n \leq 0.$ In words, patients are ordered by decreasing utility from healthcare.

The provider is a profit-maximizer. The provider divides its capacity into Free Care and Paid Care and charges price p for access to Paid Care. Those in Free Care receive q, as per the statutory provision. Those in Paid Care receive each equal share of the remaining healthcare. The way the capacity is divided depends on demand from patients. Formally, Provider maximizes:

$$(1)\quad \begin{array}{l} \max : (1-n)p \\ wrt : p, n \\ s.t. b(\frac{1-qn}{1-n}, n) - p = b(q,n) \\ Or : \\ \max : (1-n)b(\frac{1-qn}{1-n}, n) - b(q,n) \\ wrt : n \end{array} \quad ;$$

The first order condition is:

(2) $b(q,n) - b(\frac{1-qn}{1-n},n) + \frac{\partial b}{\partial r}\frac{1-q}{1-n} = 0$;

In other words, optimal price p* is:

(3) $$\begin{array}{l} p = \frac{\partial b}{\partial r}\frac{1-q}{1-n} \\ r = \frac{1-nq}{1-n} \end{array}$$

The optimal price decreases in *q*. This creates an interesting dynamics, that may have further consequences. A paying patient welfare depends on two characteristics: the price for paid healthcare and the amount of healthcare the paying patient receives. It is possible to construct a situation when increasing *q* is beneficial for some of the paying patients, who stay in Paid Care and detrimental to those who are in Free Care, because increasing *q* makes some patients leave Paid Care and join Free Care.

Sources and Bibliography

1. Altai. 1998. Administration of Altai Krai. Decree 473 July 29.

2. Antipova, N.P., et al. 1998. *Analyz upravleniya sistemoi zdrovookhraneniya v Yaroslavskoi oblasti* [Analysis of Healthcare Governance in Yaroslavl' *Oblast*]. TACIS Russia Healthcare Management Project, Moscow: http://zdravinform.ru/pub/EU.1998.A.13.R.doc.

3. Arrow, Kenneth J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review* LIII(5): 941-973.

4. Blomqvist, Åke. 2001. *Economic Efficiency and QALY-based Cost-Utility Analysis in Health Care.*
http://www.worldbank.org/hnp/hsd/documents/CUA.pdf

5. Bogatova et al. 2002. *Besplantnoe zdravookhranenie: realnost' i perspektivy* [Free Healthcare: Reality and the Future]. Moscow: IISP.

6. Boikov V. et al. 1998. "Raskhody naseleniaia na meditsinskuiu pomotsch i lekarstvennye sredstva" [Private Expenditures on Healthcare and Medication]. *Voprosy ekonomiki* 10:101-117.

7. Boikov V. et al. 2000a. "Raskhody naseleniaia na meditsinskie uslugi i lekarstva" [Private Expenditures on Health Services and Medication]. *Zdravookhranenie* 2:32-46

8. Boikov V. et al. 2000b."Uchastie naseleniya v finansirovanii zdravookhraneniya " [Participation of the Populace in Financing Healthcare]. *Ekonomika zdravookhraneniia* 7:45-50.

9. Brennan, Troyen A. 1992. "An Empirical Analysis of Accidents and Accident Law: the Case of Medical Malpractice Maw." *Saint Louis University Law Journal* 36:823-861.

10. Brown, Archie, ed. 2001. *Contemporary Russian Politics: A Reader.* Oxford: Oxford University Press.

11. Carney, Thomas F.1972. *Content Analysis: a Technique for Systematic Inference from Communications.* Winnipeg: University of Manitoba Press.

12. Chalkey, Martin and James M. Malcomson. 2000. "Government Purchasing of Health Services". In *Handbook of Health Economics, vol. 1B*, ed. Anthony J. Culyer and Joseph P. Newhouse. Amsterdam: Elsevier.

13. Chernez et al. 2003. *Finansovye aspekty reformirovaniya otraslei sotsialnoi sfery* [Financial aspects of social reform]. Moscow: Institute of Economy in Transition. http://www.iet.ru/papers/60/index.htm.

14. Chernichovsky, Dmitry et al. 1996. "Health System Reform in Russia: the

Finance and Organization Perspective.” *Economics of transition* 4(1):113-134.

15. Chernichovsky, Dmitry et al. 1998. Inequality of Health Finance, Resources, and Mortality in Russia: Potential Implications for Health and Medical Care Policy. In *Health, Health Care and Health Economics: Perspectives on Distribution*, ed. Morris L. Barer et al. London: Wiley.

16. Constitution. 1993. *Konstitutsia Rossiiskoi Federatsii* [The Basic Law of the Russian Federation].

17. Consumer Rights Protection Law. 1992. *Zakon Rossiiskoi Federatsii “O Zatschite prav potrebitelei”* [Consumer Rights Protection Law of the Russian Federation]. Law 2300-1 February 07.

18. Cullis, John G. and Philip R. Jones. 2000. “Waiting Lists and Medical Treatment: Analysis and Policies.” In *Handbook of Health Economics, vol. 1B*, ed. Anthony J. Culyer and Joseph P. Amsterdam: Elsevier.

19. Dauer, Edward A. and Leonard J. Marcus. 1997. “Adapting Mediation to Link Resolution of Medical Malpractice Disputes with Health Care Quality Improvement.” *Law and Contemporary Problems* 60(1):185-215.

20. Den Exter, Andre and Herbert Hermans. 1999. “The right to health care: A changing concept?” In *The Right to Health Care in Several European Countries*, ed. Andre Den Exter and Herbert Hermans. London: Kluwer Law International.

21. Dmitriev, Mikhail. 2004. “Bogatyi zaplantit za bednogo. A zdorovyi za bolnogo” [The rich will pay for the poor. And the healthy for the ill]. *Rossiiskaya Gazeta*. February 17.

22. Donaldson, Molla. 1991. *Assessing and Paying for Quality.* In: Jonathan D. Moreno, ed. *Paying the Doctor: health policy and physician reimbursement.* New York: Auburn House.

23. Dowding, Keith. 2001. “There Must Be End to Confusion: Policy Networks, Intellectual Fatigue, and the Need for Political Science Methods Courses in British Universities.” *Political Studies* 49(1):89-105.

24. Eisenberg, J.M. 1986. *Doctor’s decisions and the Cost of Medical Care.* Ann Arbor: Health Administration Press, MIT.

25. Elster, Jon. 1989. *Solomonic Judgments.* Cambridge: Cambridge UP.

26. Evans Robert G. 1983.”Health Care in Canada”. *Journal of Health Politics, Policy and Law* 8:1-43.

27. Federal Fund of Mandatory Health Insurance. 1995. *Vremennyi poryadok finansovogo vzaimodeistviya[...]* [Provisionary Rules of Financial Management…]. In *Obyazatelnoe meditsinskoe strakhovanie v Rossiiskoi Federatsii* [Mandatory Health Insurance in the Russian Federation], Vol. 1.

Moscow: Federalnyi Fond OMS.

28. Federal Fund of Mandatory Health Insurance. 1998. *Obyazatelnoe meditsinskoe strakhovanie v Rossiiskoi Federatsii v 1997 godu* [Mandatory Health Insurance in the Russian Federation, 1997]. Moscow: Federalnyi Fond OMS.

29. Federal Fund of Mandatory Health Insurance. 1999. *Obyazatelnoe meditsinskoe strakhovanie v Rossiiskoi Federatsii v 1998 godu* [Mandatory Health Insurance in the Russian Federation, 1998]. Moscow: Federalnyi Fond OMS.

30. Federal Fund of Mandatory Health Insurance. 2000. *Obyazatelnoe meditsinskoe strakhovanie v Rossiiskoi Federatsii v 1999 godu* (Mandatory Health Insurance in the Russian Federation, 1999). Moscow: Federalnyi Fond OMS.

31. Federal Fund of Mandatory Health Insurance. 2001. *Obyazatelnoe meditsinskoe strakhovanie v Rossiiskoi Federatsii v 2000 godu* (Mandatory Health Insurance in the Russian Federation, 2000). Moscow: Federalnyi Fond OMS.

32. Federal Fund of Mandatory Health Insurance. 2002. *Obyazatelnoe meditsinskoe strakhovanie v Rossiiskoi Federatsii v 2001 godu.* [Mandatory Health Insurance in the Russian Federation, 2001]. Moscow: Federalnyi Fond OMS.

33. France, George. 1999. “The Changing Nature of the Right to Health Care in Italy.” In *The Right to Health Care in Several European Countries*, ed. Andre Den Exter and Herbert Hermans. London: Kluwer Law International.

34. Freeman, Richard. 2000. *The politics of Health in Europe.* Manchester: Manchester University Press.

35. Fudenberg, Drew and Jean Tirole. 1991. *Game Theory*. Cambridge, Mass.: MIT Press.

36. Fukuyama, Francis. 1996. Trust: The Social Virtues and the Creation of Prosperity. London: Penguin Books.

37. Gabel, J.R. and T.H. Rice. 1985. “Reducing expenditure for physician services: the price of paying less.” *Journal of Health Politics, Policy and Law* 9:595-609.

38. Galasi, Péter and Gábor Kertesi. 1989. “Rat Race and Equilibria in Markets with Side Payments under Socialism”. *Acta Oeconomica* 41: 267-292.

39. Garfield, Signey R. 1978. “The Delivery of Medical Care.” In *Health Services Management*, ed. Anthony R. Kovner and Duncan Neuhauser. Ann Arbor: Health University of Michigan.

40. General Attorney Office. 1996. *O narusheniayakh zakonodatel’stva ob okhrane zdorov’ya detei* [Regarding violation of legislation on child healthcare]. Report 21-22-96 June 19.

41. Giddens, Anthony. 1990. *The Consequences of Modernity.* Cambridge: Polity

Press.

42. Goskomstat. 2003a. *Rossiia v tsifrakh* [Russia in numbers]. Moscow, Goskomstat.

43. Goskomstat. 2003b. *Regiony Rossii 2002* [Regions of Russia 2002]. Moscow: Goskomstat.

44. Government. 1998. *Programma gosudarstvennykh garantii obespecheniya grzhdan Rossiiskoi Federatsii besplatnoi meditsinskoi pomotschiiu.* [Program of implementation of free care entitlement for the citizens of the Russian Federation]. Decree 1096 September11.

45. Government. 2000a. Government of the Russian Federation. *Osnovnye napravleniya sotsialno-ekonomicheskoi politiki Pravitelstva Rossiiskoi Federatsii na dolgosrochnuiu perspektivu.* [Principal Long-term Objectives of Social and Economic Policy of the Government of the Russian Federation].

46. Government. 2000b. Government of the Russian Federation. *Plan deistvii Pravitelstva Rossiiskoi Federatsii v oblasti sotsialnoi politiki i modernizatsii ekonomiki na 2000-2001 gody.* [Action Plan the Government of the Russian Federation in the areas of social policy and economic modernization for 2000-2001].

47. Harrison, Michael I. "Health Professionals and the Right to Health Care." In *The Right to Health Care in Several European Countries*, ed. Andre Den Exter and Herbert Hermans. London: Kluwer Law International.

48. Health Insurance Law. 1991. *Federalnyi zakon "O meditsinskom strakhovanii grazhdan Rossiiskoi Federatsii"* [Federal law on health insurance of citizens of the Russian Federation]. Law 1499-1 June 28.

49. Health Protection Law. 1993. *Osnovy zakonodatelstva Rossiiskoi Federatsii ob okhrane zdoroviya grazhdan.* [Basic Law of Health Protection for Citizens of the Russian Federation]. Law 5487-1 July 22.

50. Hough, Jerry F. 2001. *The Logic of Economic Reform in Russia.* Washington: The Brookings Institution.

51. Immergut, Ellen M. 1992. *Health Politics: Interests and Institutions in Western Europe.* Cambridge: Cambridge University Press.

52. International Monetary Fund. 2003. *Russia Country Report No.03/145.* www.imf.org.

53. Ivánova T.G. et al. 1999. Review of Health Care Management System of Chuvash Republic. Moscow: TACIS Russia Healthcare Management Project http://zdravinform.ru/pub/EU.1998.A.4.E.pdf

54. Kanavos, Panos and John Yfantopoulos. 1999. "Cost Containment and Health Expenditure in the EU: a macroeconomic perspective." In *Health care and cost*

containment in the European Union, ed. Elias Mossialos and Julian Le Grand. Ashgate.

55. Kemerovo. 2000. Government, Kemerovo *oblast. Ob utverzhdenii polozheniia o poryadke okazaniia meditsinskikh uslug sverkh territorialnoi programmy gosudarstvennykh garantii obespecheniia naselenniia besplatnoi meditsinskoi pomotsch'iu.* [Ratification of Rules of delivery of paid medical services in excess of the Regional Program of state guarantees of free medical services to the population]. Decree 85 October 27.

56. Kessel, Reuben A. 1958. "Price Discrimination in Medicine." *Journal of Law and Economics* 1:20-53.

57. Kessel, Ronald. 1978. "The Hospital Business." In *Health Services Management*, ed. Anthony R. Kovner and Duncan Neuhauser. Ann Arbor: Health Administration Press.

58. Khabarovsk 2001. Mayor of the city of Khabarovsk. *O kontraktnykh otnosheniyakh s rukovoditelymi lechebnykh uchrezhdenii* [Regarding contracts with senior management of healthcare providers]. Decree 239 March 21.

59. Khabarovsk *krai.* 2001. Legislative *Duma* of Khabarovsk *Krai. O rekommendatsiiakh deputatskikh slushanii po teme "Ob okazanii platnikh uskug v Khabarovskom krae* [Recommendations following hearings on rendering of paid medical servces in Khabarovsk *krai*]. Decree 1347 July 25.

60. Khan, Mushtaq H. 1996. "The Efficiency Implication of Corruption." *Journal of International Development* 8(5):683-696.

61. King G., et al. 1994. *Designing Social Inquiry: Scientific Influence in Qualitative Research.* Princeton: Princeton UP.

62. Klyamkin I. and L. Timofeev. 2000. *Tenevaya Rossiia* [The Shady Russia]. Moscow: Russian State University of Humanities.

63. Kornai, János and Karen Eggleston. 2001. *Welfare, Choice and Solidarity in Transition.* Cambridge: Cambridge University Press.

64. Kornai, János. 2000. *Hidden in an Envelope: Gratitude Payments to Medical Doctors in Hungary.* Budapest: Collegium Budapest, Institute for Advanced Study, Discussion Paper Series No. 60.

65. Krizova, Eva. 1999. "The Patients' Rights as an Important Issue in the Process of Civic Emancipation in the Czech Republic." In *The Right to Health Care in Several European Countries*, ed. Andre Den Exter and Herbert Hermans. London: Kluwer Law International.

66. Kulibakin I.B. and E.A. Morozova. 1998. *Otchet po resultatam sotsiologicheskogo issledovaniaia[...].* [A Sociological Research Report [...]]. Kemerovo: Sociology Center, Kemerovo State University.

67. Laffont, Jean-Jacques and Jean Tirole. 1993. *A theory of incentives in procurement and regulation.* Cambridge, Mass.: MIT Press.

68. Laffont, Jean-Jacques. 2002. *The Theory of Incentives: the Principal-agent Model.* Princeton: Princeton University Press.

69. Landers, Renee M. et al. 1996. "Rat Race Redux: Adverse Selection in the Determination of Work Hours in Law Firms". *American Economic Review* 86(4):329-348.

70. Leningrad *oblast*. 1997. Government of Leningrad *oblast*. *O predostavlenii plantykh meditsinskikh uslug naleleniuiu* [Paid medical service delivery]. Decree 191 November 4.

71. Lindlof, Thomas R. 1995. *Qualitative Communication Research Methods.* Thousand Oaks: Sage Publications.

72. Linnako, Eero. 2002. *Health care financing in Russian Federation.* http://zdravinform.ru/pub/EU.1998.B.4.E.pdf

73. Lo Schiavo, Luca. 2000. "Quality Standards in the Public Sector: Differences between Italy and the UK in the Citizen's Charter Initiative." *Public Administration* 78(3):679-698.

74. Locock, Louise. 2000. "The Changing Nature of Rationing in the UK National Health Service." *Public Administration* 78(1):91-109

75. Makarova T.N. 2000. Implementation of new payment methods of health services: experience of Russia in 1988–1999 (Analysis of the results and proposals for further reforms). http://zdravinform.ru/

76. Mattheus, R.C.O. 1991. "The Economics of Professional ethics: Should the professions be more like business?" *The Economic Journal* 101:737-759.

77. McGuire, T.G. 2000. "Physician Agency". In *Handbook of Health Economics, vol. 1A*., Anthony J. Culyer and Joseph P. Newhouse, eds. Amsterdam : Elsevier.

78. Mehmet, Bac. 1996. "Corruption and Supervision Costs in Hierarchies." *Journal of Comparative Economics* 22:99-108.

79. Miller, William L. et al. 2001. *A Culture of Corruption? Coping with Government in Post-Communist Europe.* Budapest: CEU Press.

80. Ministry of Finance. 1999. Letter 24-02/11.

81. Ministry of Health. 1996. *Ob utverzhdenii pravil predostavleniya plantnykh meditsinskikh uslug naseleniiu meditsinskimi uchrezhdeniyami.* [Rules of paid medical services supplied to citizens by medical establishments]. Decree # 27 of January 13, 1996.

82. Ministry of Health. 1999. *Metodicheckie rekomendatsii […]* (Recommendations on methods […]) Decree 01-23/4-10.

83. Ministry of Health. 2001. *Kollegiia Ministerstva zdravookhraneniia Rossiiskoi Federatsii* Protocol 19, November 28.

84. Morozova E.A. and I.B. Kulibakin. 1998. *Otchet po resultatam sotsiologicheskogo issledovaniya "Gorodskoe zdravookhranenie"*. [City healthcare. Report on a sociological research]. Kemerovo: Sociology Center, Kemerovo University.

85. Moscow. 1996. Health Committee of Moscow Administration. Ordinance 183 March 29.

86. Moscow. 2000. Health Committee of Moscow Administration. Ordinance 330 July 26.

87. Mossialos, Elias and Julian Le Grand. 1999. "Cost Containment in the EU: an Overview." In *Health care and cost containment in the European Union*, ed. Elias Mossialos and Julian Le Grand. Aldershot: Ashgate

88. Neuhause, J.P. 1970. "Toward a Theory of Nonprofit Institutions: An Economic Model of Hospital". *American Economic Review* 60:64-74.

89. Orel. 2000. Head of Administration of Orel *oblast*. Decree 343, July 11.

90. Orel. 2000. *Perechen' platnykh uslug, okazyvaemykh zhitelyam orlovskoi oblasti v gosudarstvennykh i municipalnykh meditsinskikh uchrezhdeniiakh* [List of paid services delivered to residents of Orel *oblast* by the state and municipal healthcare establishments]. Governor Decree 343 July 11 (Appendix 3).

91. Osborne, Martin J. and Ariel Rubinstein. 1994. *A course in game theory.* Cambridge, Mass.: MIT Press.

92. Parsons, Talcott. 1975. "The Sick Role and the Role of the Physician Reconsidered." In *Action Theory and the Human Condition*, Talcott Parsons. New York: The Free Press.

93. Pauly, Mark V. 2000. "Insurance Reimbursement." In *Handbook of Health Economics, vol. 1A*., ed. Anthony J. Culyer and Joseph P. Newhouse. Amsterdam: Elsevier.

94. Pereira, Joao. 1999. "Health care reform and cost containment in Portugal." In *Health care and cost containment in the European Union*, ed. Elias Mossialos and Julian Le Grand. Aldershot: Ashgate.

95. Perm Regional Human Rights Center. 2001. *Doklad o sobliudenii prav cheloveka v Permskoi oblasti.* [Report on human rights situation in Perm *oblast*]. www.h-rights.ru/obj/doc.php?ID=150420

96. Popovich, Larisa D. 2000. *Opportunities for Differentiation in Health insurance Packages.* Moscow: Management of Social Security Development (TACIS Project).

97. Rashid, Salim. 1981. “Public Uttilities in Egalitarian LDC’s: The Role of Bribery in Achieving Pareto Efficiency.” *Κγκλοσ* 34(3):448-460.

98. Rizzo, J.A. and D. Blumenthal. 1996. “Is the target income hypothesis an economic heresy?” *Medical Care Research and Review*: 243-293.

99. Rose, Richard. 2001a. *A Decade of Change But Not Much Progress: How Russians Are Coping.* Glasgow: Centre for the Study of Public Policy (University of Strathclyde), Working Paper 349.

100. Rose, Richard. 2001b. *The impact of social capital on health.* Glasgow: Centre for the Study of Public Policy (University of Strathclyde), Working Paper 358.

101. Rose-Ackerman, Susan. 1975. “The Economics of Corruption”. *Journal of Public Economics* 4: 187-203.

102. Rose-Ackerman, Susan. 1999. *Corruption and Government: Causes, Consequences and Reform.* Cambridge: Cambridge University Press.

103. Rostov. 2001. Head of administraion of Rostov *oblast*. *Ob utverzhdenii territorialnoi programy gosudarstvennykh garantii [...]na 2001 god.* [Regional program of free care entitlement for the year 2001]. Decree 101 March 12.

104. Russia Longitudinal Monitoring Study. http:\www.cpc.unc.edu/rlms

105. Ryan, Michael. 1990. *Doctors and the State in the Soviet Union.* New York: St. Martin’s Press.

106. Ryan, Michael. 1994. *Social Trends in Contemporary Russia: A statistical Source-book.* New York: St. Martin’s Press.

107. Saint Petersburg. 2000. Public Healthcare Committee, City Administration. *O poryadke i usloviiakh organizatsii platnykh meditsinskikh uslug na gorodskoi stantsii skorooi meditsinskoi pomotschi* (Organization and conditions of paid services delivered by the city healthcare emergency center). Ordinance # 260-Ю October 4.

108. Saint Petersburg. 2001. Public Healthcare Committee, City Administration. *O predostavlenii plantykh meditsinskikh uslug naleleniuiu.* [Paid medical service delivery]. Ordinance # 225-п August 6.

109. Satarov, Georgy. 2002. *Diagnostika rossiiskoi korruptsii: sotsiologicheskii analiz* [Diagnostics of the Russian corruption: sociological analysis]. Moscow: INDEM Fund http://www.anti-corr.ru/awbreport/index.htm

110. Sergeev, N.V. and T.Yu. Sidorina. 2001. *Gosudarstvennaya sotsialnaia politika i zdorovie rossiyan. K analizu zatrat domokhozyastv na zdravookhranenie* [The social policy of the state and health of Russian citizens. Analysis of household expenditures on healthcare]. *Mir Rossii* 2.

111. Sheiman, Igor. 1998. *Reforma upravleniya i finansirovaniya zdravookhraneniya* [Governance and financing reform of healthcare]. Moscow: Rus'.

112. Sheiman, Igor. 1999. "Excessive State Commitments to Free Health Care in the Russian Federation: Outcomes and Health Policy Implications." In *The Right to Health Care in Several European Countries,* ed. Andre Den Exter and H. Hermans (eds). London: Kluwer Law International. 101-113.

113. Sheiman, Igor. 2000. *Vozmozhnye strategii reformirovaniaia systemy gosudarstvennykh obyazatelstv v zdravookhranenii* [Alternative strategies of reforming the system of state obligations in healthcare]. *Ekonomika zdravookhraneniia* 5:47-55.

114. Shishkin, Sergey V. 1998. "Priorities of the Russian health care reform." *Croatian Medical Journal*. Vol. 39(3):298-308.

115. Shishkin, Sergey V. 2000a. *Issues of Health Care Financing in Russia.* St. Petersburg: SPIDER Working Paper #16.

116. Shishkin, Sergey V. 2000b. *Reforma finansirovaniya rossiiskogo zdravookhraneniya* [Financing Reform in Russian Healthcare]. Moscow: TEIS.

117. Simek, Jiri. 1999. "The Right to Health Care in the Post-Totalitarian Czech Republic." In *The Right to Health Care in Several European Countries,* ed. Andre Den Exter and H. Hermans (eds). London: Kluwer Law International.

118. Sissouras Aris et al. 1999. "Health care and cost containment in Greece." In *Health care and cost containment in the European Union*, ed. Elias Mossialos and Julian Le Grand. Aldershot: Ashgate.

119. Starr, Paul. 1982. *The Social Transformation of American Medicine.* New York: BasicBooks.

120. Suk, I.S. 1984. *Vrach kak lichnost'* [Personality of the physician]. Moscow.

121. Sztompka, Piotr. 1999. *Trust: A Sociological Theory*. Cambridge: Cambridge UP.

122. TACIS. 1998. *Analysis of healthcare governance in the Russian Federation. General report*. Moscow: TACIS Russia Healthcare Management Project http://zdravinform.ru/pub/EU.1998.A.4.E.pdf

123. Tula. 1998. Governor of Tula *oblast. O pravilakh predostavleniia platnykh meditsinskikh uslug meditsinskimi uchrezhdeniiami oblasti* (Rules of delivery of paid medical services by medical care providers in the *Oblast*). Decree # 364 of

August 25.

124. Tymowska, Katazyna. 1999. “Changes in Access to the Health Care Services in Poland.” In *The Right to Health Care in Several European Countries,* ed. Andre Den Exter and H. Hermans (eds).. London: Kluwer Law International.

125. Volgograd. 1999. *Duma* of Volgograd *oblast*. Law 348-ОД November 25.

126. Voronezh. 1996. Department of social policy, Administration of Voronezh *oblast*. Ordinance 436 August 13.

127. World Bank. 1999. *Otchet po proektu Vsemirnogo Banka “Issledovanie rynka posrednicheskikh uslug po meditsinskomu strakhovaniiu.”* [Market Research of mediator services in medical insurance, Project Report]. Moscow: World Bank Russia.

128. Yaroslavl’.1999: Government of Yaroslavl’ *oblast*. *Ob utverzhdenii territorialnoi programy gosudarstvennykh garantii [...]na 1999 god.* [Regional program of free care entitlement for the year 1999]. Decree 29-п March 10.

129. *Yaroslavskaia oblastniaia klinicheskaia bolnitsa* [Yaroslavl’ *oblast* clinical hospital]. Web-site: www.yrh.yar.ru