2. REVIEW OF RELEVANT THEORIES

This chapter reviews the major theoretical arguments underlying economic organisation and their relevance to the vertical relationships between purchasers and providers in health care. It briefly reviews neoclassical economics and New Institutional Economics theories, and goes into more detail on Transaction Cost Economics (TCE), as it has been proposed as one of the rationales for the separation between purchasers and providers. It then reviews the topic of autonomisation of public hospitals from the perspective of TCE and other rationales. The chapter ends with a review of the theoretical models of hospital behaviour, and it highlights a major gap in research due to the lack of appropriate theoretical models, not only for hospitals in general, but for public hospitals in developing countries.

The review that follows focuses on NIE theories as the frameworks within which the data will be interpreted, as will be seen in chapter 5. These theories were selected because TCE was considered *the* adequate theory to analyse a phenomenon that was clearly related to transaction costs, i.e., a change in governance structure, or the shift from a hierarchical relationship to a contract-based one, in the expectation of improving quality and efficiency. Given that TCE is part of a larger set of theories that deal with relationships between parties, as shown in figure 2.1., it makes sense to analyze contractual relationships through the prism of these complementary theories. In fact, it is their complementarity which makes them valuable for analysing contractual relationships.

2.1. Neoclassical economics, complete contracts

The relationships between producers and consumers in perfectly competitive markets are characterised, among others, by the fact that all specifications about the exchange are implicit and the parties do not have to spend time or effort in dealing with uncertainty or information asymmetries, because these are absent.

However, perfect competition only partly explains the actual behaviour and performance of markets, not only because perfectly competitive markets rarely exist, but also because there are many other components of agent behaviour that are not captured in this neoclassical model of perfect competition, as discussed below.

Regarding the interaction between supply and demand, when transactions are completely standardised and there is no reason to suspect that one of the parties will not comply with expectations, no formal relationship between the parties is necessary and the parties do not need to agree in advance the terms of the interaction, let alone put an agreement in writing. Thus, the terms and conditions of a hypothetical contract are implicit and no written statement is necessary to make sure they will be satisfied. This type of relationship is known as a "spot contract" (Robinson, 1999, p. 70). In a spot contract type of relationship, the parties just decide on the spot, exchange money and goods/services on the spot, and the interaction ends. This is the quintessential discrete, anonymous relationship, with no costs associated with the transaction.

2.2. Agency theory

Beyond these discrete market exchanges that involve complete contracts, most relationships in the economy involve varying degrees of uncertainty and complexity, and spread over several periods. In this case, preserving the relationship depends on each party's future expectations and past performance vis à vis the other party. Thus, the parties are more likely to appeal to formal mechanisms to govern the relationship. In the presence of uncertainty about future costs and prices, quantities produced and demanded, and concern for opportunistic behaviour by either party, the supplier and the purchaser will try to reduce the risks associated with such uncertainty. Beyond the implicit contracts underlying the discrete exchanges of perfect competition, explicit contracts create the framework to guarantee that both parties comply with each other's expectations and uncertainty is reduced to the minimum possible.

However, uncertainty and information asymmetries still create problems for the contractual relationship. Agency theory provides a theoretical framework to address these problems. It starts by describing a relationship whereby an individual (the principal) engages someone else (the agent) to make decisions on his behalf, in exchange for a compensation. The basic problem is how to have the agent to maximise the principal's utility, if the agent has information advantages that can be used to maximise his own utility at the expense of that of the principal. Agency theory assumes that contingencies related to uncertainty and information asymmetries can be dealt with *ex-ante*, if the principal can arrange the adequate mix of incentives (Williamson, 1990).

If some contingencies cannot be solved *ex-ante* in the contract minute, an adequate allocation of property rights will solve the gaps, by aligning the agent behaviour with utility maximisation for the principal. And if any of the parties do not comply with contract terms, they can resort to the legal system to protect their interests in a court process. Agency theory refers both to market exchanges with principal-clients and to ownership relationships with principal-owners.

Regarding ownership arrangements, in the private for-profit sector, these agency relationships are clear: owners can assess the agent's performance through profits and stock price; despite the fact that ownership is often spread over thousands of shareholders, they all behave like a single principal because their objective function is the same: maximise the value of their property. But agency theory deals not only with relationships where the tasks can be clearly stated, the outputs and the objective function are observable and verifiable,[1] and the agent has a single principal. In the not-for-profit and public sectors the objective function is far from unidimensional. In addition, the presence of multiple principals with different goals and facing the agent with multiple tasks, makes it less likely to achieve a concrete outcome if the principals do not or cannot cooperate (Dixit, 1996).

Agency problems in the public sector are analysed by Tirole (1994), who argues that public sector organisations are characterised by having multiple goals, which are difficult to measure; the weights added to each goal cannot be clearly defined; although the community is the ultimate principal, it is diffuse, unorganised and shows various and changing preferences that translate into various and changing objective functions. In addition, intermediate principals like local politicians and the legislative bodies add to the lack of a clear and single-minded principal. The lack of a unified principal introduces problems of coordination. As the various principals cannot coordinate their emphasis on the tasks commissioned from the agent, and the tasks show differing degrees of observability and verifiability, the agent will put more effort on those tasks that are observable and verifiable, irrespective of how relevant they are. This multi-task agency problem is modelled by Holmstrom and Milgrom (1991).

[1] Observability and verifiability are two related but separate concepts. Observability refers to the possibility of making subjective assessments about the satisfaction of a given task, whereas Verifiability refers to the possibility of specifying such a task in a way that is enforceable in a contract (Chalkley and Malcomson, 2000).

This situation can be illustrated with the case of a public hospital manager, who deals with three principals: the community, the local politician and the health authority (Eid, 2001). Although maximising the community level of aggregate health seems a plausible objective function, it is less observable and verifiable than the politician's objective function of maximising constituency support, which, it can be argued, relates to employment opportunities for supporters at public organisations (Robinson and Verdier, 2003) e.g., the local hospital. Accordingly, lack of coordination between these two principals (the community and the local politician) will give managers room to select the task that better satisfies their objective function at the lowest effort.

Regarding the existence of multiple principals, and the exposure of the public official to capture,[2] Tirole (1994) also argues that if the manager or any powerful interest group is conferred the exercise of decision rights, it would open room for abusing power for private benefits. Accordingly, the public sector is characterised by a distribution of decision rights over various types of decisions, so that no one prevails and they are deliberately fighting each other for control of resources. This creates an environment of checks and balances that prevents abuse of power, but at the same time is at the root of the disfunctionality of public organisations. In fact, he argues that "No ministry's mandate is to maximise social welfare" but each ministry has to pursue its sector's goals. In addition, he posits that bureaucrats' exposure to regulatory capture and collusion is dealt with through rigid rules to avoid discretionary use of information. This rigidity also explains the lack of flexibility of public organisations as compared to private ones, to respond quickly to environmental challenges. As a consequence of these agency problems, besides other factors related to property rights that will be addressed below, public officers are usually exposed to low-powered incentives, and poor performance of public organisations is more likely to occur if no additional checks and balances and an adequate set of incentives are put in place.

One persuasive explanation about why some public organisations work better than others in a context of low-powered incentives is provided by Tirole (1994) in his argument about career concerns. The author proposes that when the agent faces inadequate incentives to carry out the relevant tasks, and the checks and balances do not work adequately, public officers can choose to put their effort on those tasks and

[2] According to Posner (1974), the concept of regulatory capture refers to the situation where "...regulatory agencies come to be dominated by the industries regulated."

outputs that improve the likelihood of achieving their long-term career goals. The outcome will not necessarily be aligned with the greatest social welfare but rather with short-term benefits that guarantee long-term effects on the officer's career. Although career concerns also play a role in for-profit firms, clearly it is much more salient in the public sector, due to the weaker effect of incentives and monitoring of inputs.

Another important implication of Holmstrom and Milgrom's (1991) multi-task agency model is that of contracting for health services where quality and cost cannot be simultaneously optimised through an incentive scheme. The problem arises because, given that quality is only imperfectly observable (because of information asymmetries), if the purchaser compensates the provider for costs incurred in order to assure maximum quality, the provider will have no incentives to control cost. On the other hand, if the purchaser gives incentives for the provider to control costs, given information asymmetries vis à vis the patient, the provider has room to cut costs by jeopardising quality. Tirole (1994) also points out that in the case of experience goods,[3] it is preferable to put low-powered incentives in place or else quality would be at risk. This problem will be addressed below in more detail, in terms of its implications for contracting health care services, where quality is partially observable and incentives for cost control can put quality at risk.

2.3. Property rights theory

Property rights theory argues that in the presence of incomplete contracts, the status of being the residual claimant allows for the optimisation of the outcome of a contract for the contracting parties, because all the efficiency gains accrue to the holder of the residual claims (Milgrom and Roberts, 1992). Property rights create the strongest incentives for good performance, not only because the residual claimant can appropriate any remaining gains after paying all debts and obligations, but also because in case of losses, they also have to be borne by the owner. Therefore, the owner faces a hard budget constraint, which will make him very careful about taking risks.

Regarding ownership, property rights mean that the owner of a given asset has the right to decide "all usages of the asset in a way not inconsistent with a prior contract, custom

[3] Experience goods are defined as those goods about which information on quality and prices can only be obtained through consumption. The other category is search goods, i.e., those goods for which quality and prices can be obtained before consumption. The terms were originally proposed in Nelson (1970).

or law" (Hart, 1995, p. 30). These are what Hart calls *residual control rights*, and takes as the definition of ownership. Being the holder of the residual control rights, the owner has an incentive to increase the market value of the asset, because "...this person can decide when or even whether to sell the asset" (Hart, 1995, p 65), and realise the long-term income of the asset.

The assumption that allocating property rights solves the *ex-post* agency problems rests on another more fundamental assumption that is taken for granted in competitive markets: the availability of a high level of information. At this point, property rights theory is inconsistent with agency theory because, according to the latter, information asymmetries allow the agent to apply a suboptimal level of effort to maximise the principal's utility, i.e., the value of property.

However, owners are not always the direct managers of their resources; instead, they commission an agent to take control of them. This separation of ownership and control seems to work fairly well in the private sector, where stock markets, the threat of a hostile takeover and the existence of a board of directors coalesce to deter managers from abusing their decision rights to their own benefit instead of that of owners (Fama and Jensen, 1983).

In the public sector, separation of ownership and control puts different challenges. By definition, public organisations do not have a residual claimant other than the community or the society at large. However, society or the community hardly work as a principal that acts in a coordinated way to hold managers accountable for maximising societal value. If managers cannot be the residual claimants of public organisations, they cannot get the benefits of good institutional performance but neither can they be held responsible for losses. A partial solution would be to allocate property rights to managers, so that the benefits of good performance yield appropriable rents to managers. However, this partial solution would create an incentive for excessive risk taking, because losses will not affect the manager in the same proportion (or may even be zero) as gains do. It would also create room for the managers to engage in self-satisfying behaviour. In addition, the lack of a single-minded principal makes it difficult to exert the ultimate decision right of a residual claimant, namely to transfer the property rights to another party.

2.4. Public choice theory

As said above, allocation of property rights does not yield the same results in the public sector as it does in the private sector. The fact that the ultimate owner of a public asset is the society makes it difficult to identify uniquely a single-minded principal that holds the agent accountable for performance. This lack of a clearly identifiable principal, in addition to the separation of ownership and control, and the lack of a unidimensional objective function, opens wide room for public officers to engage in self-satisfying behaviour.

Neoclassical economics simplified the functioning of the government by assuming that decision-makers would pursue the collective good by maximising social welfare, just following the orders of the legislators. This oversimplification has been challenged by public choice theorists, who claim that public officers are not necessarily the best guardians of the common good (Birdsall and James, 1993). In fact, this theory considers that they are also self-interested utility maximising individuals who use their public positions to increase the likelihood of being re-elected (in the case of publicly elected officers), or the size of the budget to increase personal rewards (in the case of appointed bureaucrats).

In a more specific view of merit goods, health care is seen as a proof of government's concern for the welfare of citizens (Hsiao, 2000). A direct implication of Hsiao's point for hospital policy is proposed by Healy and McKee (2002c), when they argue that hospitals can be seen as a key symbol of the survival of the welfare state. Accordingly, politicians, aiming at assuring re-election, will try to use health care spending as a proof of their concern for voters, and the possibility to close down a hospital is strongly resisted.

Another key element of public choice theory, that improves upon the limitations of the neoclassical view of government, is explained by Stiglitz (2000, p 169) with the concept of the "median voter," by which representatives try to respond to the median voter, i.e., that with the median income. After a sequential two-party game, both parties end up proposing the level of expenditure that satisfies the median voter, but given that his income falls below the average income (assuming income distribution is always skewed to the right), the median voter faces marginal benefits of additional expenditure that

exceed the marginal costs faced (i.e., the tax price, assuming a progressive or proportional taxation scheme). As a consequence, a democratic process of budgeting will always result in an amount of public expenditure that is larger than the socially optimal. This point may be useful to explain the persistence of inefficient structures in the public sector.

Another explanation for the inefficiency of the public sector is what has been called the time-inconsistency problem, which takes place when the government's optimal short run policy is not aligned with its long-run optimal policy. Majone (2001) appeals to this problem to explain the existence of autonomous institutions like central banks, which define long-term policies that would not be taken by politicians with short-term minded agendas.

However, collective decision-making does not always reflect consensus among voters, because politicians represent varying and often conflicting interests, and because special interest groups can have better resources to have politicians favor their particular interests. This competition between interest groups to capture a surplus of the public purse is also known as *rent seeking* (Krueger, 1974). Public choice theory also addresses these problems, emphasising the self-interested utility-maximisation behaviour of politicians. This additional source of inefficiency tends to favor the better off, and their probability of being favored is *ceteris paribus* inversely related to the number of constituents of the interest group, because of coordination costs.

2.5. Transaction cost economics theory

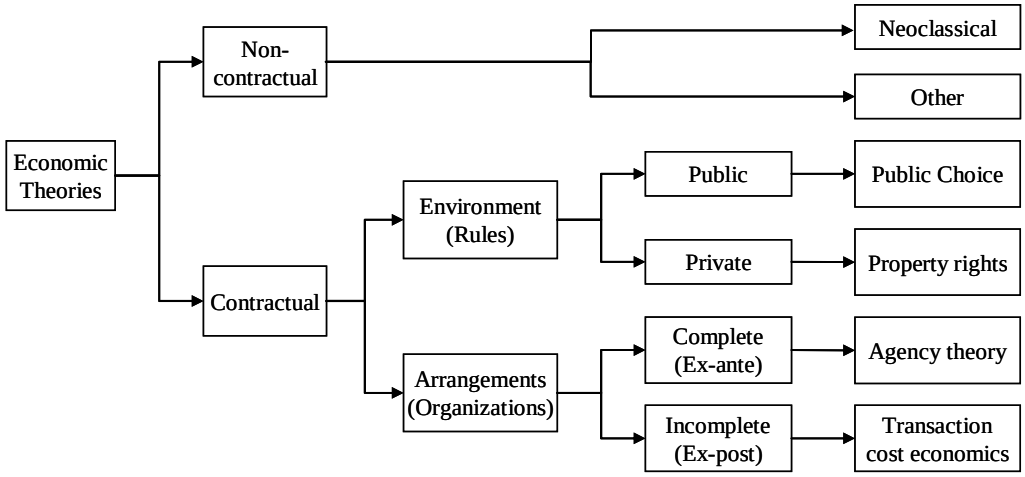
The preceding summary of neoclassical and neo-institutional economics theories has been deliberately approached through the prism of contracts. This deliberate restriction to contracts serves the purpose of providing an adequate background to understand Transaction Cost Economics (TCE). In fact, TCE addresses the various types of economic organisation by focusing on how contractual relationships take place among the different parties. Thus, the unit of analysis of TCE is *the* relationship, and economic organisation is explained according to how these relationships are governed.

Williamson (1990) lays out a summarising view of the economic theories of organisation that is useful for comparison purposes and to show what part of the

analysis is addressed by transaction cost economics (see figure 2.1.). He divides theories into noncontractual and contractual ones. The first category includes neoclassical economics and what he calls "other theories". The contractual theories are separated into those related to the environment of firms, and those related to the arrangements between organisations. The former entails the rules underlying the interactions between firms. Public choice theory involves rules mostly shaping the public sector, whereas property rights theory involves the legal frameworks that shape the private sector. Regarding the theories related to arrangements between organisations, Williamson classifies them according to how uncertainty and complexity are dealt with. Agency theory addresses the issue of how to deal with them *ex-ante* by aiming at a complete contract, whereas TCE focuses on how to deal with uncertainty and complexity when complete contracts cannot be arrived at.

Figure 2.1.

A taxonomy of the economic theories of organisation, according to Williamson (1990)



Along the lines of Williamson's taxonomy, it could be said that neoclassical economics assumes complete and implicit contracts that allow firms and individuals to engage in discrete and anonymous exchanges without much concern for uncertainty and complexity. Evidently this is not the case in many instances; thus, firms and individuals aim at longer-term relationships and seek instruments to deal with uncertainty and complexity, giving rise to the emergence of a world of contracts. Contract-based interactions require a framework to protect property and to enforce contracts (property

rights theory) but these interactions are also shaped by collective decision-making processes (public choice theory). Property rights and collective decision-making create an environment within which firms and individuals interact through contracts.

Uncertainty and complexity can be addressed by arranging an adequate set of incentives, whenever the principal is able to observe and verify the agent's performance. In this case, contingencies can be dealt with *ex-ante* in a contract, which is said to be complete because all future contingencies can be solved by going back to the contract minute. These are the basic pieces of agency theory. However, not all the times is it possible to reduce uncertainty and complexity by aiming at a complete contract, and it would be very costly to do so. Agency theory assumes that *ex-post* inefficiencies are reduced through the allocation of property rights to the agent, but TCE argues that this is not always the case and *ex-post* inefficiencies still persist. Consequently, it is necessary to create the structures and tools to deal with uncertainty and complexity *ex-post*, whatever the level of incompleteness of contracts. This is the focus of TCE as will be shown below. The next section describes the TCE framework starting from the description of the vertical chain of production and the relationships between the links in the chain. Then, the origins of transaction costs and the structures to govern these costs are described.

2.5.1. The vertical chain of production

The vertical chain of production is the sequence of transformations of raw materials, through different production processes and intermediate products, into final goods for consumption. The relationships between the links of the vertical chain have been addressed by several authors, starting from the seminal work by Coase (1937) on the firm as an alternative arrangement to govern transactions. But it was Williamson (1985) who convincingly summarised and improved the previous reflections on the degrees of integration or non-integration of the links in the vertical chain. He concisely sets a simple but striking question: "why can't a large firm do everything that a collection of small firms can do and more? (p. 131).

According to Williamson, the relationships between any two successive links of the vertical chain of production vary between the extremes of vertical integration and spot contracting. The former entails a bureaucratic arrangement in which the two links are

part of the same organisation, and decisions are dealt with through administrative orders; the latter entails the type of market relationships that are assumed in the neo-classical model of economics, namely those typical of a perfectly competitive market.

However, the spot contracting extreme is the exception; some degree of interaction between the trading parties is expected to take place through contracting processes, although this interaction is far from perfect, as seen above. Williamson argues that two basic behavioural assumptions underlie the imperfect contractual relationships between supply and demand: bounded rationality, i.e., the individual's inability to fully understand the complexities of a decision to buy or sell, and opportunism, i.e., an individual's proneness to take advantage of the other party's weaknesses. Besides these two behavioural assumptions, the following factors make it very difficult to undertake a costless transaction:

- Uncertainty about future prices and quantities of goods exchanged, or the prices of their input/output markets.
- Complexity of the products, which in turn creates information asymmetries between trading parties and makes it more difficult to:
    - Precisely define the product.
    - Observe it when delivered, so as to check for fulfilment of contract terms.
    - Verify it when disagreements arise.

Given these factors, it is very difficult to write a complete contract between supplier and buyer, in which all possible contingencies are taken *ex-ante* into account. The most commonly analysed attributes of contractual incompleteness are those related to prices and quantities, as shown, for example, in Crocker and Reynolds (1993) and Saussier (2000). Complexity, unobservability, unverifiability or difficulty to define the product, are less commonly considered as attributes of contract incompleteness.

The key factor in making spot contracting unlikely, according to Williamson, is the presence of Relationship-Specific Investments (RSI),[4] that is to say, investments whose net present value outside the relationship is lower than the opportunity cost of capital.

[4] They are also called specific assets, but the term Relationship-Specific Investments is more descriptive, and will accordingly be used in this thesis.

Assume, for the sake of the example, that A and B are two parties that are about to enter into a trading relationship that requires a RSI. A is the party incurring RSI, and B is the other party. Given such asset specificity, A expects a rate of return higher than the opportunity cost of capital. A quasi-rent is created, i.e., the difference between the net present value of the RSI within the relationship, and that of its second-best alternative, outside the relationship. The larger the quasi-rent (i.e., the lower the value of the second-best alternative) the more likely B will be tempted to renege the contract in the future and renegotiate it with the purpose of extracting the quasi-rent from A. This is because A will be better off accepting a rate of return at least higher than that of the second-best alternative for the investment. This is known as the hold-up problem (Goldberg, 1976) by which the party that incurs the RSI is exposed to the risk that the other party takes advantage of its stronger bargaining position and extracts the quasi-rent (Monteverde and Teece, 1982). The crucial role played by the act of investing in RSI has led Williamson to call this event the "fundamental transformation" of the relationship, by which an *ex-ante* competitive bargain becomes a bilateral monopoly (Williamson, 1976).

Williamson (1985, p.95) sets four types of asset specificity: 1) physical asset specificity entails relationship-specific equipment or machinery; 2) site specificity entails a "cheek-by-jowl" relationship to minimise transportation and inventory costs; 3) human asset specificity entails a learning-by-doing process; and 4) dedicated assets entails investments that are made on the prospect of selling a significant amount of product to a client. Masten et al (1991) add a fifth type: Time specificity, which entails assets whose precise timing in the production process cannot be altered, lest the purchaser incurs large losses.

In order to reduce the risk of being held up, A will try to safeguard its investment by aiming at a contract as complete as possible. But given diminishing marginal gains and increasing marginal costs in reducing uncertainty through aiming at a complete contract, there is a point beyond which it is not desirable to keep aiming at completeness (Crocker and Reynolds, 1993). Given that it is not possible to write a complete contract, A would expect a commitment from B that it will not take advantage of the risk of hold up to which A is exposed. Such commitment could be expressed as a risk premium or a hostage, which, in practical terms, is an additional cost to the transaction. Nonetheless, A still has the option to underinvest in RSI (Lyons, 1994) in order to

reduce the loss in case of being held up by B. On the other hand, A can also hold B up, by reneging on the contract and increasing the price, so as to extract B's quasi-rent (Crocker and Reynolds, 1993). Thus, the interaction of contract incompleteness and asset specificity is the key factor in making it difficult to opt for a pure spot contract relationship as the only alternative to a hierarchical relationship (Joskow, 1988). Addressing vertical relationships through pure spot contracts, on the expectation that disputes will be solved in the courts, would make it prohibitively costly to start relationships with high levels of complexity and *ex-post* uncertainty. Therefore, appeal to courts is reserved as a resource of last resort.

An additional variable that is given emphasis in Williamson's framework is the frequency with which the transaction in question is undertaken. In the presence of RSI and/or contract incompleteness, a more frequent transaction is more likely to be vertically integrated than mediated by a contract. Frequency is not in itself a source of transaction costs, but a covariate; that is to say, it is its interaction with uncertainty and asset specificity that counts.

From a legal perspective, MacNeil (1974) points to the need for flexible relationships that make adaptation through time possible; his classification of contracts into classical, neo-classical and relational is based on uncertainty and the difficulty in "presentiating" all the possible contingencies and making relations discrete (MacNeil, 1978). Classical contracts are those akin to spot contracting, where the concept of "sharp in by clear agreement, sharp out by clear performance" (MacNeil, 1974, p 738) holds unquestioned. Neo-classical contracts are those where room for flexibility to cope with uncertainty is created *ex-ante*, while relational contracts are those where there is such a large degree of uncertainty that the only thing the parties can agree is to develop a relationship to deal with future contingencies, something like "an agreement to agree." This concern for the role of uncertainty is similar to Williamson's, but the latter's improvement of the theory by the addition of the role of RSI gives a more consistent theoretical framework for the understanding of the vertical relationships in the chain of production.

2.5.2. Hierarchies vs. markets

Within the continuum between hierarchical and market-based arrangements, every type of relationship between any two links of a chain of production has intrinsic transaction costs. On the side of hierarchical relationships, the low-powered incentives inherent to salaries induce the agent to limit effort to maximise the principal's goals. Influence costs are also present in hierarchies, when bureaucrats or internal interest groups use political manoeuvring instead of efficiency and quality as a strategy to obtain preferred treatment in internal budget allocation decisions (Milgrom and Roberts, 1990, cited by Harding and Preker, 2003). This opens room for internal divisions to demand additional resources, which in the public sector is particularly complex, because divisions face soft budget constraints. Diseconomies of scale are also present in large bureaucracies, when an individual enjoys information surpluses that can be used to her own advantage and not to that of the organisation (McAfee and McMillan, 1995). When these costs of hierarchies outweigh the benefits of in-house production, the theoretical framework of TCE provides a strong rationale for the separation of components of a vertical chain of production (Williamson, 1985).

On the side of market-based relationships, the presence of contract incompleteness and RSI are a source of transaction costs. These costs are: 1) the costs of writing, negotiating, monitoring and enforcing contracts, 2) the costs of renegotiation of contract terms when conditions change, 3) the costs of underinvestment in RSI on the prospect of hold up, 4) the costs of credible commitments or risk premiums by one of the parties to stimulate the other to invest in RSI, and 5) the costs of distrust that lead the parties to a) avoid sharing information useful for quality- and efficiency-improving purposes, and b) make redundant investments to protect themselves from being held up (Williamson, 1985). Figure 2.2. shows a schematic view of the components of contract incompleteness and RSI and their consequent transaction costs in market-based relationships.

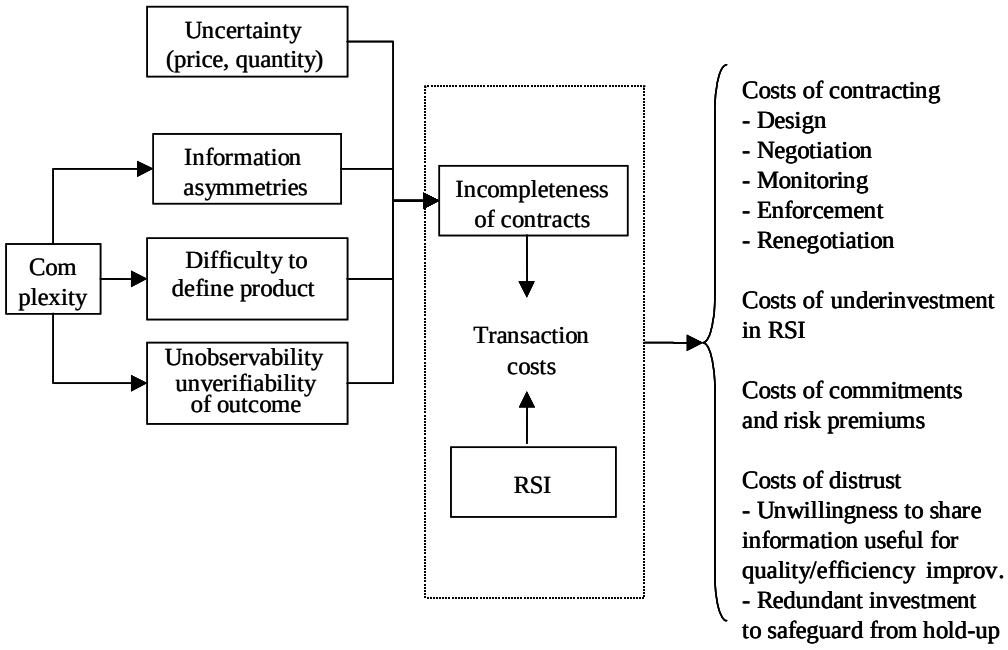
It is then expected that in the presence of contract incompleteness and/or RSI, a frequent relationship between purchaser and provider will not be of a spot contract type. Instead of that, the relationship has to be governed by the parties, so it will evolve towards a specialised governance structure that minimises the transaction costs derived from contract incompleteness and RSI. These specialised governance structures can be short-term contracts when RSI and contract incompleteness are not very relevant. Or they can be of a long-term type or exclusive contract when there are important RSI and contract

incompleteness. In the case of extreme uncertainty, the relational type of contract proposed by MacNeil is more likely to be present, and in the presence of *ex-ante* bilateral monopolies it is more likely that contracts are inevitably exclusive.

Figure 2.2.

Transaction costs of market-based relationships

Within a context of bounded rationality and opportunism, contract incompleteness and RSI create transaction costs



Here is where TCE shows its robustness as a theoretical explanation of the vertical boundaries of the firm; its central hypothesis, as stated by Williamson (1993), is: "Transactions, which differ in their attributes are aligned with governance structures, which differ in their costs and competencies, in a transaction cost economising way." Nevertheless, this statement rests on the assumption that firms are profit maximisers or cost minimisers, which is not necessarily true for other types of organisations, for instance, hospitals.

Distrust, as a key source of transaction costs, can be analysed from other perspective though. Klein (1997) shows that promises between contracting parties usually work, no matter that contract incompleteness is pervasive. He refers to an "invisible eye" as a

complement to Adam Smith's invisible hand. This is an important factor that forces individuals to act correctly. Indeed, such correct behaviour has important implications for the reputation of the parties, which in turn reduces transaction costs without appealing to vertical integration as shown by Holmstrom and Roberts (1998) in the case of Japanese industry. In that country the risk of hold-ups due to the presence of RSI and overt contract incompleteness is not reduced through vertical integration but through the reputation of the trading parties that has been built during years of relationships. In fact, the relationships become long-term and more akin to relational contracts, which means these governance structures are obvious substitutes to vertical integration. A direct application of the role of trust as a strategy to reduce transaction costs in the NHS is carried out by Goddard and Mannion (1998). The authors show that no matter purchasers and providers are locked in to contract with each other, trust arises as a response to problems of contract incompleteness and RSI.

2.6. The vertical chain of production in health care

The analysis of the vertical chain of production in health care requires an accurate description of the core functions that make up each of the links of this chain. The framework proposed by WHO (2000) gives normative conceptualisation of a vertical chain composed by the functions of: 1) resource generation, 2) provision and 3) financing, with a stewardship function that involves all the other functions. The financing, provision and resource generation functions are arranged in a way that resembles more closely a chain of production, while the stewardship function is not in itself part of the chain. Regarding the financing and the purchasing function, these show diverse degrees of integration within public health systems or a public-private mix (Murray and Frenk, 2000). Regarding the provision function, the most clearly identifiable part of the vertical chain, it is separated between personal and non-personal services; the former are more easily related to the vertical chain in which producers sell health care services to purchasers (insurers) and buy inputs from resource generators. Regarding the resource generation function, it corresponds to the upstream providers of inputs, namely, manpower, equipment and consumables.

2.6.1. Transaction costs in the vertical chain of health care

It is thus clear that the hospital, as a component of the vertical chain of production of health care services, uses inputs (consumables, manpower, medications, etc) to produce outputs (discharges, days of stay, surgical procedures, etc) that are sold to downstream purchasers (the government, insurers or households). The characteristic role of the physician as the decision maker about services to be provided by the hospital makes the vertical chain of production rather unique (Robinson, 2001a), as compared to most sectors of the economy where relationships are bi-partite. The relationships between the hospital as a supplier and the government and insurers as purchasers show diverse degrees of integration, and these variations can be assessed from the perspective of TCE. It could be hypothesised that, in the presence (absence) of contract incompleteness and RSI, the likelihood of vertical integration between purchaser and provider is higher (lower), in order to economise on transaction costs.

But the particular features of the health care industry warrant a more detailed analysis of the sources of transaction costs. The presence of RSI is considered a necessary condition to raise transaction costs, whereas contract incompleteness is considered a complementary condition. This is clearly stated by Williamson (1985, p 56): "Just as the absence or differential risk aversion would diminish if not vitiate much of the recent incentive work on contracting, so would the absence of asset specificity vitiate much of transaction cost economics." This strong emphasis remains largely unmodified today, as evidenced from Joskow (2005). It could be argued, though, that it is a common finding in other industries, where information asymmetries are not large enough as to cause concern for transaction costs by themselves. However, the implications of information asymmetries in health care are much larger than in other industries (Dranove and Satterthwaite, 2000), and the question remains as to what extent they are large enough to cause relevant transaction costs even in the absence of RSI.

Arrow (1963) highlighted the nonmarketability of risks associated with health care as one of the reasons for market failures in this industry. This nonmarketability stems from the fact that uncertainty involves not only the probability of occurrence of disease or trauma, but also the amount of resources that will be consumed to restore the patient's health. But the decisions regarding resource consumption are made by physicians, whose information advantage opens room for wide variations in practice styles (Wennberg et al, 1973). When resource constraints force purchasers to reduce unnecessary care but physicians do not face those constraints directly, it is likely that

information asymmetries will make it difficult for the purchaser to restrict physician's decision-making with regards to unnecessary care. It would then be very costly for the purchaser to collect the necessary information to bridge the asymmetry, and it is impossible to contract *ex-ante* the mechanisms to deal with the uncertainty associated with it. Thus, it could be hypothesised that if uncertainty has to be dealt with *ex-post*, and if it is very high, then it is a relevant source of transaction costs (irrespective of the presence or not of RSI) that determines the governance structure that minimises them.

2.6.2. Property rights and agency approaches to the vertical chain of health care

Other authors have addressed the issue of vertical relationships between purchasers and providers, particularly regarding the government as a purchaser of health care services, from the perspective of property rights theory. Hart, Shleifer and Vishny (1997) develop a theoretical model to analyse the purchase of public goods by the government, and propose that the allocation of residual decision rights on the provider side has key implications for the outcomes of the government-provider relationship. The problem underlying their analysis is that of multi-task agency, posed by Holmstrom and Milgrom (1991), particularly the case where cost control is achieved at the expense of quality. Hart, Shleifer and Vishny conclude that, given difficulties in defining quality, which are at the root of noncontractibility of quality and cost innovations, private contractors are more likely to devote excess efforts to reduce costs, even at the expense of jeopardising quality.

On the opposite side, public providers within a hierarchical relationship will be less likely to compromise quality on behalf of cost control, but will also have less incentives for efficiency. However, Hart, Shleifer and Vishny conclude that when patronage and powerful union groups limit decision makers' efforts to optimise quality and costs, it is better to privatise. In addition, when noncontractible factors like quality and cost innovations have a minor effect on outcomes, it is also better to contract the activity with private providers. These conclusions raise a paradoxical situation for public hospitals: if powerful unions and patronage make a vertically integrated hierarchy less desirable, it would be better to contract out. But contractors are more prone to cost cutting and compromising quality. Thus, both alternatives (vertical integration or

contracting out) would lead to poor quality, whereas vertical integration would lead to higher costs.

Chalkley and Malcomson (1998) analyse this agency problem in similar terms, but they separately address the effects of having three types of providers: a purely selfish, a purely benevolent, and one in between. They conclude that in the first case the government cannot achieve both quality and cost-reducing effort above the minimum contractible level. In the case of the purely benevolent provider, both objectives can be achieved with a block grant contract. In the intermediate case, including some degree of cost reimbursement as an incentive to improve quality can be optimal, at least if, at the margin, the disincentive for efficiency does not outstrip the incentive for better quality.

Eggleston and Zeckhauser (2001) use a model similar to that of Hart, Shleifer and Vishny, but with more precise applications to contracting out health care services to private providers. They add to the model the effect of competition, incentives of payment mechanisms, and soft budget constraints, as well as the use of non-profit providers as an alternative to contract with, besides private for-profits. They acknowledge that services contracted vary in characteristics, which make some services, e.g., dental care, more amenable to contracting out than others, e.g., mental health care. Although their analysis is restricted to for-profit and non-profit providers (an addition to Hart, Shleifer and Vishny, who only consider private for-profit providers), they consider public providers rather briefly in their analysis.

Regarding the effects of payment incentives, Eggleston and Zeckhauser suggest that private providers will exhibit stronger responses as compared to public and non-profit providers to the incentives inherent to payment mechanisms, namely, demand inducement as a response to fee-for-service (FFS) payment, and skimping on care/cream skimming as a response to prospective payment. With respect to unobservable quality, they predict that private providers are more prone to exploit user's perceived quality, but underperform in the less observable dimensions of quality. A last point raised by Eggleston and Zeckhauser is the role of soft budget constraints. Given government's commitment to be the provider of last resort, and managers' lack of residual control rights, particular challenges of soft budget constraints will arise in vertically-integrated public health care services. This gives for-profit and non-profit providers a comparative advantage as they face hard budget constraints, although chronic underfunding of public

networks (as has been the case in Eastern Europe countries) can exert a much stronger effect to keep overall spending low.

Both Eggleston and Zeckhauser, and Hart, Shleifer and Vishny approaches miss an important point regarding autonomous public hospitals. The point is that no matter they are autonomous and engage in contractual relationships with purchasers, autonomous hospitals are still public organisations. In this sense, Bech and Pedersen (2005) go a step forward to propose the specific characteristics of contracting with public providers that have diverse degrees of autonomy. They point out that the choice of payment mechanism is a sort of governance structure that aims at reducing transaction costs between the public purchaser and the public provider. Thus, payment mechanisms are not just a family of strategies to transfer money to providers in exchange for a basket of health services. Line-item budgets are the most hierarchical type of governance structure, whereby the government defines the hospital production function. Global budgeting is less hierarchical, whereas per-case payment is more market-like.

2.7. Rationales for hospital autonomy

Granting autonomy to public hospitals can be justified from the point of view of transaction costs associated with vertically integrated structures. However, there are other rationales for vertical dis-integration that are not based on TCE arguments but on other more pragmatic reasons. This section reviews the TCE-related rationales and two other relevant rationales: the need to mobilise resources to ease the fiscal pressure on central budgets, and the need to shift expenditures from cost-ineffective hospital care to cost-effective primary care interventions.

2.7.1. TCE-related rationales

As shown above, when the costs of a hierarchical arrangement are higher than the gains of integrated production of any two successive links of the vertical chain of production, it makes sense to separate them and shift to a market-based relationship. Accordingly, the previously mentioned problems pervading vertically integrated structures in public health care networks are argued to justify the separation of the producers of health care outputs from downstream purchasers (WHO (2000), World Bank (1993), Preker and

Harding (2002), Mills (1997)); this entails the shift from a hierarchy-based relationship to a market-based one, or at least a contract-based one.[5]

Taking some elements of TCE, the New Public Management (NPM) approach has advocated the separation of the purchasing function from the actual provision of services for the public sector in general. This entails the introduction of contracting, either competitive or non-competitive, between these two parties. Such a separation, as predicted by TCE, would lead to improvements in efficiency and quality if it is true that inefficiency pervades vertically integrated public enterprises and the transaction costs of contracting are not higher than the inefficiency costs of bureaucracy (Walsh, 1995).[6]

The separation of providers and purchasers is known in general as the Purchaser-Provider Split (PPS), and has been implemented in many other sectors with a strong influence of the NPM approach. In developing countries, the PPS has been promoted as part of the structural adjustment programs that have been applied to them (Batley, 1999). A major application of the PPS in health care has been the transformation of public hospitals into autonomous entities, with varying degrees of decision rights regarding budget setting, utilisation of user-fee revenues and expenditures on inputs and human resources. This makes the concept of autonomy a continuum rather than a discrete status, because there are differing degrees of decision rights across each of the fields of hospital management (Collins et al, 1999).

It is important to make clear the difference between contracting out and the PPS. Contracting *out* refers to a given organisation signing a contract with an outside upstream provider or downstream distributor, or an organisation closing a vertically integrated division and starting a contract with its equivalent in the market. It also can be applied to the government, when it closes one production unit and contracts it with an outside producer, or when it wants to increase output but does not want to increase capacity. By contrast, the PPS refers to a government shifting from a hierarchical relationship with providers to a contract-based one. Therefore, the PPS is different from contracting *out* because it does not imply a competition-based approach to contracting,

[5] As will be shown below, the separation of purchaser and provider can result in a bilateral monopoly, where no market exists at all, and the only difference is the presence of a contract between both parties.
[6] Although Walsh only briefly mentions the role of transaction costs (see p. 32), it can be inferred that the optimism of NPM towards the separation of purchaser and provider and contracting-out rests on the unproven assumption that transaction costs must be lower as compared to the vertically integrated governance structure. Thus, the search for a different governance structure goes in line with TCE.

and it has profound implications for the contract-based relationship in terms of exit options (Goddard and Mannion, 1998). It is thus adequate to refer to these two processes as contracting *in* (the PPS) and contracting *out*. The internal market reforms of the UK are an example of contracting in.

Granting autonomy to hospitals implies a shift away from an administrative unit within a hierarchical organisation, toward an independent unit where managers can manage their hospitals. It also implies a shift of day-to-day decisions to their locus (Harding and Preker, 2003). Although some commentators argue that the PPS is a step towards backdoor privatisation of hospitals (Armada et al, 2001), it does not necessarily mean privatisation but the shift from a hierarchical relationship to a contract-based one, while keeping public ownership and orientation.

A key feature of hospital autonomy that is rarely pointed out is the shift in the ways the government controls the hospital. In the vertically integrated structure, control is exerted through inputs, via budgets and plans, but there is little concern for output. Conversely, in the purchaser-provider split, the government is supposed to shift from an input-focused to an output-focused type of control. It implies that the hospital will be controlled via its ability to negotiate contracts for delivering health care services (Suriyawongpaisal, n.d.)[7]

Autonomisation has to be differentiated from corporatisation. Whereas autonomous and corporatised units retain their public ownership characteristic, a corporatised unit faces a hard budget constraint and full financial accountability while fulfilling social and public obligations (Harding and Preker, 2003). Eid (2001) calls it a "hybrid organisational form" that displays characteristics of both private sector and public sector organisations.

However, in terms of the TCE framework, any governance structure that falls between the two extremes of vertical integration and spot contracting is a hybrid organisational form. Accordingly, hospital autonomisation is a halfway departure from vertical integration towards spot contracting, i.e., a hybrid organisational form. This hybrid feature allows for the coexistence of market-based relationships mediated by contracts,

[7] The paragraph is based on the opinion of McPake, Hanson and Lake included in a report by Suriyawongpaisal for a roundtable on hospital autonomy in Thailand.

with hierarchical relationships mediated by internal communications or ownership structures.

From a general approach to transaction costs, Frant (1996) argues that the application of TCE to the public sector is made difficult because this theory has been developed for the private sector. To make his argument, he proposes that the equivalent to profits in the public sector is politicians' desire for reelection. Thus, if a given government function has information problems, exposing it to politicians' control would result in an undesirable result due to the unchecked effect of the high-powered incentives of reelection, as much as the high-powered incentives of profits do in the private sector. The solution to avoid the effects of high-powered incentives in the private sector is to abolish profits, so non-profit organisations emerge. Frant proposes that the equivalent in the public sector is to de-politicise the government function. Although Harding and Preker (2003) favor hospital autonomy as a strategy to avoid influence costs, which seems to be in the lines of Frant's argument, it remains to be seen if politicians' undue influence is adequately restricted in the PPS.

2.7.2. Non TCE-related rationales

Besides the TCE-based rationale for hospital autonomisation of the NPM approach, two different rationales have dominated the wave of hospital autonomisation and corporatisation reforms in the developing world: the need to mobilise resources through user charges in order to reduce fiscal pressures on the central government, and the need to concentrate expenditures on the guaranteeing of a cost-effective basic benefit package of health care interventions (Govindaraj and Chawla, 1996)[8]. From a critical point of view, Polidano (1999) argues that autonomisation and decentralisation in developing countries have not been preceded, as was the case in developed countries, by a strengthening of central government's capacity to perform the stewardship function. Thus, Polidano argues that this leap forward is more the result of pressures to reduce public spending. On this lines, Castano et al (2004) suggest that "…it could be argued that autonomisation is more an act of de-regulation by which the government capitulates on its responsibilities with social functions in order to respond to external pressures for

[8] See also the examples of Malawi (Shehata and Cripps, 2000), Jordan (Banks et al, 2000),Thailand (Charoenparij et al, 1999), Kenya (Collins et al, 1999), Indonesia (Lieberman and Alkatiri (2003) and Bossert et al (1997)), and Abrantes's (2002) analysis of the Chile, Argentina and Uruguay cases. All of them underline the importance of budget constraints at the national level as a major reason for hospital autonomisation.

fiscal balance." Hanson et al (2001) also point to the fact that government's weak regulatory capacity opens room for hospitals in developing countries to engage in cream skimming behaviour to the detriment of the poor.

Regarding resource mobilisation, Low Income Countries (LIC) with stringent fiscal pressures face hospital financing as a heavy burden on central government budgets. Thus, prompted by multilateral agencies to reduce the fiscal deficit, governments in these countries have resorted to user fees and cost-recovery fees as an additional source of hospital financing, particularly for tertiary care centers (Batley, 1999). However, a given hospital has no incentive to collect these fees unless it can retain them and make a surplus for its own benefit. If collected surpluses were transferred to underfinanced hospitals, it would discourage collection of fees and no resource mobilisation would be possible. Therefore, if the government wants to mobilise resources it has to give hospitals the incentive to collect user fees and cost-recovery fees, and the best incentive is to grant autonomy to the hospital to use surpluses as it sees fit.

Regarding the allocation of resources to cost-effective interventions, it is widely recognised that tertiary care does not have the same impact on a country's burden of disease, when compared to primary care and public health interventions. However, the fact that hospitals consume a large share of health care budgets, despite the availability of more cost-effective interventions, reveals the high political visibility of public hospitals (McPake, 1996). Given this political stance, it is hard for central governments to abandon public hospitals to redirect spending towards cost-effective interventions, so granting them autonomy seems to be an alternative in the middle.

In summary, hospital autonomisation in developed countries is grounded on the TCE rationales that underlie NPM-type policies. In Low-Income Countries (LIC), hospital autonomy is grounded on more pragmatic rationales, namely resource mobilisation and cost-effective spending. In contrast, it could be argued that Middle Income Countries (MIC) seem to be in the middle of this spectrum. Although financing hospitals in MIC also entails a heavy burden on public budgets, fiscal pressures are not as tight as in LIC. Thus, granting autonomy to hospitals is a plausible strategy which is expected to yield efficiency gains that will eventually reduce the fiscal burden to the central government and at the same time yield quality gains.

2.8. Potential drawbacks of hospital autonomisation

Castano et al (2004) summarise a set of potential drawbacks that would result when a vertically integrated public network of hospitals is converted into a dispersed group of autonomous hospitals. The most relevant drawbacks are briefly reviewed in this section. Regarding resource mobilisation, when hospitals have the incentive to collect fees, they will put more effort to attract paying patients at the expense of the poorest patients; this raises equity concerns that can be politically unsustainable (Kamwanga et al, 2003).

Equity concerns also relate to the amount of uncompensated care that the hospital has to provide. Depending on how the contracts specify the relationship between payments and outputs, public hospitals that operate as the safety net for the poor are more or less likely to provide uncompensated care (McPake et al, 2006). The emergence of a contractual culture that makes hospital managers more conscious about reimbursement would cause denials of care to the poor if no payment is received in exchange. No matter a basic-benefit package is created to explicitly state what services will be subsidised, denying care for benefits outside the package is very hard and politically costly. This hurts particularly the poor, and the hospital will be in a complex blind alley whereby it prefers to close down services that are more likely to attract uncompensated care (Bazzoli et al, 2005). On the opposite side, it could be argued that a large share of uncompensated care can be taken by a given hospital manager as a justification to evade responsibility for poor performance, because responsibility for not paying for that care is outside the realm of the manager's autonomy and in the hands of the health authority.

Regarding accountability, the lack of an identifiable principal and a single-minded objective function that characterises the public sector opens room for the manager of an autonomous hospital to engage in self-satisfying behaviour (Eid, 2001). It is thus necessary to create accountability devices to make sure this will not happen; market and contract accountability are not enough by themselves. A board of directors emerges as a key accountability device, although it has the same limitations derived from the lack of a clear principal and objective function. Since the board can be captured by interest groups within or outside the hospital, it does not guarantee that it will pursue the best interest of society at large, or ensure the manager is accountable for the maximisation of such social welfare.

A third potential drawback of autonomisation is the loss of coordination devices that work more easily in a hierarchy. One of these devices is the ability to work as a network of providers that exploits economies of scope and complementarities between them to provide seamless care. Autonomous hospital managers can be tempted to increase capacity or upgrade technology, which can result in redundant supply and either idle capacity or excess demand inducement (Arroyo, 1999). It is also possible that hospitals that are more successful in mobilising resources will attract the most qualified staff, exerting a crowding-out effect on the less successful hospitals. They can also close down unattractive services, either those that are reimbursed at a loss or those that attract patients for uncompensated care. This creates access barriers within the safety net of public hospitals.

Coordination costs are also related to lost economies of scale, specifically with regards to procurement, labour union negotiations, high-specialty care, and risk pooling. The possibility to re-unite hospitals to take advantage of these economies of scale depends on a hospital manager's willingness to engage in cooperative strategies, or on a health authority's ability to create the incentives for them to cooperate.

Another aspect of coordination regards the bottom-up flow of information for policymaking. It is possible that a manager of an autonomous hospital finds it strategically optimal to retain information that increases bargaining power vis à vis the purchaser, just as pointed by McAfee and McMillan (1995) for large organisations. A weak oversight capacity at the local authority would make it unrealistic to resort to contract clauses to force the hospital to provide this strategic information. Information problems also arise when purchasers rely exclusively on "hard" data that can be potentially "massaged" by providers to satisfy the demands of the purchaser. Goddard et al (1999) point at this potential problem in the NHS, and propose the use of "soft" data as both a complement and a substitute of hard data to assess the performance of Trusts.

2.9. A reference framework for the evaluation of hospital autonomy

Harding and Preker (2003) set out a conceptual framework for the evaluation of hospital autonomy. They propose two sets of determinants of hospital behaviour: on the one hand, they include external forces regarding ownership, market exposure (both to patients and to third-party purchasers) and stewardship. These have been traditionally

considered the relevant levers to handle in order to shape hospital responses. On the other hand, they propose that besides these levers, internal organisation is also important, but has been overlooked. These refer to: decision rights, residual claimant status, market exposure accountability and social functions. These can be summarised as follows:

- Decision rights: refers to the possibility to decide about assets, inputs and other aspects of day to day management as will be shown below.
- Residual claimant status: refers to granting the hospital the right to use any remaining resources after paying all the costs.
- Degree of market exposure: refers to the extent to which hospitals face competition once they shift to a contract-based relationship, and to what extent the market rewards them for good performance or punishes for bad performance.
- Availability of accountability mechanisms: refers to the presence of devices, such as a board of directors, a contract, monitoring and enforcement structures, etc., that prevent the manager from engaging in self-satisfying behaviour.
- Extent of unfunded mandates: refers to the extent to which the hospital has to provide uncompensated care or incur costs that it has no way to recover.

Regarding decision rights, this is the most obvious way in which autonomy can be materialised. The separation of ownership and control sketched by Fama and Jensen (1983) clearly describes the space for management: whereas owners specialise in diversifying risks, managers specialise in decision-making to increase the value of the owner's property. Although public hospitals have no owners, it is still possible to think of the separation between ownership and control, at least regarding the "decision space" that hospital managers are given to maximise social value. This decision space refers to the following items:

- Setting the strategic plan.
- Budget setting and flexibility of budget execution.
- Discretionary use of user-fees and other sources of revenue.
- Setting the marketing mix (products, prices, promotion, and placing).
- Medical management strategy for the production of clinical services.

- Human resources management.
- Procurement of drugs and other supplies.
- Outsourcing of services.
- Purchasing and selling-off of fixed assets.

Harding and Preker's framework starts from the assumption that in a continuum between a budgetary unit (no autonomy) and a privatised firm (full autonomy), a higher level of autonomy is more desirable. Nevertheless, this claim is not substantiated, given the lack of clear evidence about the effects of autonomy and the lack of a theoretical framework to guide research, as will be shown below.

Over and Watanabe (2003) created a detailed list of indicators based on Harding and Preker's framework, in order to evaluate the nature and extent of autonomy in a given organisation. Although it is an exhaustive checklist of the multiple issues involved in the process of granting autonomy to a public hospital, these indicators have not been used extensively. Paradoxically enough, the case studies reported in a review of experiences edited by Preker and Harding (2003) do not use the exhaustive list of Over and Watanabe but rather provide a qualitative gross evaluation of the five determinants of hospital behaviour.

2.10. Theoretical models of hospital behaviour

Regarding the PPS in the context of middle income countries (where the improved-efficiency rationale seems a more plausible driving force), it is assumed that granting diverse degrees of autonomy to hospitals and exposing them to the competitive pressures of markets will put the forces in place to make them improve efficiency and quality (World Bank, 1993). However, this expectation is based on the assumption that public hospitals will respond to market signals in the same way that is predicted by the standard theory of the firm. Alternatives to this expectation are based on models of hospital behaviour that were developed in the 70's and 80's. The most relevant models are proposed in Newhouse (1970), Pauly and Redisch (1973) and Harris (1977).[9] These and other models will be briefly reviewed in the following paragraphs. The critique is followed by new alternative approaches to modeling hospital behaviour.

[9] The comparison here between these three models draws on Folland et al (2004) chapter 13, Sloan (1980) and McPake and Archard (2002).

Newhouse's model proposes that the hospital's objective function is defined in terms of a tradeoff between quality and quantity, which reflects the utility of the decision makers, either physicians, administrators or trustees. The budget constraint is determined by the fact that revenues must equal costs, so the hospital aims only at breaking even. This relationship causes a unique shape of the budget constraint, which is upward sloping to a certain point and then backward sloping. The optimum level of production occurs where the utility maximising combination coincides with the quantity-quality frontier depicted by the budget constraint.

Newhouse's model does not explicitly consider the dominance of any of the decision makers. In contrast, Pauly and Redisch's model explicitly takes into account the fact that the hospital is controlled by physicians, which allows them to use the hospital to pursue their income maximising goals. In this vein, physicians aim at maximising average net revenue by restricting the number of staff appointments. If they cannot control the number of appointments, the number of staff will be higher but the net average revenue will not be inferior to the individual physician's supply curve.

While Pauly and Redisch's model improves Newhouse's omission of physician control of decision-making, Harris has gone further to propose that the hospital is not one single organisation but two separate ones. He proposes that the physician staff act as a demander of inputs for the treatment of patients and the administrative staff act as the supplier of those inputs. Given time-related asset specificity, physicians cannot bargain for prices, quantities and quality, so they have to make sure all the necessary inputs (capital and non-physician labour) are available when required. An important implication of this model is that government regulation to control for hospital cost escalation has to take into account two different targets: the hospital as the provider of inputs, and the physician as the demander of larger or smaller quantities of those inputs.

Another interesting theoretical model has been proposed by Muurinen (1986). In this model, physicians and administrators are modeled as powerful groups within the hospital, and the relative power of each group vis à vis the other leads to different outcomes. In one extreme, if doctors wield the highest bargaining power and they belong to the same group, the model would result in the Pauly and Redisch model.

The lack of relevance of these models (except Muurinen's) for developing countries seems to be largely the result of their applicability to the US context. However, other health systems in the world show features that make them radically different from the US system, and make these models less able to predict the effects of policy. Health systems in developing countries that have evolved towards insurer-and-provider competitive approaches (e.g., Colombia) also find these theoretical models less relevant because the purchasing power of the insurer strongly restricts physician autonomy to bring patients to the hospital and to use resources without restriction. If managers of a public autonomous hospital have power to hire and fire physicians, and to negotiate contracts with insurers, the remaining challenge is how to align physicians' incentives with the interests of the manager. The challenge remains because no matter physicians are salaried and managerial restrictions are put in place, they are still the clinical decision-makers, so that resource use still depends on their bedside decisions. This clinical decision-making cannot be transferred to the hospital manager.

At this point, the argument seems to turn upside down: no matter that physicians are more restricted as public hospital employees, their autonomy for clinical decision-making creates pressures on the hospital budget. The fact that the hospital is public, softens the budget constraint, so physicians still exert a strong influence to expand the budget (Lesur, 2002). But even in the infrequent case that the public hospital makes a surplus, given that it cannot be distributed, it is used to increase employees' salaries or perks, or to increase capacity, either in terms of beds or new technologies. This expectation creates a different set of incentives in public autonomous hospitals: if they are autonomous to use the surpluses, managers will try to reduce costs, increase prices, or both. They will be more able to align physician incentives, because they retain the power to hire and fire and do not depend on physicians for referrals. In this respect, an autonomous hospital's behaviour would be closer to that of a profit-maximising firm, or a not-for-profit firm that acts as a for-profit (Feldstein, 1993, p. 237). The key difference that keeps the autonomous hospital away from the property rights approach to profit-maximising firms is the lack of a residual claimant with a single-minded objective function, and the incomplete allocation of residual rights, including the right to sell the asset and realise the profits of its higher market value.

On these lines, McPake and Archard (2002) propose an alternative approach. Starting from the TCE view of the firm as a nexus of contracts, and based on the principal-agent

and property rights theories, the authors contend that a theory can be built that helps to better predict hospital responses to reforms. Thus, they propose that upstream, downstream and internal relationships can be explained as principal agent relationships and that the differing types of property rights allocation yield the corresponding types of relationships.

Regarding principal-agent theory, they propose that it is up to the principal to implement the policies that lead hospitals to align manager's and doctor's incentives so that social welfare is maximised. Nevertheless, this is a necessary but not sufficient condition, because welfare maximisation also depends on the objectives of the ideal principal. At this point McPake and Archard's approach assumes the existence of that ideal principal, the society, and the existence of a single-minded objective function that exhibits contractible attributes; as seen above, this is not the case for organisations in the public sector. Moreover, their proposal of the Minister of Health or other health authority as the ideal principal overlooks the fact that these bureaucrats cannot be more than agents to a diffuse principal and that they can use their power to advance their private agendas, as predicted by public choice theory. Perhaps the only situation in which public officers would perform as the ideal principals, would be when their career concerns and/or individual values coincide with long-term social values.

From the viewpoint of property rights theory, McPake and Archard argue that hospital responses depend on the distribution of property rights, and hospital reforms aim at altering such distribution. This is also emphasised by Harding and Preker (2003) when they point to the importance of granting residual claimant status to hospitals and managers as an important determinant of hospital behaviour. Bech and Pedersen (2005) argue that the strength of the incentives associated with residual claims depends on the hospital's freedom to dispose of surpluses and the hardness of the budget constraint, i.e., not only to use surpluses but face losses. It could be argued, following McPake and Archard, and Harding and Preker, that the adequate granting of property rights to the hospital -namely, residual claimant status and decision rights- and the adequate use of accountability devices and funding of social functions, would lead to the optimisation of social welfare. But transferring property rights to physicians and managers via performance-based incentives does not guarantee that social welfare is maximised, because good hospital performance is not necessarily aligned with the best social

welfare outcome. In addition, the difficulty to contract for best health outcomes opens room for maximising residual income at the expense of technical quality.

The problem of aligning a hospital's objective function with a social welfare function lies in the contradictions that both objectives imply. On the side of the purchaser, it could be argued that improving the aggregate level of health through cost-effective care would be the overall goal. However, society not always values the trade-offs involved in this objective when it means reducing curative-tertiary care to provide more preventive and primary care services (Goddard et al, 2006). The so-called "rule of rescue" (Hadorn, 1991) clearly points to the contradiction in social objectives when they have to be met with scarce resources. Thus, the lack of an identifiable ideal principal and a single-minded objective function makes it unlikely that allocating property rights and setting incentives to reduce agency costs yield an ideal level of hospital performance.

Interestingly, however, Harding and Preker (2003) propose that granting autonomy to public hospitals shifts the residual claimant from the public purse to the hospital. But given that the hospital itself has no "owners" it is unlikely that the ability to retain surpluses will make managers and physicians act as residual claimants in the strict sense of an ideal principal. This is because residual claims include two different types of rights that are treated as occurring simultaneously: residual control rights, and residual income rights (Hart, 1995). Although they may occur simultaneously in for-profit firms, they certainly are separated in public organisations: whereas residual income rights are in the hands of the manager or whatever powerful group can capture rents within the hospital, residual control rights are in the "hands" of a diffuse principal, the society. It could be said that managers hold residual claims if their jobs are at risk for poor judgment. But this may be true only with regards to the property rights of the manager (i.e., income and wealth), not those of the hospital (i.e., the market value of the organisation), because the ultimate residual control right is the right to sell the asset.

An additional point that is raised in McPake and Archard is that the hospital model is endogenous to the reform process. Accordingly, McPake and Archard propose that hospital behaviour should be modeled in the form of a generic model for the behaviour of mangers and physicians. This is a crucial departure from the prevailing models of

hospital behaviour, because it acknowledges that hospitals are means to an end of the two basic groups that operate within them.

Consequently, if each group's influence is modeled based on their profit-maximising objective, the hospital response to incentives can be better predicted. For example, if the purchaser decides to pay on a fee-for-service basis, both physicians and managers will have an incentive to maximise output if managers are able to set up a productivity-based incentive scheme for salaried physicians, or if they can shift employed physicians to a piece rate payment. Physician's output will be restricted by ethical, legal (the threat of a malpractice suit) and time constraints (Liu and Mills, 2005), so that demand inducement depends on the strength of these constraints. If the purchaser decides to pay the hospital on a prospective basis and the manager transfers the incentive to physicians to reduce output, the hospital will consequently reduce output. Or if the purchaser decides to keep the inertial historic budgeting independent of outputs, managers and physicians will try to minimise effort and the hospital will not increase output. Physicians' temptation to skimp on care or reduce output in these two situations will also be subject to ethical and legal constraints. It is thus more likely that the behaviour of a hospital dominated by two profit-maximising groups will be better predicted according to property rights theory. In this sense, the two groups will try to extract a rent as large as possible, and the share of the total rent will depend on the relative power of one group vis à vis the other.

Alternative approaches to modeling public hospitals in developing countries have not been thoroughly developed by economics theorists to include all the possible variations among health care systems. McPake and Archard's proposal is an introductory reflection but it has had little further elaboration. Therefore, the empirical analysis of public hospital behaviour lacks an adequate theoretical framework, a problem that is even worse when it comes to public hospitals in developing countries. A better knowledge of hospital behaviour in developing countries is warranted to inductively inform a process of theory formation for future research.

2.11. Conclusions

The separation of purchasing and provision in health care in middle-income countries seems to follow, at least partially, the logic of TCE: a search for a governance structure

that minimises transaction costs. However, this logic starts from the assumption that a vertically integrated structure shows higher transaction costs than a contract-based relationship. Nonetheless, agency problems in both types of governance structures and the lack of adequate allocation of property rights, make it difficult to predict which governance structure will have the lowest transaction costs. In addition, an inadequate theoretical modeling of hospital behaviour does not allow to develop a systematic approach to empirically test the assumptions on which the PPS rests. The next chapter reviews the empirical evidence found in the scientific literature on hospital autonomy, not only from the perspective of TCE but also from the perspective of the other rationales analysed in this chapter.