The optimal use of hospital capacity in the presence of stochastic demand and output heterogeneity

by

David Hughes

A dissertation submitted for consideration for the qualification of Doctor of Philosophy at City University, London.
Department of Economics
December 1997.

Contents

List of Tables and Figures

Figures **Page**

Acknowledgements

I would like to thank the staff at CHKS limited, in particular Erika Brown, whose invaluable efforts in collating the majority of the data used in this thesis were greatly appreciated. I would also like to acknowledge Ivan Csaba who provided data and also other valuable academic input to the thesis. Thanks also go to the staff at City University, London, who have sat through a number of seminars and listened to my ramblings on the subject matter contained within this thesis.

I would like to thank my advisor, Professor Alistair McGuire, who has been a constant source of advice, support and friendship. I have learnt much and benefited enormously from his input, as has this thesis.

Finally, I would like to thank my friends and family for their support, particularly my mother for proof-reading the whole thesis. But most of all I would like to thank Elaine, without whose endless support, encouragement and understanding this thesis would not have been finished.

Abstract

Hospitals provide a substantial proportion of the care services within health care systems: around 55% of total NHS expenditure is on hospital care. The majority of hospital care is undertaken on an inpatient basis requiring patients to be admitted to hospital. The range and type of admissions to hospitals, however, vary widely: Demand for care is heterogeneous. Hospitals, however, have limited space available to treat all the demands for care and excess demand is considered to be an endemic problem in the NHS revealed by the existence of long waiting lists.

Given the conflicting demands for care, hospitals must determine the mix of patients that they will treat given available capacity. As hospitals have limited capacity available to treat patients who demand services, choices must be made regarding the use of scarce resources in the face of these conflicting demands for available capacity. In addition hospitals face uncertain demand for some of their services, most notably those seeking emergency care. The clearest choice facing the hospital is whether to use available capacity to reduce waiting lists and treat elective cases, or keep capacity available to treat emergency cases should the need arise.

It is our intention to address the issue of how hospitals allocate capacity under conditions of demand uncertainty and output heterogeneity, in this thesis. In particular, we will consider whether hospitals appear to respond to the inherent trade-off in a way that can be explained by economics in terms of the behaviour of a rational economic agent. In order to investigate this we will attempt to formulate an integrated approach to considering the behaviour of hospitals given the nature of the environment within which they operate.

The thesis provides a review of the current strands of literature within hospital economics that deals with issues related to capacity utilisation and identifies different strands in the literature. We also review empirical estimates of hospital costs; these identify the theoretical foundations of the empirical studies to date.

The thesis is then separated into theoretical and empirical chapters. The theoretical chapter constructs a formal model of hospital behaviour building on theoretical foundations, to enable the problem of hospital capacity allocation decisions to be viewed within standard economic theory. The basic model introduces output heterogeneity, separating output into two, planned and unplanned (elective and emergency). In this model we also consider a utility function including a wider social perspective. Uncertainty is then introduced into the model, allowing stochastic demand for one of the outputs. The impact of uncertainty is considered in a formal manner by drawing on current theoretical knowledge regarding the influence of demand uncertainty on the production responses of the firm. A mathematical approach is specified, which allows an optimal allocation of capacity to be identified. This highlights the empirical content necessary to identify a fully specified model of hospital allocation decisions.

The empirical section estimates a cost function that is consistent with the theoretical specification identified in the previous section, by adjusting for demand uncertainty and output heterogeneity. The implications for the estimation of a hospital cost function are then considered and, particularly, the implications for standard economic theory of cost analyses. A number of different estimation approaches are assessed. The Box-Cox model is the preferred approach and this performs well, most notably, the demand variable included to pick up the influence of demand uncertainty is both significant and of the correct hypothesised sign.

The final part of the thesis attempts to bring together a more fully specified empirical solution to the problem, by focusing on the social costs of turning emergency patients away and leaving elective patients on waiting lists. Two empirical elements are identified from the theoretical model as being important: The probability of turnaway, and the relationship between waiting lists and this probability. The probability of turnaway for each hospital for each month is estimated. A number of different estimation techniques were employed to estimate the latter including a Tobit, Heckman two-stage and Box-Cox analysis. Based on these estimates a fully specified empirical model allowed the calculation of the implied marginal social costs of turning an emergency patient away. These were estimated to be around £300, and the implied marginal social costs of placing a patient on the waiting list were estimated to be just over £1. These represent indicative values and are based on aggregate estimates across all hospitals, nonetheless, they provide the first estimates of this kind adjusting for demand uncertainty and including output heterogeneity.

Chapter 1: Introduction

Hospitals provide a substantial proportion of the care services within health care systems. In 1994 the NHS spent £21,657m on hospitals representing around 55% of total NHS expenditure. This is approximately £880 per household in the UK. The majority of care is undertaken on an inpatient basis requiring patients to be admitted to hospital. In 1994 5,968,000 people received in-patient hospital treatment in the UK. The obvious consequence is that hospitals must have beds available in order to provide the necessary inpatient care. The total bed capacity in the UK in 1994 was 294,000. The range and type of admissions to hospitals, however, vary widely. The ICD three digit classification lists over 1,000 different disease states; demand for care is heterogeneous. Each disease differs in terms of treatment and severity and, potentially, each requires a different length of time in hospital. Hospitals, however, have limited space available to treat all the demands for care, and excess demand is considered to be an endemic problem in the NHS revealed by the existence of long waiting lists. Given the conflicting demands for care, hospitals must determine the mix of patients that they will treat given available capacity.

The reforms of the NHS gave hospitals greater autonomy in terms of choosing the type and scope of services provided, allowing hospitals to operate with greater independence. Hospitals still, however, operate under intense political scrutiny and their performance is monitored on a number of levels. Two issues have attracted particular attention in the UK. The first is the size of waiting lists for elective care. In the UK in 1997 there were over 1,200,000 patients waiting for hospital treatment, attracting adverse political attention. Hospitals are, therefore, under pressure to use capacity to the limit in order to treat as many patients as possible given available capacity. However, given the heterogeneity of demand for services and that some element of hospital demand is truly stochastic, hospitals are faced with the problem that as pressure builds up on capacity there is less scope for reducing uncertainty through the existence of excess or standby capacity. This leads to the second issue: Hospitals are under pressure to be able to respond to the fluctuating need for urgent care, such that treatment should always be available and patients seeking emergency care should not be turned away. When patients are turned away from hospital facilities this attracts adverse publicity. Therefore, whilst waiting lists exist for some forms of care (i.e. planned elective care) other forms of care are provided for immediately. This trade-off has been explicitly recognised recently by the NHS Executive in EL(97)42, which emphasised the need to look into the balance between meeting emergency requirements and the size of waiting lists.

The two problems co-exist yet appear to call for different solutions. The problem of waiting lists suggests an increased throughput and full use of existing capacity. The potential problem of turning emergency patients away, however, requires that reserve capacity to be held in order to maintain the flexibility to respond to random demand for services. There exists, therefore, an inherent conflict within the system.

As hospitals have limited capacity available to treat patients who demand services, choices must be made regarding the use of scarce resources in the face of conflicting demands for available capacity. The clearest choice facing the hospital is whether to use available capacity to reduce waiting lists and treat elective cases, or keep capacity available to treat emergency cases should the need arise. Staffed hospital beds are costly and leaving large amounts of capacity empty for large periods of time will incur costs without any apparent benefit being accrued. Nonetheless, given the less than full utilisation of hospital capacity over long periods of time, i.e. average occupancy rates below 100%, this is what hospitals appear to do. On the other hand, the seasonal peaks in emergency demand can lead to excessive pressure on the system, with 100% capacity levels being exceeded on occasion.

The fact that hospitals normally operate within capacity constraints indicates that hospitals are trading-off these two issues. This suggests that the hospitals may be weighing up the costs and benefits of allocating capacity between the conflicting demands. It is our intention in this thesis to address the issue of how hospitals allocate capacity under conditions of demand uncertainty and output heterogeneity. In particular we will consider the optimal response to the inherent trade-off in terms of the behaviour of a rational economic agent. We will develop a model that attempts to formulate an integrated approach to considering the behaviour of hospitals, given the nature of the environment within which they operate.

Economic theory suggests that the rational economic agent will utilise scarce resources in order to achieve the maximum benefit from these available resources. Consequently, if the hospital is acting as a rational economic agent, it should decide how to use available capacity based on the costs and benefits of providing care. The detailed specification of these costs and benefits raises three key elements for analysis: The first is the definition of the general aims of the hospital, that is, given the non-profit nature of hospitals what costs and benefits are they likely to consider? The second is, once the relevant economic variables have been identified how do they interact? The third involves providing some empirical content to the problem through estimating the magnitude and empirical relationship between the economic variables.

The aim of this thesis is to pursue the investigation of the use of hospital capacity in the face of demand uncertainty and heterogeneous output, through analysing each of these three dimensions in detail. In doing so we will develop an economic theory of hospital behaviour that allows the development of empirical hypotheses that can be tested using available data. This will improve existing knowledge of hospital behaviour, as well as stimulating new ideas in an area that has an important bearing on health care policy. We are specifically concerned with the production responses to demand uncertainty and output heterogeneity. As such we will develop a theoretical model that distinguishes between two broad types of hospital demand; emergency and elective treatments. Demand uncertainty characterises emergency treatments, and the production of both types of output must take account of this stochastic element of total demand. This theoretical model then allows the development of a cost function that is consistent with hospital behaviour, where behaviour is determined by production responses to demand uncertainty. Finally, other aspects of hospital behaviour relating directly to the trade-off between the use of capacity in treating the two broad categories of demand are modelled.

To address these issues this thesis will be structured in the following way. Chapter 2 will review existing work in the area, noting the fragmented nature of current research. In particular, we will highlight two distinct strands of literature that have developed, almost in isolation to each other; the first are essentially static theories of hospital behaviour, the second refer to simulation and empirical models of behaviour. These will be reviewed with specific focus on the issues highlighted above: Heterogeneity, demand uncertainty, and capacity utilisation.

The static behavioural theories of hospitals allow us to consider the factors that may enter the hospital objective function, given the nature of the environment within which hospitals operate. This enables us to identify the theoretical trade-offs and arguments that enter the hospital's objective function. The simulation models of hospital behaviour allow us to formalise the relationship between the objective function and the decisions faced by hospitals, in particular focusing on the issue of limited capacity where our interest is in the use of capacity given heterogeneous output and stochastic demand. Finally, the review of empirical studies allows us to consider how previous authors have estimated cost functions where attention will be paid to the theoretical construction of cost functions, and how this fits in with the particular issues we are focusing on.

Chapter 3 will develop a model of hospital behaviour that is consistent with an environment in which production must respond to a stochastic element of demand. We will use a standard production possibility framework but will introduce uncertainty. This is the first time uncertainty has been introduced into such a framework. Building on this model, we develop a mathematical representation of the problem consistent with this approach, considering output heterogeneity, demand uncertainty, waiting lists, and turnaway rates. This will enable the identification of a theoretical optimal allocation condition and will identify the first order condition necessary to provide an empirical solution to the theoretical model. Chapter 4 will provide the first part of the empirical content of this general approach by estimating a cost function that is consistent with the theoretical specification, once more paying attention to the issues of heterogeneity and demand uncertainty, and focusing on the issues of duality and cost minimisation. Chapter 5 will provide the second part of the empirical content and will estimate the social costs implied by observed behaviour of hospitals in the NHS that which is consistent with the hospital's basic trade-off in the use of capacity when faced with output differentiation and stochastic demands. Chapter 6 will draw together some conclusions.

Chapter 2: A review of theoretical and empirical analyses of the hospital

Section 1: Introduction

This thesis concentrates on three main issues; the stochastic demand for hospital services, output heterogeneity, and the utilisation of hospital capacity. A large literature exists on general hospital economics but very little deals with these specific issues. This chapter will review the literature that considers capacity utilisation in some detail. However, it is still worthwhile looking at the general hospital literature to give some background, in particular to the consideration of the formulation of objectives and the behaviour of the hospital as an economic agent. This provides some reference point to the model developed in Chapter 3.

There have been two distinct strands of analysis that have developed within the field of hospital economics and these will be reviewed in sections 2 and 3. Section 2, itself, will be broken into two parts, where the first part considers analyses that do not explicitly deal with capacity utilisation, but focus on theoretical constructs of behaviour. The second part will deal with a smaller, more specialised literature that considers capacity utilisation and stochastic demand. Section 3 will review the empirical models of hospital behaviour emphasising the cost function with regard to stochastic demand.

There has been little integration of these strands within the literature, and this has led to little formalisation of the capacity optimisation problem. The lack of formalisation across the two sections of literature causes problems of integration between the sections, but the primary root of the problem is due to the theory itself being inadequately formalised without reference to mathematical or empirical models of behaviour. A final section, Section 4, will draw together the key issues when considering the development of an integrated model of hospital behaviour.

Section 2.1: Theories of hospital behaviour

This section will focus on the purely theoretical constructs of hospital behaviour and theories of non-profit institutions. As the objective function is an important determinant of hospital behaviour, in general, and a potentially important influence on capacity utilisation decisions, in particular, we highlight the lack of agreement across the literature over this fundamental aspect.

A number of possible objective functions for non-profit enterprises have been suggested in the literature, including revenue maximisation, quantity maximisation, quality maximisation, and utility maximisation. These provide grand theories, which largely follow neo-classical lines, and consequently, they tell us little about internal allocation within the hospital or capacity utilisation. Most of these studies abstract from the complexities of the hospital sector and few, for example, have introduced heterogeneous output or recognised that stochastic demand exists for some services.

The failure to consider the stochastic nature of demand potentially causes problems, since this is one of the key issues in determining capacity utilisation. Whilst there have been several attempts to develop operational models of hospital behaviour, these have failed to incorporate fully the fundamental nature of demand and output; i.e. endogenous output for some services and stochastic demand for others. Furthermore, largely due to the dominance of theories coming from the US, they have not dealt with capacity constrained systems and, therefore, tell us little about internal capacity utilisation.

While the objective function of the hospital is clearly an important factor in defining hospital behaviour, the determinants of this function may also be linked to ownership. In broad terms, hospitals can be separated into three forms of ownership. Hospitals may be privately or publicly owned, or operate as charitable organisations. In addition to which hospitals may operate on a for-profit or non-profit basis within each form of ownership. The majority of hospitals in the UK are publicly owned and run on a non-profit basis.

The difference between the objective functions of non-profit and for-profit organisations has received some attention in the literature, most notably in the US, since there exists the full spectrum of ownership, and it is often hypothesised that different ownership types will result in different objective functions. It has been suggested by Pauly (1987), however, that much of the discussion about the objective function of non-profit enterprises is irrelevant, and that for practical purposes the different objective functions collapse into three basic forms:

- maximisation of money income of a set of agents (most notably physicians).
- maximisation of profits-in-kind to managers/decision makers.
- maximisation of output or quality.

Pauly concluded that any income maximisation model or profit-in-kind model will yield the same quantity, quality, and price as profit maximisation (although costs may differ since profit will be represented as costs in profit in kind models) as all models are concerned with residual maximisation. He suggested that only if quantity or quality directly entered the utility function would an observable behavioural difference be apparent. Even here output and quality maximisation could, under certain conditions, be compatible with residual maximisation. Pauly concluded that theoretical and empirical work might actually suggest that ownership is less important than it initially seemed. However, since the majority of objective functions in the literature propose that quantity, quality, or both directly enter the utility function, Pauly's conclusions are less than satisfying. Furthermore, the evidence Pauly draws on only considers the objective function in terms of the impact on efficiency, and, most notably, whether or not non-profit firms cost-minimise. That is, he does not consider issues of capacity utilisation.

Perhaps the most straightforward yet appealing objective function suggested in the literature is that of quantity maximisation. Quantity maximisation has been suggested by a number of authors as the relevant objective function for non-profit enterprises. Long (1964), for example, suggested that non-profit enterprises would aim to maximise the number of patients seen subject to budget, capacity and quality constraints. However, straightforward analysis of quantity maximisation, despite its appeal, fails to tell us anything about the internal allocation of capacity within a hospital. It fails to recognise the heterogeneity of hospital output and clearly, in the short-run, if capacity acts as an aggregate constraint, choices regarding capacity utilisation have to be made. Moreover, study of output maximisation, per se, considers total capacity rather than issues of internal capacity allocation across different demands.

With a straightforward quantity maximisation objective the hospital would, in theory, simply treat as many patients as capacity would allow, at least cost. If, however, we allow for heterogeneity of output, then the hospital would prioritise patients on the basis of, for example, expected length of stay, treating those patients with the shortest length of stay first in order to maximise throughput (and, therefore, the number of patients treated). The rational hospital would effectively categorise output on the basis of length of stay, treating all patients within a given 'band' of length of stay, until demand was exhausted for the treatments in question, and then move on to the 'next best' category of patient, defined by length of stay.

Therefore, a quantity maximisation objective function implicitly gives an equal weight to all treatments with the same length of stay, regardless of cost or outcome. That is, without reference to the nature of the condition being treated. Long (1964), amongst others, recognised the importance of this noting that some treatments may require immediate treatment. This, however, suggests that some sort of weighted quantity maximand would be more appropriate, recognising the heterogeneity of output within the hospital.

Reder (1965), developing the quantity maximising theme, suggested precisely this; that is, the objective function should be one of weighted quantity maximisation. Quantity, in his model, was weighted on the basis of professional prestige as determined by physicians. Ginsburg (1970), in a similar manner, suggested a weighted quantity maximisation objective function subject to a budget and capital constraint. Weighted quantity maximisation is more informative in that it recognises heterogeneity of output and that the hospital (decision makers/physicians) may derive more utility from treating some patients than others, (although the exact link between the utility function of the hospital and patient is often vague). It can also potentially tell us more about the process of capacity utilisation within the hospital, since the relative size of the weights attached to each output will be an important determinant of capacity utilisation decisions within a capacity constrained system.

The question of how the weights attached to output are determined is not, however, straightforward, and depends ultimately on the definition of the appropriate decision-maker in the hospital. Reder (1965), for example, suggested that the weights are likely to be determined by physicians and placed the emphasis on the role of physicians as the dominant decision-makers within the hospital. This is consistent with Pauly and Redisch's (1973) argument that the hospital acts as a physicians' co-operative, and that the physicians dominate the decision making process, such that the aim of the hospital becomes one of maximising physicians' income.

Pauly and Redisch (1973) used their model to consider optimum levels of medical staff, and suggested that the problem the non-profit hospital faces is identical to the profit maximising solution with hospitals as cost minimisers. With respect to issues of capacity utilisation, their model is similar to the weighted quantity maximisation model proposed by Reder in as much as physicians, who decide which patients to admit, will allocate capacity to cases where their fee is highest.

There are, however, problems with this model. First, unless there is sharing of income between physicians, there may be internal conflict as physicians compete for limited capacity. This is equally true of Reder's model, where physicians are responsible for weighting output by 'prestige' attached to treatment, and where the incentive may be for each physician to overstate the relative prestige attached to his/her own cases. Secondly, the applicability of the model outside the US is limited. This is particularly true in the UK where hospital physicians are largely salaried and, consequently, the model looks less convincing, although it may be possible that increasing caseload and prestigious cases may lead to faster promotion and higher salaries (for example, through merit awards). The main problem is that it is difficult to view the hospital as a co-operative when there are so many apparent internal conflicts.

Holtmann (1988) criticised Pauly and Redisch on the grounds that their model does not explain the existence of non-profit enterprises, and their analysis is essentially a profit maximisation model. Nonetheless, the basic idea that output is weighted by price and that price is linked to relative worth is an improvement on earlier models. Indeed, authors such as Ben-Ner (1986) have suggested that the hospital may in fact be a consumers' co-operative, in which case the weights would be related to utility of patients rather than physicians. However, the actual mechanism for revealing patients' valuations is unclear in a healthcare system where insurance and a principal-agent relationship predominates.

In a variant of the straightforward quantity maximisation model, Newhouse (1970) suggested a quantity-quality trade-off might exist. He postulated that the first element of concern to non-profit enterprises is the quantity of services provided, and the second is quality, where quality is linked to professional excellence and the ability of the hospital to attract staff. Furthermore, he suggested that administrative staff, medical staff, and hospital trustees, may attach different weights to quantity and the elements of quality, although this does not affect the theoretical construct of his work.

In his model, quality can be introduced as either a choice variable or a constraint. Newhouse considered it to be a choice variable, and proposed a budget constrained quantity-quality maximand (although in a cash-limited system such as the UK it may well be more appropriate to consider quality as a constraint rather than a choice variable). He suggested a number of ways quality could be measured: personnel/patient ratio, professional perks/patient ratio, or the extent of laboratory facilities available. However, he recognised that these may represent capital/labour substitution rather than quality per se.

By letting demand be a function of both price and quality, and proposing that cost be a function of quality, Newhouse determined a cost-quality vector by assigning an arbitrary ordinal set of numbers to quality, allowing costs to be used as a direct measure of quality.

Identifying the output associated with any given level of cost/quality allowed the quantity-quality locus of equilibria to be defined, where the rational decision-maker will choose the point that gives highest utility. Newhouse concluded that as long as resources are constrained the quantity-quality trade-off is an inherent one; these are the two commodities to which resources can be allocated.

Newhouse explicitly ignored the problems posed by the model (notably the problems of inefficiency and maximum quality constraints), and confined his analysis to a single product firm, thus side-stepping the problems associated with the heterogeneity of output. Nonetheless, the model is reasonable, as far as it goes, if both quality and quantity enter the utility function. However, there are a number of problems associated with defining quality, measuring a cost-quality relationship, and applying weights to relative quantity-quality combinations. Some quality is intrinsic to health care itself, some is related to hotel services. Other quality measures may refer to the general performance of the hospital, for example related to waiting lists and turnaway rates. The problems of measurement are not confined to quality alone, as Newhouse recognised; the measurement of the output of the hospital depends on the nature, and severity of the disease and a simple measure of patient days may not accurately capture this; nonetheless, this is the output measure used. The main problem is operationalising the Newhouse approach since it leaves many issues unresolved. Furthermore, since the model avoids the issue of output heterogeneity, it provides little to assist in answering this question of how hospitals allocate capacity.

Other authors have defined the objective function within a general utility maximising framework. Lee (1971), for example, assumed that hospital administrators attempt to maximise utility, where utility was defined as a broad concept including all the variables that affect the administrators well-being; such as salary, prestige, security, power, etc. He based his assumption on organisational theories of the firm, where the status and prestige of the administrators are dependent on the status of the hospital itself, and where hospital status in turn is a function of variety, quantity, and complexity of inputs. Therefore, the utility function effectively becomes one of status maximisation, where the hospital attempts to minimise the gap between actual and desired status.

Clarkson (1972) suggested that the non-profit hospital managers would have a weaker association between their own wealth and that of the hospital organisation. The fact that there are no share options or profit related pay means that the cost to the manager of not maximising the organisation's wealth is less in the non-profit sector. Furthermore, the unconstrained non-profit manager could choose better office facilities, hire more congenial colleagues, have more relaxed personnel policies, and other non-pecuniary benefits. This is likely to be a function of the defined benefits; i.e. non-profit managers may take advantage of non-pecuniary benefits, and this conforms to Pauly's (1989) notion that ownership, per se, may not affect behaviour.

Holtmann (1988) suggested Clarkson's view of non-profit firms was in fact a special case of a Williamson utility maximising model where non profit firms have the scope for opportunistic behaviour, allowing managers to gain utility. However, it is not clear whether this is a theory of non-profit, non-competitive, or unconstrained behaviour. As Clarkson recognised, if there is scope for opportunistic behaviour this may potentially lead to tighter controls on behaviour.

Holtmann (1988) in a review of earlier work, suggested it was possible to set up the Newhouse model using a Williamsonian framework, and criticised Newhouse for not developing a mathematical model. A mathematical representation could present the non-profit firm as a utility maximiser that is subject to a break-even constraint. Holtman set up the problem as a Lagrangian, presented below:

$$L = U(Q_1, Q_2) + \lambda(P_1Q_1 + P_2Q_2 - C(Q_1, Q_2)) \qquad (2.1)$$

where utility is a function of quantity and quality, represented by $U(Q_1, Q_2)$, where Q_1 is quality, Q_2 is quantity, P_1 is the price of quality, P_2 is the price of quality, and $C(Q_1,Q_2)$ is the cost function. The solution of which tells us that the non-profit enterprise will produce where product and quality prices are greater than their marginal costs; the extra quality being justified by the managers' utility gain.

This, however, implicitly assumes that the hospital is a type of managers' co-operative, and that consumers' utility is a by-product after managers have chosen quantity and quality. Alternatively, the situation may be represented by managers acting as imperfect agents for consumers, as from the consumers' perspective there is excessive quality and output, given that the level of quality is greater than the willingness to pay. Furthermore, it is not clear how the market operates in this model, since in a competitive market we would expect quality to be reduced, i.e. the market would not withstand excessive quality.

Holtmann suggested that many theories of non-profit enterprises ignore the widespread social acceptance and emergence of non-profit firms. Implying that it is unlikely that non-profit firms would be acceptable if its managers were behaving opportunistically. This ignores the fact that opportunistic behaviour may not be obvious, and those non-profit firms may only operate in social markets where for-profit firms would be less acceptable.

Weisbrod (1975) suggested that non-profit firms may emerge to provide services with 'public good attributes' (although strictly they are not public goods since they have rivals and are excludable) and evidence suggests non-profit firms provide more 'public good' services, such as hospital emergency room services than for-profit firms provide. This indeed may be observed in the UK, where private, (non-profit), firms only provide elective care. Davis (1972), in her review of the theory and empirical evidence of the economic behaviour of hospitals, found evidence to suggest that a greater proportion of non-profit hospitals operate specialised facilities such as intensive care units and post-operative recovery rooms than for-profit hospitals. It may also be possible, however, that this reflects the difference in objective functions between for-profit and non-profit firms.

Ben-Ner (1986) viewed non-profit firms as consumers' co-operatives that produce output as an individual consumption good, and quality as a public consumption good for all consumers of the product. However, he viewed quality as non-separable, unlike Newhouse. Holtmann (1988), developing this theme, suggested using willingness to pay as an index of welfare, such that the welfare function can be written as:

$$W = \int_{0}^{Q^*} P(Q_1, Q_2)dQ_1 - C(Q_1, Q_2) \qquad (2.2)$$

Where P is price, Q_1 is quantity, Q^* is quantity of output produced, Q_2 is quality and $C(Q_1,Q_2)$ is the cost function. Non-profit organisations maximise consumer surplus subject to a break-even constraint. However, Ben-Ner (1986) suggested that in an ordinary co-operative the power may become concentrated in the hands of a few, and this will lead to a conflict between consumers' interests and those controlling the organisation, which fits in with Pauly and Redisch's (1973) study.

This analysis, however, depends on the structure of the market and what enters the decision-maker's utility function. If consumers' welfare is the sole argument in the producer's utility function, then producers' and consumers' interests will coincide, and there would be no conflict.

All of the preceding work highlights the debate over the objectives being pursued by the hospital and the resultant behaviour. These theories are consistent with cost-minimising behaviour, but say little about the cost generating process, i.e. the production process of the hospital. Thus, while we can differentiate extensively amongst the specific details of this theoretical literature, there is little insight into the production process and hence little knowledge gained about the capacity utilisation issue. Given these limitations there is little that can be operationalised from these models, as they are no more than descriptions of variants of behaviour stemming from the different specifications of objective functions, which in any case, as Pauly pointed out, may be little different from one another.

Harris (1977) noted that most models of the hospital are too simple, and fail to capture the more complicated features of the sector, e.g., the absence of equity capital, regulatory controls, insurance subsidisation. The nature of the health care product has lead to an insurance market developing, due to the potentially catastrophic costs of illness and the uncertain timing of episodes of illness. It is useful to consider how the hospital responds to such a market, and this is where the main contribution of Harris' work lies.

The main thrust of Harris' analysis was that the hospital is really two firms in one. The hospital as a firm ,requires complicated decision processes due to the complex and uncertain nature of illness. This requires an organisation that can respond and adapt to changing circumstances, such that hospitals operate a specialised system of very short run internal resource allocation procedures, and this forms the basis of the split organisational structure. Harris emphasised the need for hospitals to maintain flexibility in the face of uncertainty.

Here Harris touched on a very important point related to the nature of demand for hospital services and the response of hospitals to this. In essence there are two arguments in the paper:

i) that hospitals have an implicit contract to treat patients; and

ii) flexibility is required in the face of uncertainty.

Harris referred to the latter as 'standby capacity' (capital, labour and capacity) required in order to respond to uncertainty and to have the facilities available to treat all patients. Due to the flexibility of labour, he implies that standby bed capacity is likely to be the focus of standby capacity requirements. The implication is that hospitals derive some sort of utility out of treating patients and disutility if they fail to treat patients.

Harris described a situation whereby physicians make spot markets for ancillary services, and this creates the dual nature of the hospital that Harris refers to, where an 'internal market' is created for these services operating within the hospital. This, however, simply reflects the uncertain nature of demand. Harris' argument referred to uncertainty at a micro level, regarding specific treatments, but this argument equally applies at a more aggregate level where demand for health care is equally uncertain. The emphasis placed on supply assurance (i.e. the aim to keep facilities available to treat patients), and the need for physicians to obtain access to ancillary services, implies that there is a disutility associated with not treating patients.

Harris described the capacity problem in hospitals as 'a complicated system of rationing, rules and manoeuvring'. He explicitly recognised the fixed nature of certain parameters such as beds, and, in the very short run, ancillary services. He suggested that capacity decisions should not be considered as different from any other inventory problem, that is, the joint probability density functions of demand for inputs as well as the left and right hand loss functions of excess and insufficient capacity must be known. Therefore, the fundamental problem highlighted by Harris is that demand is uncertain and, given a fixed capacity, hospitals must decide how to allocate this capacity in the short-run.

The situation becomes interesting when the heterogeneous nature of hospital services is taken into account. The problem becomes one of weighting different treatments and weighting the 'cost' associated with turning different patients away. Harris recognised as much, when he suggested that medical and administrative staff would attach weights to the right hand and left hand loss functions. He suggested that administrative staff will consider the cost of holding excess capacity in terms of revenue forgone, and that this will be heavily weighted; however, medical staff will be more concerned with not turning patients away than with costs or empty beds. This was, therefore, a description of the bargaining process as related to excess capacity and turnaway, but without formalisation and the optimal solution was not outlined.

The description is, however, somewhat limited. For example, the assumption that physicians will be less concerned about empty beds than not turning patients away is an over-simplification. Physicians will only be unconcerned about empty beds if all demand is being met. If, however, as is the case in the UK, there is excess demand for services, then physicians may be concerned about empty beds, particularly if, as Harris suggests, bed capacity is constrained. However, empty beds may well be a concern to some physicians but less so for others, and conflict may arise between physicians. For example, an empty bed in an elective specialty with a waiting list represents an untreated patient; however, an empty bed in an emergency ward represents flexibility to treat more patients should the demand arise unexpectedly, and a reduction in the likelihood of having to turn patients away. Therefore, the Harris model, whilst recognising some of the complexities of the hospital, does not deal with the issue of how constrained capacity is allocated between potentially competing demands and suffers from the lack of formalisation of the problem.

The theories discussed thus far represent fairly grand and abstract theories, in most cases following neo-classical lines. They are little more than descriptions of outcomes attained under a variety of assumed patterns of behaviour. The description by Harris represented a step forward in so much as he recognised that one of the key issues facing a hospital is demand uncertainty, and that this will be one of the main considerations when determining capacity utilisation decisions under fixed capacity constraints. However, as with the other theories, there is no formal link to the production process and output heterogeneity is widely ignored.

There have, however, been some attempts to simulate and empirically model hospital behaviour. Whilst these models move away from a theoretical discussion, it is worth considering the theoretical underpinnings of these models here as they offer some valuable insights into some of the issues faced when operationalising maximands; they also highlight some important issues associated with modelling hospital capacity utilisation decisions.

Section 2.2: Simulation models of capacity utilisation

All the above studies are deficient in terms of capacity utilisation being left unanalysed, with the exception of Harris. However, as noted above, Harris failed to formalise the problem and, therefore, does not allow identification of the optimal solution to the issue of capacity utilisation. This section will consider some models that present more formal analyses and potentially allow optimal solutions to be identified.

One of the first models to consider hospital capacity decisions was developed by Shonick (1970). This model was primarily concerned with the rising costs of hospital care in the US and hospital efficiency in general. Shonick suggested that one of the most important economic characteristics of hospital operations is the relatively large proportion of cost that remains fixed in the face of variability in the proportion of occupied beds. He identified two important issues in his model. First, the fact that unoccupied beds do not earn revenue. Secondly, that hospitals may have to cover costs and set prices accordingly - unoccupied beds result in fixed costs being allocated over fewer cases and this increases average costs per case, which has implications for prices. He suggested that if the number of beds could be 'safely' reduced without reducing the number of occupied beds this would reduce costs. Shonick addressed the question of how many beds a community needs to meet its demand for hospitalisation. The decision context for his model was the medium to long term, as it is widely accepted that bed numbers are fixed in the short run.

Shonick recognised that before the question of hospital efficiency could be addressed a determination of the criteria of efficiency to be satisfied is required. He noted that a common measure of hospital efficiency was the average occupancy rate. He also noted that in many cases the only criterion used was the frequency with which all beds were occupied. Shonick developed this theme and suggested that high occupancy rates and low overfill rates represent two 'basic' criteria whose optimisation might be the goal of the hospital.

He suggested that concentrating solely on minimising the frequency of overfill implies a large number of beds relative to the average daily demand. However, if reducing costs is of importance, then this would suggest keeping the number of unoccupied beds low, and this implies a small number of beds relative to demand. Thus, using these two criteria together creates a potential conflict; there does not exist a choice of beds which will simultaneously minimise the overfill rate and empty bed rate. He suggested that, consequently, the hospital must choose an acceptable level for each of these, whilst not achieving an optimal level for either. The focus, therefore, was on the trade-off between these two criteria, such that once the inherent trade-offs between the factors are recognised then it is possible to identify an optimal solution, although Shonick actually set up the problem as one of maximisation/minimisation rather than optimisation.

It should be noted, however, that it is not immediately clear which of the criteria are objectives and which operate as constraints. For example, if the sole reason for including occupancy rate is as a proxy for average costs, as Shonick implies, then it is not clear whether this should be included as a constraint or an argument in the objective function. This raises the issue of what factors enter the hospital's utility function, and in particular, why occupancy rate should directly enter the utility function?

Occupancy rate may enter the utility function for two reasons: First, because the higher the occupancy rate the larger the number of patient days that can be supplied for any given bed number; this argument focuses on quantity of output. Secondly, because of the relationship between occupancy rates and costs, i.e. the higher the occupancy rate the lower average costs are likely to be and the lower prices will be, therefore, the higher demand, and, hence, output will be. The latter argument is the one Shonick focused on.

Shonick was also one of the first authors to note the importance of the stochastic demand for services. In specifying his model he assumed that there was a stream of patient arrivals with a random number arriving daily. This formed an important part of the model specification, where expectations are formed with regard to the random nature of arrivals. He noted that since the criteria variables set up in his model were random variables they would have probabilities attached to them. Shonick went into some detail regarding the precise specification of demand, the calculation of the mean census, and the expected loss function. This allowed him to calculate the criteria of efficiency for any choice of demand and bed numbers.

He considered the optimisation of two principal criteria: maximising the expected occupancy rate and minimising the expected proportion of the time the facility will be full. Using the properties of the Poisson distribution, such that expected daily census E(C) is:

$$E(C) = \gamma; \qquad \sigma_c = \sqrt{\gamma}; \tag{2.3}$$

Where σ_c is the standard deviation of C, and the coefficient of variation (σ/μ) is given by

$$\text{Coeff of Var}(C) = 1/\sqrt{\gamma} = k_c \tag{2.4}$$

He showed that if demand, γ, is sufficiently large, then k, the coefficient of variation is small. Therefore, it is possible, in the presence of a large average daily demand, to choose the number of beds, such that it is not much larger than γ, and yet have a relatively small rate at which the facility is full. Thus, under this model the prerequisite for being able to satisfy both efficiency criteria is a large effective demand. Shonick suggested this was borne out by the often observed high correlation between large hospitals and high occupancy rates.

It would appear then, from Shonick's model, that the size of demand determines whether hospitals can satisfy the criteria. Unfortunately this tells us little about the internal allocation processes. The extent to which hospitals can satisfy the joint objectives of maximising occupancy rate and minimising turnaway rate depends, primarily, on the size of demand and this is assumed to be outside the control of the hospital.

There is, however, still an inherent conflict between the two objectives, since as the number of beds rises, the expected occupancy rate and the expected turnaway rate will both fall. The hospital must still choose the level at which the criteria are optimised, recognising the inherent trade-off between the two. The model gives no indication of how this will be done. Shonick did note that it is common for planning agencies in hospitals, particularly in the US, to assert that hospitals should operate at a predetermined occupancy rate[1]. His model, however, points in a different direction. He suggested that desirable levels of the efficiency criteria should be administratively determined rather than through regulation. Furthermore, that this would especially be the case if the turnaway rate was set at a maximum level. This implies that there would be constraints introduced into the system. If this were the case, in Shonick's model once the turnaway rate is determined everything else falls out, since hospitals would simply satisfy the maximum turnaway rate and this would determine the occupancy rate. This once again returns us to the issue of why occupancy rate enters the utility function directly. Shonick touches on this in his introduction, stating that a low occupancy rate will increase average costs. He does not, however, make any further reference to the cost of providing the spare capacity in his model.

He did, however, implicitly recognise the trade-off between turnaway rate and occupancy rate, but does not fully draw out the issues. He noted that there are potential costs, in terms of lost utility, associated with turning patients away, and that the probability of turning a patient away is inversely related to the bed numbers, however, the more beds there are the lower the occupancy rate, and the lower the occupancy rate the higher the costs. Therefore, the trade-off is between higher costs and reduced turnaway. An optimal decision must be made with reference to the expected rate of turnaway, the associated costs of turnaway, and the cost of providing empty beds at the margin.

Consequently, the model highlights some important issues regarding the inherent trade-offs that exist, but cannot select between the many different solutions due to the lack of formalisation of the optimisation process.

Nevertheless, Shonick was one of the first authors to note the importance of the stochastic demand for services. He also recognised the importance of the heterogeneity of demand. In specifying his model he assumed that there were two annual streams of patient arrivals, with the number arriving daily from each represented by random variables Xp and Xn, where Xp referred to 'physician generated' or elective arrivals, and Xn referred to 'nature generated' or emergency arrivals. He assumed both are Poisson distributed with means λp and λn respectively.

---

[1] For example, the Hill-Burton regulation in the US calculated the total 'need' for beds based on an occupancy rate of 80%.

Furthermore, he assumed that if an emergency arrival applied for admission when all beds were full they would be turned away. If an elective arrival applies for admission when all beds are full they are put on a waiting list. This led to his specification of the model differentiating between turnaway rate for emergency cases, and waiting times for elective cases.

Shonick calculated a loss function based on the type of arrival. The model also allowed Shonick to calculate the expected time for admissions, and the average daily loss of patients for elective and emergency cases, respectively, for a given number of beds.

There are, however, issues raised by the split between elective and emergency cases concerning the use of the Poisson distribution. He showed that the average daily loss was given by:

$$L = P_n \gamma g(s) \qquad (2.5)$$

and the expected waiting time for admission was:

$$E(W) = 1/\mu(s-\gamma_p)g(s) \qquad (2.6)$$

where s is the number of beds, $1/\mu$ is length of stay, γ is average daily demand, and g(s) is the expected time the facility will be full. However, the model requires a potentially limiting restriction that $\gamma_p<s$, that is the average daily demand from the elective source must be strictly less than the number of beds. This clearly imposes a restriction which

may not be sustainable in other health care systems outside the US where, particularly the UK, there is excess demand for certain, mainly elective, services and queuing is a regular occurrence. For Shonick's calculations to be valid queuing must be an irregular occurrence.

In discussing the model's, assumptions Shonick noted that investigators have indicated that the distribution of daily census in certain facilities is often well approximated by a Poisson. However, the conclusion drawn by some investigators has been that if the facility is rarely full, then the distribution of daily demand may be well described by a Poisson, but if the facility is often full then the Poisson is not a good description of the system. This has led some writers, notably Blumberg (1961), to suggest that if the facility is mostly non-scheduled or emergency admissions, then Poisson will hold, otherwise not. Once more highlighting the rather restrictive assumption that queuing rarely occurs.

Shonick picked up on this point and criticised some authors for using a Poisson distribution to describe fluctuations of the hospital daily census generally. He showed that if the facility was often full, then this would bias the observed mean average daily census, and he went to some length in the paper to prove that. This may have implications for Shonick's model as it will only be applicable for specialties, or distinctive patient facilities (DPFs), where the majority of admissions are non-scheduled, or the facility is rarely full.

Nonetheless, Shonick made a useful distinction between emergency and elective arrivals. His simulations were based on a total demand that divided into elective and emergency and, varying the proportion that was elective and emergency, he calculated the associated turnaway rates and occupancy rates. He concluded that the crucial factor in determining the efficient number of beds is determined by P_n, the proportion of arrivals which are emergency. He did not, however, attempt to assess the size of the emergency demand, rather he provided various simulations for different values of P_n. He recognised this in his conclusions, stating that a significant improvement on the model would involve a complete study of the census fluctuations over a period of time that should at least segregate data by DPFs, tabulate daily arrivals, and classify, however rough, arrivals as emergency and elective.

The main criticism of Shonick is that he did not provide an explicit model of hospital allocation decisions and failed to formalise the trade-offs in a way that allowed an optimal solution to be identified. The model, therefore, became little more than a simulation game.

Shortly after Shonick developed this model Joseph and Folland (1972) developed a model considering how stochastic demand for care affected the optimal size of the hospital. Therefore, as with Shonick, the decision context was the medium to long run since the number of beds was allowed to vary. Joseph and Folland recognised the role unused capacity performs due to peak demands and the inherent trade-off between unused beds, costs, and the cost of turning patients away if beds are full. They suggested that the hospital would attempt to achieve a balance between the costs of excess capacity and the 'hazards' of insufficient capacity.

The problem was set up in terms of a hospital planner's dilemma. Joseph and Folland suggested that increasing the number of beds would increase satisfaction in terms of reducing the rate at which patients are turned away and increasing the number of patients treated in total. They suggested that, in theory, the dilemma be resolved by increasing the number of beds until the marginal cost of providing the extra bed equals the monetary value imputed to the corresponding reduction in turnaway. The formulation of the problem is useful in that it suggests that the hospital planner will aim to maximise a utility function, where the rate at which patients are turned away, and the number of patients treated, both enter as arguments in that function. This suggests that the objectives are linked to social welfare.

Joseph and Folland also recognised the role of uncertainty in the provision of health care and the stochastic demand for care. They acknowledged the role unused capacity performs due to peaks of demand and the trade-off between unused beds, the cost of bed provision, and the costs attributed to turning patients away if bed capacity is not available. The question they addressed was whether a balance could be struck between the costs of excess capacity and the 'hazards' of insufficient capacity, noting that the daily census fluctuates with uncertainty over the year, but that the size of the hospital is more or less fixed during the year, resulting in low occupancy in some periods and turnaway in others.

As with Shonick, they noted that, given the probability mass function of the daily census, it was possible to calculate the expected turnaway for any number of beds, S. Once more, as in Shonick's specification, the larger S the smaller the rate of turnaway, for any given demand.

Given the objective function, increasing S will increase satisfaction in two terms entering the objective function; by simultaneously reducing the turnaway rate and serving more patients in total. However, Joseph and Folland also noted that increasing S increases total costs. They suggested that, in theory, this dilemma be resolved by increasing S until the marginal cost of providing an extra bed equals the monetary value imputed to the corresponding reduction in turnaway. They then set out to prove this solution mathematically.

They made a number of assumptions. First, that the daily census was Poisson distributed; this was consistent with Shonick. Secondly, they assumed that the long-run average total cost function was constant over the range of hospitals. Thirdly, they assumed that the hospital segregated patients into wards, and has the flexibility to change the size of wards to meet demands of daily census. This allowed them to ignore the heterogeneity of output and treat hospital demand as homogeneous.

The aim of the model was to attempt to determine the optimal size of the hospital. The planner faced a daily census with mean, L, and average total cost, C. The expected number of patients turned away each day for a hospital size, S, is:

$$\sum_{R=S+1}^{\infty}(R-S)P(R;L) \qquad (2.7)$$

Where R is the number of patients who demand a bed each day. The expected number of patients turned away for a hospital size S-1 is:

$$\sum_{R=S}^{\infty}(R-S+1)P(R;L) \qquad (2.8)$$

Therefore, the reduction in expected number of patients turned away each day that is attributable to the Sth bed is:

$$\sum_{R=S}^{\infty}P(R;L) \qquad (2.9)$$

which, the authors pointed out, is the probability of obtaining a census of S or above. The problem facing the planner was to choose a value for S. Corresponding to this choice is a daily cost for the last bed and the reduction in the number of patients turned away.

They suggested that, if the hospital planner is a utility maximiser, then they will choose S such that the monetary value of utility lost by the expected turning away of a patient equals the marginal cost of preventing turnaway. On the basis of total costs taken from published data in Iowa they calculated the cost per bed day, and, using results from a study by Feldstein (1961), they calculated the cost of an empty bed. Using these figures they calculated the cost incurred to avoid turning the marginal patient away (i.e. to treat the marginal patient) to be $49,200. They presented this as the cost incurred in reducing the number of patients turned away by one.

However, they stated that it is possible that hospital planners may not be aware of the costs incurred, and if made aware they may alter their decisions. Thus, their calculations are based on an implied value approach, however, they assume that B is chosen on the basis of the cost of providing an extra bed. Assuming costs are equal they find that B is almost the same for all hospitals. If the planner does not know the costs then B will be set arbitrarily. Their findings would then require another explanation. It could be that costs are not in fact equal, although hospitals set B based on what other hospitals do, or that they have a target turnaway rate regardless of costs.

As an extension to their model Joseph and Folland presented a mathematical model of the behavioural model as an appendix. The authors suggested that it is not crucial to the paper, and other models with different assumptions may also be consistent with their model. Nonetheless, this mathematical description highlights an important approach to the problem, and allows further insight into their approach, so is worth discussing in some detail here.

They set up the problem by assuming that the hospital planner is a utility maximiser, subject to a break-even constraint, where utility is a function of the expected number of patients turned away, T(S), and the expected occupancy of L(P) patients. Such that:

$$U = U[T(S), L(P)] \tag{2.10}$$

where utility varies inversely with T(S) and directly with L(P), which is inversely related to P, the daily charge. Therefore, in their mathematical expression the quantity of patients treated and the expected number of patients turned away directly enter the utility function.

However, this is inconsistent with the way the model is developed initially, since the expected number of patients turned away did not directly enter the function initially, i.e. there is no negative utility attached to turning patients away. Rather, the discussion is of the willingness to pay (through increase in costs) to treat one extra patient which, whilst based on the probability of a patient arriving (and hence if a bed was not available the probability of turning a patient away), does not make any reference to the expected number of patients turned away once the bed level is chosen. Although it could be argued that, by implication, if the hospital chooses a bed level, S, they are not willing to incur the implied extra cost to treat one more patient, therefore, they implicitly are taking into account the utility lost in not treating the patient and the disutility of associated with turning a patient away. This is not explicitly recognised in the utility function specified in the mathematical model. This led to problems in the solution derived by Joseph and

Folland; the importance of which should become clear in the discussion of the mathematical solution presented below.

A cost function was outlined, where expected costs for a hospital size S, with expected occupancy of L patients per day is:

$$C = C[L(P)/S].S \tag{2.11}$$

and the hospital was subject to a break-even constraint, such that:

$$PL(P) - C + G = 0 \tag{2.12}$$

Where, PL(P) is expected revenue, C is costs and G is expected gifts of government appropriations. A simple Lagrangian was set up such that:

$$M = U - \lambda[PL(P) - C + G] \tag{2.13}$$

The hospital planner then must select S and P by following the first order condition, which were presented as:

$$\frac{\partial M}{\partial S} = \frac{\partial U}{\partial S} + \frac{\lambda \partial C}{\partial S} = 0 \tag{2.14}$$

$$\frac{\partial M}{\partial P} = \frac{\partial U}{\partial P} + \lambda[\frac{\partial PL(P)}{\partial P} - \frac{\partial C}{\partial P}] = 0 \qquad (2.15)$$

and

$$\frac{\partial M}{\partial \lambda} = -[PL(P) - C + G] = 0 \qquad (2.16)$$

Where the last term simply states that the daily charge should equal the daily cost minus gifts divided by the number of patients.

From the first term they derived the following expression:

$$\lambda = -\frac{\frac{\partial U}{\partial S}}{\frac{\partial C}{\partial S}} \qquad (2.17a)$$

which is:

$$\lambda = -\frac{\partial U}{\partial C} \qquad (2.17b)$$

They suggested, therefore, that the unspecified multiplier could be interpreted as the marginal utility of money (or more precisely costs).

Then by the chain rule:

$$\frac{\partial U}{\partial S} = [\frac{\partial U}{\partial T(S)}].[\frac{\partial T(S)}{\partial S}] \qquad (2.18)$$

and this allows the first term in the Lagrangian to be rewritten as:

$$\frac{-\frac{\partial U}{\partial T(S)}}{\lambda} = \frac{\partial C}{\partial T(S)} \qquad (2.19)$$

They suggested that this equation gives the monetary value of utility lost by the expected turning away of a patient when the hospital is full as equal to the cost that would be incurred to prevent the patient from being turned away.

Further they state that the second term in the Lagrangian can be rewritten:

$$\frac{\frac{\partial U}{\partial P}}{\lambda} = \frac{\partial PL(P)}{\partial P} - \frac{\partial C}{\partial P} \qquad (2.20)$$

which, they suggested, states that the monetary value of the utility lost by increasing the price, and reducing the number of patients treated, is equal to the sum of the changes in revenue and costs.

This, however, is only half a solution and results in problems reconciling the mathematical model with the theoretical solution. In the mathematical model the solution is driven by the utility (or disutility) attached to turning patients away. However, in the intuitive solution, outlined earlier, it is driven by the probability that the bed will be filled and the utility attached to treating one extra patient, without reference to the number of patients who are turned away. This solution is not entirely consistent with the way the problem was set up originally. That is, utility is derived from two sources: the reduction in the probability of turning a patient away and the increase in number of patients treated as a result of the provision of the extra bed. These two appear to have been collapsed into one in the mathematical and intuitive solution proposed by Joseph and Folland. However, providing one extra bed will simultaneously reduce the probability of turning a patient away and increase the expected number of patients treated, for any given demand. The second term is missing from their solution. Consequently, if monetary values were attached to preventing turnaway and treating patients, this would lead to an underestimation of the optimal bed numbers for any given demand.

Let us reconsider the mathematical specification of the model and the reason for these problems should become apparent.

First, Joseph and Folland introduced a break-even constraint into the model, which necessarily involves introducing price and cost. When specifying the model, they suggested that demand is a function of price, and, therefore, the occupancy, or number of patients treated, is a also function of price. Furthermore, they suggested that the expected number of patients turned away from the hospital is a function of the number of beds.

Both these assertions are correct. However, as we know from their previous discussion, the expected number of patients turned away from a hospital is also a function of demand, and it is intuitively obvious that the expected number of patients treated is also a function of the number of beds available, S. Therefore, their specification of the mathematical model is incomplete. The expected number of patients turned away, T, should in fact be represented as T(S,P), and the expected occupancy, or the number of patients treated, should be represented as L(S,P). It should be clear that the expected number of patients turned away, and the expected number of patients treated, are determined by the same factors. Reducing the expected number of patients turned away by increasing the number of beds must simultaneously increase the expected number of patients treated, for a given demand. Similarly, reducing the demand for treatment must simultaneously reduce the expected number of patients treated and the expected number of patients turned away for a given bed supply.

Therefore, the model should be specified as:

$$U = U[T(S,P), L(S,P)] \qquad (2.21)$$

We know that utility is derived inversely with T(S,P) and directly with L(S,P), therefore, we will denote this by negative and positive terms respectively in our equation.

Let us reconsider the solution of the mathematical model using these revised terms. Let us consider the first term:

$$\frac{\partial M}{\partial S} = \frac{\partial U}{\partial S} + \frac{\lambda \partial C}{\partial S} = 0 \tag{2.22}$$

By the chain rule we know:

$$\frac{\partial U}{\partial S} = [\frac{\partial U}{\partial T(S)}].[\frac{\partial T(S)}{\partial S}] \tag{2.23}$$

However, rewriting this using the re-specified equation gives:

$$\frac{-\frac{\partial U}{\partial T(S)}.\frac{\partial T(S)}{\partial S} + \frac{\partial U}{\partial L(S)}.\frac{\partial L(S)}{\partial S}}{\lambda} = \frac{\partial C}{\partial S} \tag{2.24}$$

This equation suggests that hospitals will provide more beds up to the point where the monetary value of the combined marginal utility associated with treating more patients and reducing the disutility associated with turning patients away equals the marginal cost associated with providing the extra bed. As more beds are provided the probability of these two events, i.e. turnaway and treatment will fall, and, therefore, the expected utility will fall. The difference is that the two terms that directly enter the utility function form part of the solution rather than one term in Joseph and Folland's solution.

Similarly:

$$\frac{\frac{\partial U}{\partial P}}{\lambda} = \frac{\partial PL(P)}{\partial P} - \frac{\partial C}{\partial P} \tag{2.25}$$

now becomes the monetary value of the utility lost by increasing the price and reducing the number of patients treated, and the utility gained (or disutility reduced) by reducing the number of patients turned away, is equal to the sum of the changes in revenue and costs.

Therefore, the first order conditions are specified in very general terms, and the first order derivatives of the cost function are not actually calculated. Furthermore, the cost function is specified very loosely as being a function of occupancy rate and number of beds. This is likely to be an under-specification, although it is a useful way of considering the problem.

There are other issues that the mathematical model raises which are worth mentioning briefly. In this model costs are not linked to occupancy rate. This may miss a potentially important relationship between higher occupancy rate, and lower costs, and increased probability of turnaway, which the authors originally recognised. Furthermore, the model suggested that the level of beds is chosen based on the turnaway rate, however introducing a break-even constraint may significantly alter this, for example it may not be possible to satisfy both simultaneously. Additionally, since arrivals are stochastic, the mathematical model should be set up using expected utility theory, since the decisions are all based on expectations not certainties.

Finally, Joseph and Folland's actual empirical model suffers from not actually specifying a cost function, but nonetheless, resting many of their findings on assumptions regarding the cost function. For example, the hospital planner's dilemma is set up as one where they will increase S until the marginal cost equals the money value imputed to the corresponding reduction in T(S), these assumptions are fairly important. Furthermore, they made the assumption that the long-run average total cost function for hospitals is constant over the sample in their data. They suggested that if the cost of an extra bed is constant, then the optimal values of S corresponding to different expected daily censuses will all give the same probability of obtaining a census of S or above. This allows a constant value for B to be imposed across all hospitals. The fact that their empirical results confirm this does not necessarily support their assumptions regarding the cost function, or indeed that all hospitals have the same target turnaway probability. Nonetheless, Joseph and Folland's work did represent an important step forward and led to the one of the seminal works in the analysis of hospitals; the work by Joskow (1980).

Joskow's work highlighted the domination of analysis of hospital behaviour and performance in the US as one of concern with hospital beds and occupancy rates. He explored the characteristics of hospital bed supply planning in the context of a simple queuing model. The model was used to examine hospital bed supply decisions and what he termed the 'reservation quality' of the hospital.

A key issue in the work was the identification of the optimal amount of capacity for a hospital, or area, and the idea that this was determined by the appropriate probability that a hospital will be full and patients turned away, or queued. Joskow suggested that the appropriate value for this probability depended on five factors: the kind of patients served by the hospital, i.e. emergency or elective; the distance to other hospitals and availability of services to which the patients might turn; the admission or queuing discipline used by the hospital; the value patients put on a rapid admission; and the costs of maintaining various levels of hospital capacity. Therefore, this work drew together many of the strands of work identified by other authors.

Joskow identified a number of important issues, noting that the hospital's allocation decision will in some way be attempting to optimise an (unspecified) objective function. He noted that the different types of demand and the costs associated with turning patients away or queuing patients may be important determinants of an optimal allocation rule, and that maintaining excess capacity may be costly.

He briefly discussed the literature concerning hospital objective functions noting that there had been a wide range of hypothesised objectives. He noted that, generally, the hospital is viewed as a monopoly supplier 'characterised by some objective function over quantity, quality, and scope of services'. He suggested that, for his model, the specific form of the objective function was not particularly important, and that all that was required was a constraint that was binding, and that the value of the objective function increased as the number of patients or quality/scope of services increased.

Joskow considered a hospital facing an exogenous demand for admissions and assumed that, when full, the hospital turns patients away and that the hospital has an objective of achieving a target turnaway probability. Therefore, despite suggesting that the specific form of the objective function be unimportant, he implied a general objective function of quality/quantity maximisation. Joskow, in fact, specifically imposed an objective function where utility is derived solely from a target rate of turning patients away. Thus, he specified a very explicit objective of achieving a target turnaway probability, which he termed 'reservation quality'.

Joskow suggested that an accurate estimate of the reservation quality might be useful for making welfare judgements about the level of bed capacity in the US system. He stated that, in general, as the reservation quality increases the marginal cost of treating an extra patient increases. To illustrate this point he calculated the marginal cost per additional patients admitted for a range of different values of reserve capacity. He suggested that the marginal cost per additional patient treated could easily reach a level where legitimate questions could be asked about the inefficiencies resulting from the provision of too many hospital beds.

Joskow recognised that the demand for hospital services is stochastic, which was the starting point of his analysis. Demand uncertainty, he suggested, meant that the hospital must attempt to meet peak demands and, as a result, may be operating at full capacity during only a few days of the year, turning some patients away or increasing delays in admission during these periods.

He noted that the appropriate value for the turnaway probability depends, amongst other factors, on the kind of patients served by the hospital, i.e. emergency or elective. Although, as with Joseph and Folland, Joskow viewed the hospital as a single organisational entity where all beds substitute perfectly for each other and this, he recognised, was one of the weaknesses of his approach. Therefore, despite recognising the importance of heterogeneity of output and demand, the model failed to fully incorporate these factors. In reality he notes that DPFs may exist that are earmarked for particular treatments. As a result the demand characteristics and bed supply of each individual facility are relevant for planning purposes, rather than aggregate demand as a whole.

Joskow found that there are potentially interesting implications of the relationship between demand and occupancy rate. For a given reserve capacity, as the average daily demand increases so does the occupancy rate, and, therefore, the average cost of maintaining spare capacity falls. This, Joskow noted, has implications for regulatory efforts to set a standard occupancy rate and favours relatively large providers.

Joskow's model was very similar in specification to Joseph and Folland's in that he specified an exogenous demand for admissions arriving according to Poisson process.

Using the properties of the Poisson he suggested that it is possible to set the probability of the hospital being full at any level by choosing how many standard deviations, k, away from the mean (ADC) the number of beds (BEDS) is to be:

$$\text{BEDS} - \text{ADC} = k\sqrt{\text{ADC}} \tag{2.26}$$

Thus he specifies the reserve margin of the hospital, R where:

$$R = k\sqrt{\text{ADC}} \tag{2.27}$$

The greater k, the larger the average reserve margin, and the smaller the probability that the hospital will be full and patients turned away. In the context of his model k becomes a target for the hospital. He stated that, in general, as the value of k increases, (as additional beds result in a smaller number of patients not turned away), the marginal cost of turning away fewer patients increases. He gave this some empirical content by calculating the marginal cost per additional patient admitted to show how the marginal cost increases with k. He suggested that the optimal value of k depends on the direct comparison of marginal cost and the value to the marginal patient of being admitted rather than turned away. This is the same presentation of the problem as Joseph and Folland.

The maximisation problem, therefore, became one of weighing up the marginal cost of treating the extra patient and the value to the patient of being treated rather than turned away. Therefore, this is exactly the same specification as Joseph and Folland in that two elements are included in the value of the providing the extra bed; the value attached to treatment and the reduced disutility associated with reducing the number of patients turned away.

Joskow also dealt with the issue of constraints to behaviour. He suggested that there were four characteristics of the market that may constrain the ability of the hospital to maximise its objective function: break-even constraint; the nature of demand; medical technology; and government regulation. He focused on two factors that may affect the hospital supply decisions: inter-hospital competition and government regulatory efforts to constrain the expansion of hospital beds. However, in considering regulatory effects Joskow's model suffers from its lack of exposition of the underlying economic theory, and he did not really consider the objective function of the hospital and the interaction with the constraints he identified.

He highlighted the certificate of need regulation (CON) introduced in the 1960s, the aim of which was to constrain hospitals from building facilities that were not needed. Given demand the CON agencies tried to ensure that the number of facilities satisfied demand at a minimum cost, the prevailing utilisation criteria was generally based on 80-90% occupancy rates. Joskow stated that in the context of his model CON regulation should reduce reserve capacity. Furthermore, this type of regulatory constraint may even move the focus of the problem from one of attempting to determine the optimal numbers of beds, as originally set out by Joskow, to one of determining the optimal use of existing bed capacity.

The other regulatory control Joskow considered was that of using a prospective reimbursement system. Joskow noted that, in theory, if a prospective reimbursement system succeeds in constraining a hospital's expenditure in some way then a supply response would be induced. In particular, he noted that combining an utilisation criterion (e.g. occupancy rate) with a reimbursement formula could provide strong incentives[2]. Joskow recognised that reimbursement formulas provide incentives. However, what he did not draw out was that the incentives they create are also affected by other factors, which he earlier recognised as being important determinants of bed supply decisions. That is, the objective function, and the constraints within which the hospital operates.

Joskow also considered non-price competition between hospitals. Assuming hospitals have some objective function over the quality and quantity of services provided and, given that patients have some choice among hospitals in their area and have extensive insurance coverage such that prices have an insignificant effect on hospital choice, Joskow suggested that quality is likely to determine hospital choice. He suggested that the more intensive

[2] There exists a branch of literature that considers reimbursement issues within the context of the principal-agent relationship, see for example Berki (1983), Ellis and McGuire (1986 and 1991), and Eby and Cohodes (1985) and Morrissey et al (1984) for a review of studies. This literature will not be considered here as it concentrates on optimal reimbursement strategies under quantity/quality objectives, rather than the reaction of capacity utilisation to changes in reimbursement, and focuses mainly on the impact on costs.

the competition in an area the greater the quality of services are likely to be. Furthermore, he suggested that the probability of being turned away and the expected admission delay are likely dimensions of quality from the viewpoint of the patient, hospital and physician. Therefore, in the context of his model, the greater the competition the higher the reserve capacity will be.

Joskow estimated an empirical model to consider the impact of competition and regulatory controls on reserve capacity in the hospital, and found support for his theory that competition and CON regulation affect the reserve capacity of hospitals. However, as noted above, his model suffers from the lack of foundation in theory and failure to fully integrate into the model the influences he highlights as important considerations of supply decisions. Having noted the various constraints he does not actually attempt to set the model up as a maximisation problem subject to any of these constraints.

Rafferty (1971) developed a model with the aim of identifying and explaining short-run variations in case-mix of a hospital's census. He suggested that case-mix is an important characteristic which, thus far, had received limited attention. He built his work on a theoretical definition of hospital output, drawing on Feldstein (1967), which recognised that the marginal social value of hospital treatment will differ for different treatments, and that this will be in direct proportion to the degree to which treatment is more or less discretionary. This potentially allowed, what Rafferty termed, a 'maximising tendency' to be identified from which the resulting patterns of case-mix variations may be predicted. This suggested that the social value of hospital services would be maximised if, under the constraint of limited bed capacity, admission priorities were allocated on the basis of medical need.

This suggested that hospitals may pursue an objective that could be loosely defined as welfare maximisation. Whilst not explicitly defining an objective function, the implication was that with a fixed bed capacity, hospitals will allocate beds on the basis of social value, and this will be directly related to the type of case treated, where case type can be broadly split into 'necessary' and 'discretionary'. However, it is not immediately clear that the efficient outcome of this decision rule would imply that those in greatest medical need would receive care first, as suggested by Rafferty, as it fails to consider cost. A case with twice the social value of another case would have to have a cost less than twice that of the less valued case for this allocation rule to be efficient.

Rafferty stated that evidence suggests physicians do commonly consider the availability of beds before admitting patients. As a consequence, an increase in the incidence of illness and the resulting increase in occupancy rate could be expected to reduce the number of discretionary admissions, and the case-mix properties would respond to variations in the relative availability of beds.

This recognised a potentially important trade-off within the hospital when considering bed allocation decisions. That is, hospital admission policies will depend on the demand for different treatments. Rafferty suggested that as the demand for all types of care increases less elective patients would be admitted. He also recognised that aggregate analyses of hospital demand may miss the essential nature of the hospital output, and that knowledge of case-mix is essential in addressing issues of efficient and optimal use of hospital capacity.

Despite this, he failed to fully recognise demand heterogeneity. What may be more accurate is that as the demand for *emergency* care increases less elective patients are admitted. This hints heavily at an objective function of the hospital that is linked to supply assurance for emergency patients, and where not turning patients away enters the utility function directly.

The implication is that hospitals will be attempting to maximise the social value of treatments, however, this was not explicitly drawn out in Rafferty's work. For example, he did not consider why hospitals operate with lower occupancy rates in some periods than others if there is a pool of discretionary demand they can call on, or what drives the hospital to change admission policy when occupancy rate rises.

If we follow this line of argument, then it seems likely that there are two admission sources that roughly correspond to necessary and discretionary admission; these are planned and unplanned admissions. In the Rafferty model, when the occupancy rate reaches a certain critical level, the hospital would stop taking any more planned admissions and devote all beds to unplanned admissions, although this would involve cancelling planned admissions, and would suggest that in fact the planning had not been particularly successful. Additionally, it fails to recognise that the hospital may be concerned with the extent to which it cancels planned admissions and, by implication, the size of waiting lists.

The problem is apparent; the model failed to incorporate the stochastic nature of unplanned admissions, which intuitively drives the whole problem. The fluctuating occupancy rate may be related to the desire not to turn emergency patients away and, rather than there existing a critical occupancy rate when discretionary patients are no longer admitted, it may be that hospitals plan for fewer discretionary admissions to fit in with their expectations regarding emergency arrivals. Therefore, any observed fluctuations in casemix may be due to hospital planning.

However, without specifying an explicit objective function that recognises the two types of admission and their characteristics, it is difficult to determine the appropriate criteria for identifying optima. As with other authors, his model suffers from a lack of a formal link between the theory and the model. He suggested that estimated hospital cost functions have found costs vary with case-mix, and that since some case-mix patterns found in his analysis are related to occupancy rate, that part of the change in costs could be due to different services being provided at different utilisation levels.

Rafferty concluded that, given the increased intensity of use of hospital facilities, it seems reasonable to reduce admission for certain cases and, conversely, when bed capacity increases, or incidence of illness declines, it seems reasonable to increase admission of discretionary cases.

He found evidence to support his theory, whereby admissions for less serious conditions fell during periods of high occupancy, and he suggested this was quite probably due to conscious rationing on behalf of medical staff. Furthermore, he found that the overall number of admissions was not very closely related to occupancy rate, rather he found that changes in length of stay were the major cause of changes in occupancy rate. This has implications for subsequent analyses of length of stay, which is often treated as an exogenous determinant of hospital utilisation. Rafferty suggested that length of stay may not only determine occupancy rate, but also be induced by changes in occupancy rate. However, as he himself recognised, as occupancy rate rises and rationing of elective cases occurs, if priority is given to emergency cases with a greater length of stay, then increased occupancy rate may occur without any change in total admission or discharge policies.

This raises the issue of displacement of planned cases, and whether hospitals will be concerned about the rate at which this occurs. Once more, based on expectations of emergency demand, hospitals might have some idea of the extent to which this event will happen. Furthermore, there must be a limit to how many beds they can make available at very short notice. They can discharge early, or stop admitting planned admissions up to a point, but the extent to which this is feasible will depend on the number of patients scheduled for admission and for discharge, and it is likely there will be a limit to the extent to which patients can be discharged early based on medical requirements. If hospitals do not want to turn serious cases away, and they also have a concern about the length of waiting lists, or the number of cancellations, then these factors must be weighed against each other. That is, they should enter the objective function directly. Hospitals must also make ex ante decisions about allocations, and the requirements for reserve capacity, based on expectations of demand. If the demand for emergency care changes then this will alter these allocations. This may be the source of fluctuations observed by Rafferty.

It seems intuitively obvious that the hospital will have some idea about the level of unplanned admissions and will allocate capacity accordingly, however, implicitly in this model, the hospital does not want to turn away serious cases due to the high social value attached to treatment. Therefore, hospitals may displace discretionary cases when they consider that the probability of turning patients away is getting too large, based on the expectations of unplanned demand.

The main problem with the model is, once again, its lack of exposition of the underlying economic theory. Many of the parts are in place; the discussion of objectives of the hospital; recognition of the importance of heterogeneity of demand; and the essential nature of the trade-off between using capacity to treat serious and non-serious cases. However, he failed to bring these parts together in a formal model of behaviour.

In an attempt to develop a theoretical model for explaining regional differences in the utilisation of hospitals, Chiswick (1976) noted that the general concern with the level and distribution of hospital services stemmed from the adverse effect of delayed treatment arising from insufficient bed capacity on one hand, and the costs of maintaining unused beds on the other. The aim of his work was to develop a theoretical model for explaining regional differences in the utilisation of general hospitals.

The model assumed that hospital bed capacity is fixed in the short-run, and he graphically represented the short-run demand and supply for hospital admissions on a standard demand/supply graph. The price of an admission was defined in what Chiswick termed 'its broadest sense', that is, it included the value of extra discomfort, loss of earnings, and curative costs caused by delayed admission and the poorer quality service that may arise from crowded hospitals. The demand for admissions was assumed to be negatively related to price. Supply was assumed to be positively related to price with an absolute upper limit based on the capacity constraint.

Chiswick's model, in its attempt to remain within the simple demand-supply framework, rapidly becomes over-complicated. For example, it is not immediately clear, and Chiswick does not draw out, why supply will respond to any element of price other than the monetary element, although he specified a positive relationship between shadow price and quantity supplied.

In his specification of the problem he suggested that if short-run fluctuations in demand lead to a shift in demand upwards to the right then this will increase price, due to the increased probability of delay or denied treatment, but at the higher price hospitals will supply more admissions and increase occupancy rate. However, the mechanism through which the supply response operates is not at all clear.

There is a possible explanation why hospital supply may respond to the shadow price that depends on the objective function. If the hospital takes account of the social costs, as assumed in this model, then they must, by definition, have an objective function that aims to maximise social welfare, such that supply not only responds to monetary price, but also to other factors that enter the utility function. Whilst the mechanics of response are not clear, the implicit objective function appears to be one of social welfare maximisation.

Chiswick noted that the randomness of short-run demand for admission is an essential aspect of the analysis of occupancy rates and bed rates. In the model he suggested that the height of the short-run demand curve is, at any moment in time, a function of what he termed 'systematic' and 'random' elements. Such things as the demographic characteristics of the population and the extent of health insurance determine the systematic elements. At any moment in time, at each price of admission, each individual has a probability of demanding an admission. The random elements are said to be due to 'the aggregation across individuals of the outcome of this random process'. Consequently, the short-run demand curve fluctuates randomly about its expected value.

Chiswick, as others have done, assumed perfect bed substitutability within hospitals, but not between hospitals in a geographic area, citing three reasons for this: differentiation among hospitals in the demographic characteristics of the patients they admit; the fact that patients may view hospitals as imperfect substitutes; and physicians may have a limited number of affiliations thus restricting choice of hospital.

However, there is another reason why beds may not be perfect substitutes within a health region, that is the heterogeneity of product and the geographic nature of some markets. This is most notable in the market for emergency services, where delay can result in very high costs, and patient choice is often limited to those hospitals within a very limited geographic area. Furthermore, this only really applies to emergency cases. Therefore, it is important to distinguish between emergency and non-emergency cases. Chiswick did, albeit in passing, note that the main costs of increased travel time would be imposed on emergency cases. However, he does not fully develop the issue of demand heterogeneity.

Furthermore, the model focused on the issue of how many beds are required by a community, based on the probability of turning patients away and the cost of providing an extra bed. This, however, does not deal with all the arguments entering the utility function and fails to deal with capacity utilisation. In systems such as the UK one of the key issues is how to allocate scarce resources. With beds limited in the short to medium term, the relevant question is how best to use existing beds, not how many beds are required. This leads to different concerns such as the trade-off between different types of care.

Once more, Chiswick's model, as others before, failed to formalise the problem in a way that drew out all the relevant arguments in such a way as to allow an optimal solution to be identified.

Few of the models have looked at capacity utilisation, demand uncertainty, output heterogeneity and their impact on production. The very few that did, which we have detailed here, have not presented formal models. All the models are either ad hoc or empirical in their approaches. This lack of formalisation has led to their being little ability to generate testable hypotheses with the result, as we shall see, that most of the empirical analyses relating to cost and production functions remain ad hoc also.

Section 3: Empirical estimates of hospital cost functions

Interest in specifying and estimating cost functions for hospitals has grown in the last three decades, largely reflecting a greater availability of data. The aim of this section is to provide an overview of the theoretical underpinnings of the empirical estimates of hospital costs, in particular focusing on recent advances in the area that incorporate production responses to demand uncertainty.

There have been a number of extensive reviews of hospital cost function analyses. Breyer (1987), Cowing et al (1983), and Wagstaff (1989) in the UK, provided a fairly comprehensive review of over 80 studies that have attempted to estimate hospital cost functions, providing a useful chronicle of how cost function estimation has evolved. It is not our intention to replicate these studies here, nor is it necessary. Rather we will provide highlights of the development of the different methods employed by considering the theoretical and empirical bases. First let us consider the theoretical foundations.

In the neo-classical theory of the firm the cost function describes the minimum cost of producing a given output as a function of the exogenous vector inputs of prices. Duality theory demonstrates a one-to-one correspondence between the production function for a good and the respective minimum cost function, and as such represents a technical relationship. This concept of the cost function has important consequences for the specification of a regression equation if observed costs are interpreted as minimum costs.

Independent variables may only comprise output quantities and input prices. Other potential independent variables, such as capacity utilisation, cannot be included in this type of specification because they do not determine the minimum cost. Rather they explain deviations from the theoretical minimum. Examples of this approach include Conrad and Straus (1983) and Cowing and Holtman (1983). In both cases the authors assume that the observed data represent cost-minimising behaviour (at least in the short-run).

There has, however, been a movement away from the traditional production-cost duality condition through the employment of behavioural cost functions, which recognise, inter alia, that demand could be endogenously determined. Evans (1971), most notably, developed an approach, which he termed a 'behavioural' approach, that expresses the idea that the cost function represents a behavioural and not merely a technical relationship. The argument is that a wider range of effects needs to be taken into account and, by definition, cost minimising behaviour is no longer pursued. Evans moved away from the idea that demand was exogenously determined. Rather, he suggested that a more realistic model would recognise that the demand for hospital services is generated by medical practitioners who weigh up the relative costs and benefits of hospitalisation (to themselves and the patient). Thus, creating a situation where inputs and outputs are determined jointly.

This behavioural theory was in many ways an extension of earlier work by Liebenstein (1966) and Cyert and March (1963), which moved away from maximising behaviour. Key notions of this approach focused on satisficing behaviour, multiple goals, organisational slack and other behavioural characteristics.

De Alessi (1983) built on this work and proposed an alternative approach to neo-classical theory that took account of institutional constraints, focusing on property rights and transaction costs. The theory can be generalised to include any constraints that affect decision makers and limit choice. The individual characteristics and the economic context according to Liebenstein, determine the degree of deviation from maximising behaviour. The application of some behavioural theories to the hospital sector did not so much stimulate output, but led to a diminution in the attention paid to cost function specifications.

More recently, however, the duality approach had been revived through the use of flexible functional forms that allow approximations to the minimum cost to be specified. If the production function is unknown, then it is impossible to specify the form of the minimum cost function accurately, increasing the risk of misspecification. Specifying a flexible functional form allows an approximation to the minimum cost function to be estimated and the restrictive assumptions, like separability, can be dispensed with (see, for example, Cowing and Holtman, 1983; and Conrad and, Straus 1983). The most commonly used flexible functional form is the transcendental logarithmic function, where all the variables are transformed logarithmically and a Taylor series expansion approximates the 'true' cost function. The increased flexibility is, however, at a cost; the number of parameters grows almost proportionately to the square of the original number of regressors. This is a particular problem when estimating multiproduct cost functions, (see Caves et al., 1980), and can cause problems of multicollinearity between the regressors, (see Vita, 1990; and Dor and Farley, 1996). The translog, nevertheless, remains one of the most popular functional forms in UK studies (see McGuire and Westoby, 1983). The Cobb-Douglas specification has also been used, (Feldstein, 1967), in UK based estimates. The latter specification represents a restricted version of the flexible functional form estimated by a translog specification and, obviously, implies a particular functional form. The issue of functional form clearly relates to the production process and this is difficult to specify. Wagstaff (1989) highlighted the problems of using a Cobb-Douglas specification when estimating factor substitution. Estimates by Feldstein (1967.) and Lavers and Whynes (1978) both utilised a Cobb-Douglas specification and found exactly the opposite results regarding factor substitution between nurses and doctors, i.e. Feldstein's results indicated too much was being spent on nurses relative to doctors and Lavers and Whynes found exactly the opposite.

Econometric considerations such as these have led to a particularly constrained choice of output. In most cases studies have relied on measures of throughput such as patient days or cases treated and, therefore, specifyied a homogeneous output. Clearly, where output is not homogeneous the implication is, as highlighted in the discussion earlier, that equal weights are attached to each case. This has led to fairly aggregate patient classifications to reflect the multiproduct nature of hospital production. For example, Cowing and Holtman (1983) divided patient days into five diagnostic categories, Conrad and Straus (1983) by three based on patient age. Feldstein (1967) divided patients into groups according to admissions department, and then weighted each case by expected average cost. More recently, studies have adjusted for output heterogeneity by including a casemix variable to adjust for case complexity (see, for example, McGuire and Williams, 1986).

The apparent dilemma, where an accurate reflection of patient heterogeneity seems to be too demanding for the flexible functional form, has led to authors such as Granneman et al (1986) to specify what they term a 'hybrid functional form'. The regressors include variables other than output quantities and input prices, but the functional form is homogeneous in factor prices. However, Breyer (1987) criticised this approach noting that linear homogeneity is not strictly imposed. Thus, most work has relied on conventional assumptions in the estimation of cost functions. Econometric, rather than economic, considerations have dominated with cost minimisation and duality assumed.

One of the fundamental aspects of the environment hospitals face is uncertainty over demand for services. Despite this, the majority of estimated hospital cost functions have assumed that demand is known. The neo-classical theory of hospital cost and production requires that production be technically efficient, that is that production occurs on the boundary of the production possibility frontier. When a firm faces stochastic demand this assumption no longer holds.

Feldstein (1967) was one of the first to consider the general issue of technical efficiency, noting that when the hospital operated off its estimated production frontier that this indicated whether the hospital was technically more efficient (or inefficient) than the average hospital. He, therefore, suggested that the residual of the production function could be used as a measure of technical efficiency (see also Wagstaff, 1987, for an example of this approach). Hospitals with positive residuals could be considered to be producing more than expected, those with negative residuals, less than expected. Feldstein used the ratio of actual output to the level of output predicted by the production function as a measure of technical efficiency, proposed by Farrell (1957). The main problem with frontier analysis of this kind is that it fails to define the cause of the technical inefficiency, i.e. there is no theoretical underpinning, and it implicitly assumes that all cross-sample variation is due to technical inefficiency. Furthermore, using of a ratio measure of actual to forecasted output will not pick up the absolute size of the divergence from the production frontier, although it does allow ranking of units by efficiency.

The first authors to explicitly recognise the issue of demand uncertainty were Friedman and Pauly (1983), who noted that since hospitals face demand uncertainty then it is not possible to assume that inputs were hired to produce at minimum cost. Rather inputs have been hired to minimise ex ante expected costs, but, crucially, these may differ from ex post realised costs. The divergence between the ex ante expected costs and the ex post observed costs will depend on the error in forecasting admissions and, therefore, any econometric model of the realised costs should consider the impact that unexpected demand, i.e. unforecasted demand, has on costs.

Friedman and Pauly highlighted the apparent anomaly in empirical studies where short-run marginal costs appear to be below average costs. This result is at odds with a cost-minimising decision-maker that has perfect knowledge of demand conditions that would operate at the efficient minimum point rather than on the downward part of their short-run average cost curve. However, this may well be explicable as a rational response to demand uncertainty when some of their costs are fixed. That is, if decisions regarding inputs are made ex ante against the background of uncertainty and some of these costs are fixed ex post, i.e. there is no ex post adjustment, then this may be rational. Against this background, Friedman and Pauly model the expected cost function as:

$$E\{C_t\} = K + wL_t + s\int_{F(L_t)}^{\infty}[R(q_t) - L_t]h(q)dq_t \tag{2.28}$$

where $E\{C_t\}$ is expected costs, K is fixed charges, L_t is quantity of inputs, w is cost per unit of inputs, $R(q_t)$ is the mapping of outputs to inputs, $[R(q_t)-L_t]$ is the input shortfall, and s is the rate per unit of the value of this loss and h(q) is the probability density of demand. The hospital chooses L_t to minimise this expectation.[3]

Gaynor and Anderson (1995) adopted this approach modifying a model developed by Duncan (1990) for the telecommunications sector. Their primary aim was to provide an estimate of the cost of an empty bed. They specified the hospital's problem as one of minimising cost subject to a production function that incorporates an adjustment for demand uncertainty, and a constraint that the probability that the hospital is full at any given time does not exceed a pre-determined target level. It was assumed that the hospital can adjust its variable input on the spot market once demand was realised. This differs from Friedman and Pauly who allowed quality to deteriorate, (resulting in delays for treatment), and that this will manifest itself in the form of 'latent penalties' (reduction in unit price attainable, or reduction in goodwill of physicians). Nonetheless, the concept was very similar. The cost function has all the usual properties except duality, since the firm is constrained to have the capacity to meet randomly fluctuating demand with some probability, and will, generally, not be producing on the production possibility frontier.

Carey (1996) adopted a similar estimation approach as Gaynor and Anderson, and Friedman and Pauly to consider a question originally addressed by Joskow (1980) concerning the socially optimal level of excess capacity. This involved estimating the

[3] Pauly and Wilson (1986) extended this work using different data, but the underlying theory was the same.

cost of an empty bed and included forecasted demand in the cost function estimate. Carey, however, developed the model a stage further and set up an optimisation problem whereby the hospital will attempt to minimise expected costs, where costs include the social cost of turning patients away, which given the distribution of demand will occur with some probability. This represents a subtle development from the model Friedman and Pauly specified as the model includes social costs as well as private costs, explicitly recognising that the objective function of the hospital includes arguments outside the narrow private costs and benefits usually assumed.

Setting the problem up in a similar manner to Friedman and Pauly, Carey derived the first order condition. The optimal solution occurs where the ratio of the hospital costs of staffing the last bed to the social cost of not having a bed available, is equal to the probability of turnaway. The optimisation problem that the hospital attempts to solve in Carey's model is to choose the number of staffed beds that will minimise expected costs E(C).

$$E(C) = K + r(B)B + v(V)\int_0^y g(y)f(y)dy + s\int_y^\infty [h(y)-B]f(y)dy \qquad (2.29)$$

Where K is capital, B is staffed beds, V is the variable input, f(y) represents the probability density function of demand, g(y) is the level of occupancy variable input V required to treat output y, h is the mapping of output to beds, and s is the social cost of not having enough beds to treat all demand. The hospital must, therefore, choose the number of beds to optimise this condition, where the hospital provides output to treat demand y. The first order condition derived by Carey suggests that the optimal condition is where:

$$[(\delta r/\delta B*B + r)/s] = [1\text{-}F(y)] \qquad (2.30)$$

That is where the marginal cost of the last bed equals the social cost of not having a bed available.

However, the optimal condition is actually more complicated than this, since Carey failed to recognise that the demand the hospital can treat, y, is determined by the number of beds available. Consequently, B enters as a limit in the integrals as well. Therefore, differentiating with respect to B yields a more complicated first order condition[4].

Furthermore, whilst Carey represents a move forward in terms of integrating theoretical and empirical work the model only allows for one output, treating output as homogeneous. It is widely accepted that some demand can be queued and some cannot. If output is treated as heterogeneous then demand should be separated into, at least, two.

---

[4] Carey sets up the optimisation problem as:

$$E(C) = K + r(B)B + v(V)\int_0^{y} g(y)f(y)dy + s\int_{y}^{\infty}[h(y)\text{-}B]f(y)dy$$

Where K is capital, B is beds, v is variable input, f(y) is the pdf of demand, h(y) is mapping of output to beds, s is the social cost of having insufficient beds, g(y) is the level of occupancy variable input required to treat output y. She then sets up the problem where the capacity is constrained based on the hospital's choice of beds, B, where the hospital will choose to serve demand . However, when differentiating with respect to B she fails to recognise that $\dot{y}$ is also determined by B (i.e. $\dot{y}$ = B/LOS). Thus, the third term in her equation also features B as a limit to the integral. Therefore the first order condition she derives is incorrect and missing one term.

If the reason for spare capacity is primarily to prevent turning patients away, the probability of turning away different cases may have different social costs attached. Therefore, the model can only deal with total bed supply rather than allocation of internal capacity, and in this context it is somewhat limited.

Nonetheless, the work by Carey is consistent some of the earlier work identified in section 2.1 (most notably Joskow, 1980; and Joseph and Folland, 1972) and draws together some of the empirical work by Friedman and Pauly and Gaynor and Anderson into a single model. As such this represents the first work of this kind.

Section 4: Objective functions, model specification and modelling issues

What is clear is from section 2.1 is that there is no general consensus about the appropriate objective function. There is also a problem operationalising objective functions in general. Having noted this however, almost all commentators suggest that quantity and some other factor will enter the utility function, which implies a trade-off between quantity and other factor(s). Utility maximisation, regardless of whether producers' or consumers', features largely in most specifications.

The main focus of this thesis is on the heterogeneity and uncertainty of demand for hospital services. This makes some of the hypothesised objective functions redundant since they assume a single output firm. However, some of the specifications have offered useful insights into how decision-makers may deal with heterogeneous outputs, although not many consider the stochastic nature of demand, with one notable exception. Harris analysed the nature of the hospital firm with specific reference to the uncertainty, and offers a useful starting point to any specification of the objective function given our own objectives. He suggested that hospitals might be interested in things other than output per se, most notably, supply assurance.

Whilst most commentators specifications can be viewed from a Williamson perspective of utility maximisation there is no general agreement on what enters the utility function, or how it is specified. However, most analyses, which consider heterogeneous output, do suggest that there are trade-offs between different services and the output level will depend on the relative weights attached to each output. If we integrate these approaches with Harris and apply this to our original problem, we arrive at a specification of the objective function of the hospital specified in its broadest terms, i.e. utility maximisation, with explicit trade-offs between those factors entering the utility function and weights attached to each factor.

These grand theories of the objectives of non-profit institutions are, however, often difficult to operationalise and tell us little about internal capacity utilisation decisions. Operational maximands have been dealt with variously in the literature and some have not drawn out the issue fully. This is partly due to the focus of the models, i.e. with total bed requirements, rather than capacity utilisation; nonetheless, they do offer useful insights into the problem and raise a number of important issues.

It has widely been acknowledged that the random nature of demand for hospital care is a crucial part of the model specification. How this actually effects the allocation process, however, depends on how patients are treated, and it is likely that it is possible to place some patients in a queue, whilst others require immediate care. This brings us to another of the key issues, heterogeneity of demand. The issue of output heterogeneity has been dealt variously within the existing models. Despite the promising early work by these authors, most notably Shonick, their work was not really built on.

The theoretical models have largely been developed in isolation without reference to the more complex theories of the non-profit firm, which have developed in tandem. This is partly due to the difficulty in operationalising the grander theories, but also reflects the lack of integration in general within the field of hospital economics. The lack of formalisation of the problem has led to descriptive models that have been unable to identify optimal solutions to the problem. Since the promising early work by Shonick, which recognised many of the important features of the environment within which hospital operates; uncertainty, demand heterogeneity, etc, progress has been slow. The work by Joskow marked an important step forward in trying to empirically analyse the reasons why hospital held reserve capacity and the different impacts of the level of reserve capacity.

The empirical estimates of cost functions have, largely, been estimated without reference to the theoretical or mathematical work. The theories underpinning the estimation processes have, upon closer inspection, often been untenable given the environment within which hospitals operate. Only recently have their been moves away from traditional neo-classical theory to attempt to take account of, in particular, demand uncertainty and the impact this has on cost function estimation.

This recent work has attempted to bring together all three strands of hospital economics. The study by Carey has gone some way to integrate the three distinct strands that have developed, however, this work is still has some outstanding problems. Most notably, it failed to recognise output heterogeneity and only allows consideration of total number of beds rather than the use of beds within a hospital; this may stem from the lack of formal consideration of the theoretical basis of the model. Furthermore, the optimal condition specified in the paper, and consequently, the derived first order condition, is not fully specified. Notwithstanding these issues, Carey's work does represent a development within the field, which should point the way forward for future work. The intention of the remainder of this thesis is to consider a theoretical, mathematical, and empirical specification of the issue of optimal use of hospital bed capacity within the framework of demand uncertainty and output heterogeneity, building on the existing work reviewed in this chapter.

Chapter 3: An integrated approach to the theory of hospital behaviour

Section 1: Introduction

The previous chapter described some of the approaches adopted by other commentators when modelling hospital behaviour. Three areas were highlighted in this discussion as important when considering the behaviour of hospitals; the objective of the hospital; the heterogeneous nature of services; and the stochastic demand for services. However, none of these studies have attempted to build an integrated approach to modelling hospital behaviour from a solid theoretical background allowing an empirical estimation that is consistent with the theoretical model. Furthermore, previous models have not dealt with the issue of internal capacity utilisation given the nature of hospital demand and output, which is key to the whole exercise here. The aim of this chapter is to bring together the main theoretical features of a model of hospital behaviour, and consider the optimal allocation of available capacity given these features.

The conceptual issues regarding the objective of the hospital, the nature of the demand for hospital services, and output of hospitals, have been dealt with varying degrees of sophistication, as noted in Chapter 2. The objective of the hospital has, nonetheless, widely been considered either implicitly, or explicitly, as some sort of welfare maximisation. However, this has too often been vaguely specified leading the authors to ignore the interactions between the different arguments in the objective function. It is widely recognised that the stochastic nature of demand is an essential part of modelling hospital behaviour. Less emphasis has been placed on the heterogeneity of demand, although some authors have highlighted important issues regarding bed substitutability, and particularly the useful distinction between emergency and elective cases. However, this issue has largely been ignored in the maximisation and modelling process. Trade-offs exist between emergency treatments and turnaways, elective treatments and queuing.

Hospitals face a range of demands for their services and, in the UK, the capacity to treat these demands is constrained, i.e. in aggregate there is an excess demand for hospital services. Furthermore, we know that demand for at least an element of hospital services is stochastic. These are the key features of the NHS hospital. Therefore, the hospital must decide how best to utilise existing capacity given the nature of demand. The question then becomes what is the optimal use of hospital capacity given demand heterogeneity and demand uncertainty? This chapter will build a theoretical model of hospital capacity utilisation based on standard theories of production, and will consider the impact of two main issues; output heterogeneity and demand uncertainty with respect to the situation facing NHS hospitals in the UK. Section 2 will consider this issue building on underlying economic theory using a standard analysis of production theory. Section 3 will consider the specification of a general objective function consistent with hospital behaviour, and develop a geometric representation of hospital capacity utilisation choices. Section 4 will bring together the arguments in the objective function, and develop a four segment representation of capacity allocation. Section 5 will formalise this model in a mathematical representation allowing the optimal allocation condition to be identified. Section 6 will draw some conclusions.

Section 2.1 A conceptual framework for considering capacity utilisation in hospitals

Analyses of hospital resource allocation decisions have concentrated largely on intuitive rather than theoretical analysis. Formal models have largely ignored theoretical foundations and the particular problems facing hospitals. Before considering the full theoretical model incorporating the objectives of the hospital, issues of uncertainty, and output heterogeneity, it is useful to outline the theoretical approach using a more constrained analysis. As such, we will initially focus on the impact that stochastic demand has on a hospital's production process from a private perspective. Later, we will consider the same issue but take a wider social perspective and consider social costs and benefits.

Before considering the model it is worth outlining the main assumptions we will make regarding the features of the hospital firm and the demand for hospital services.

i) The hospital as a multiproduct firm. Hospitals produce many different goods and services sharing a common means of production. However, for presentational purposes, hospital services will be separated into two outputs here; elective procedures and emergency procedures. Whilst this represents a considerable simplification it is nonetheless, both convenient and intuitively appealing, for example, reimbursement of these two types of service differ in the UK.

ii) The demand for services. We will assume that the demand for emergency services is stochastic and that there exists an excess demand for elective services.

iii) Fixed capacity. We will assume that there is a fixed capacity which is based on current endowments of capital and labour. We also assume that the hospital must decide, ex ante, what level of resources to allocate to each service before demand is revealed and that beds are perfectly substitutable between different outputs.

iv) The hospital utility function. Initially we will adopt a private perspective and assume that the hospital is a surplus maximiser. In section 3 we will develop a more sophisticated von Neumann-Morgenstern expected utility maximiser that takes account of social costs and benefits.

These simplifications allow us to consider the allocation issues within a standard economic framework, outlined in section 2.2, below.

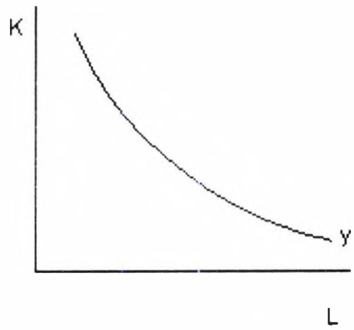
Section 2.2: Standard analysis of production theory

The starting point for standard analysis of the firm's production decision is the problem of minimising the cost of production of a given output, subject to available technology. The production function is a purely technological relationship representing the association between inputs and outputs. If we initially restrict the analysis to a firm with two inputs (K, L) producing a single output (y), then output can be represented as a function of the two inputs:

$$y = f(K, L) \tag{3.1}$$

where the factor inputs are combined to produce output which can be represented by a production isoquant reflecting all the technically efficient combinations of factors of production for a given level of output. This can be represented by an isoquant, such as Figure 3.1, which denotes the combination of capital (K) and labour (L) to produce a given level of output, y (where the smooth isoquant assumes substitutability between K and L over a certain range).

Figure 3.1: Production isoquant

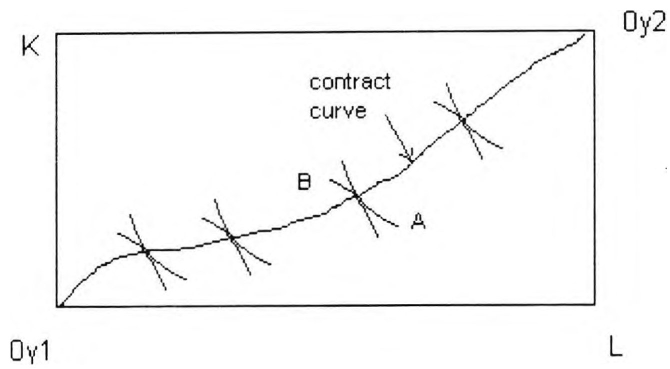


If we then extend the analysis to a multiproduct firm we can define the production function for two products, y_1 and y_2 where, as above, each product is assumed to be produced by two factors, K and L, such that:

$$y_1 = f(K_{y1}, L_{y1})$$
$$y_2 = g(K_{y2}, L_{y2}) \quad (3.2)$$

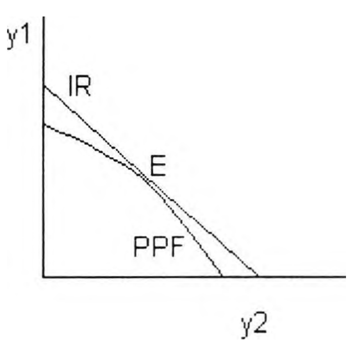
Each production function may now be represented by a set of isoquants. This allows us to represent the production possibility curve of the firm using an Edgeworth-Bowley box. Assuming that the firm has initial endowments of capital and labour, 0K and 0L respectively, the production function for y_1 is represented by isoquants A in Figure 3.2, and the production function for y_2 is represented by B. The points lying on the contract curve represent points of efficient production.

Figure 3.2: Edgeworth-Bowley box



To determine the optimal choice of outputs, y_1 and y_2, it is necessary to derive the production possibility frontier (PPF). The PPF represents the combination of outputs that can be produced by the firm using efficient combinations of inputs, and is derived directly from the contract curve in the Edgeworth-Bowley box. This is represented by PPF in Figure 3.3, below (which is usually represented as concave reflecting diminishing returns to scale and reflects the rate of transformation of y_1 into y_2).

Figure 3.3: Iso-revenue curve and PPF



The optimal combination of outputs is determined by the indifference curve of the firm, which represents the firm's utility function. Standard analysis represents this through the iso-revenue curve, which is the locus of points of various combinations of output that yield the same total revenue for the firm. For example, in Figure 3.3 where the slope of the iso-revenue curve (IR) is equal to the ratio of the price of the two outputs. (The iso-revenue curve as represented by IR in Figure 3.3 assumes that prices are constant over the entire output, i.e. marginal revenue = average revenue, which is consistent with perfect competition). The optimal combination of outputs, for a revenue maximising firm, is the one that yields the highest revenue, given available inputs. This is located at the point of tangency between the iso-revenue curve and the PPF, such as point E in Figure 3.3, where the firm attempts to attain the highest iso-revenue curve (i.e. furthest away from the origin).

Furthermore, the equilibrium level of output for a profit maximising multiproduct firm coincides with the revenue maximising level determined by the above analysis. The profit maximising level of output, given the constraints of factors of production, is where the marginal rate of product transformation between y_1 and y_2 (represented by the slope of the PPF,) and the ratio of the relative prices of the output are equal. At the point of tangency the slopes of the iso-revenue and product transformation curves are equal, such that:

$$-\frac{dy_1}{dy_2} = MRPT_{y_2,y_1} = \frac{p_{y_2}}{p_{y_1}} \tag{3.3}$$

where, $MRPT_{y2,y1}$ is the marginal rate of product transformation between the two outputs, $\partial y_1/\partial y_2$ is the slope of the PPF at the point of tangency, and p_{y1} and p_{y2} are the relative prices of y_1 and y_2 respectively.

Any movement along the PPF, away from this equilibrium condition, will reduce the total revenue received whilst leaving total costs of production unaltered.[1] Therefore, assuming that the quantity of factors and their prices are given, the maximisation of profit is achieved by maximising revenue. Standard economic analysis, thus, allows us to determine the optimal allocation of available resources for a two-output firm, where optimal output mix is determined by the production process and by the relative return to each product.

This approach can be extended to the hospital sector. If we consider the basic production processes set out in an Edgeworth-Bowley box in the context of fixed amounts of capital and labour, which are used to produce two hospital services; emergency and elective treatments, then an efficiency locus can be identified. This allows a production possibility frontier for the possible combinations of emergency and elective treatments to be identified, such as PPF in Figure 3.4.

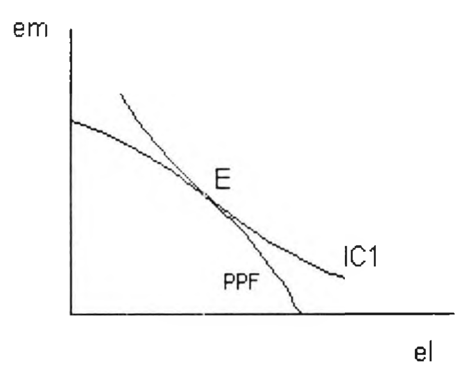
The optimal combination of outputs is determined by the hospital's objective function. The hospital decision maker's preferences over the combinations of the two illnesses can be represented by standard isoutility, or indifference curves. At this stage the

[1] See Koutsoyiannis (1979) for a proof.

objective function of the hospital need not be more sophisticated that specifying that the hospitals utility function is positive in both outputs. The optimal combination of outputs is identified at the point of tangency between the PPF and the highest iso-utility curve.

The optimal attainable combination of emergency and elective treatments is represented by E in Figure 3.4, at the point of tangency between the PPF and highest indifference curve IC1, (where em represents emergency cases treated and el represents elective cases treated and IC1 is convex to reflect diminishing marginal utility).

Figure 3.4: Optimal combination of hospital output



Therefore, hospital output can be represented using the framework outlined above, where the hospital employs existing endowments of capital and labour to produce combinations of elective and emergency treatments, where the indifference curve represents the preference function for the two outputs.

However, in order to incorporate some of the complexities of the production process, detailed earlier, it is necessary to introduce two complications; demand uncertainty and a more sophisticated specification of the hospital's objective function. First, let us consider the impact of demand uncertainty before considering a fuller specification of the objective function.

Section 2.3: Demand uncertainty

A theoretical analysis of production responses to demand uncertainty can be usefully initiated using the analysis of Aiginger (1985). He showed, without any reference to the firm's attitude towards risk or their preference function, that a producer, facing conditions of uncertainty arising from stochastic demand, would produce less than under conditions of certainty. The inference being that there is an inherent 'cost' associated with uncertainty. Aiginger distinguished between quantity produced and quantity sold in his model. He considered a one period single output model and assumed that unsold goods have no value; i.e. they are non-storable and cannot be backlogged, such that no value can be attached to unsatisfied demand. For simplicity, he assumed risk-neutrality as we will do here.

His model assumed that the firm is an expected profit-maximiser (although in principle any objective function can be applied), where profits result from the difference between costs and expected revenue; expected revenue being determined by the product price and the quantity produced, or demanded, whichever is smaller. This gives:

$$E\pi = p\min[q,x] - c(q) \qquad (3.4)$$

Where Eπ is expected profits, p is price, q is quantity produced, x is quantity demanded, and c(q) represents costs. He further assumed that p is given, which is reasonable if one believes that one reaction to demand uncertainty is to accept the previous period's price.

This can also be expressed in the following way:

$$E\pi = \int_0^q pxf(x)dx + \int_q^\infty pqf(x)dx - c(q) \qquad (3.5)$$

where the first term on the right hand side represents the expected level of quantity sold for any level of quantity produced, q, for expected demand, x, given the distribution of demand, f(x). The second term represents the expectation that for any chosen level of quantity produced the good will be sold (i.e. demand will be at least equal to q) and c(q) is defined as a function of quantity. Where q is determined ex ante, before demand is realised and, ex post, cannot be altered.

Differentiating profits with respect to chosen output, q, gives the optimal allocation condition, where we have to employ the Leibniz's formula to differentiate the second term since the parameter q features in both the integral function and as a limit, where the general Leibniz's formula for differentiating a function of the form:

$$F(t) = \int_{a(t)}^{b(t)} f(t,x)dx$$

where f(t,x), a(t) and b(t) are differentiable functions is:

$$F'(t) = f(t,b(t))b'(t) - f(t,a(t))a'(t) + \int_{a(t)}^{b(t)} \frac{\partial f(t,x)}{\partial t} dx$$

Thus, differentiating with respect to q gives the following, which is most clearly represented by separating the into three parts, where δEπ/δq is:

$$\frac{\partial \int_0^q p \cdot xf(x)dx}{\partial q} = p \cdot qf(q) \cdot 1 - p \cdot 0f(.0) \cdot 0$$

and, using Leibniz's rule:

$$\frac{\partial \int_q^\infty p \cdot qf(x)dx}{\partial q} = p \cdot q \cdot f(\infty).0 - p \cdot qf(q) \cdot 1 + \int_q^\infty pf(x)dx.(1)$$

and

$$\frac{\partial c(q)}{\partial q} = c'(q)$$

which is:

$$\frac{\partial E\pi}{\partial q} = \int_q^{\infty} p \cdot f(x)dx - c'(q) \tag{3.6}$$

where c′(q) represents the marginal costs of production.

This can be abbreviated to:

$$p \cdot [1 - F(q)] = c'(q) \tag{3.7}$$

or:

$$p \cdot = \frac{c'(q)}{[1 - F(q)]} \tag{3.7a}$$

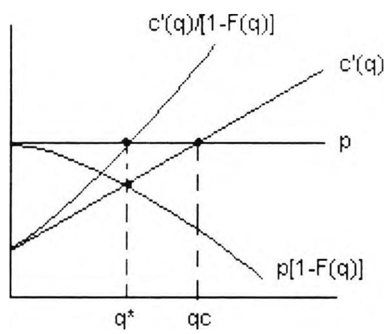
where [1-F(q)] is equal to the probability that output will be sold (and equals the value of the integral between q and infinity). Such that the optimal level of production is determined by the equality:

$$p = c'(q) + p \cdot F(q) \tag{3.8}$$

where F(q) represents the probability that q cannot be sold and gives an expression for the marginal costs of uncertainty. Since, under conditions of certainty, the profit maximising firm will produce up to the point where price equals marginal cost it is clear that uncertainty introduces an extra element of cost, i.e. p.F(q).

Therefore, the reduction in the output of the firm under conditions of uncertainty, as compared with conditions of certainty, is as a result of the extra cost component, which shifts the marginal cost curve upwards reducing the optimal output level. This can be represented using a simple diagram, such as Figure 3.5, where uncertainty can be introduced on the production or revenue side. In Figure 3.5 q_c represents output under conditions of certainty, and q* under conditions of uncertainty, where uncertainty shifts the marginal cost curve, (or marginal revenue curve), such that the optimal output under conditions of demand uncertainty is less than under conditions of certainty.

Figure 3.5: Output under demand uncertainty



Just to reiterate, the uncertainty is introduced as a production response to demand uncertainty and does not depend on the firm's risk preferences.

Once more, this general approach can be extended to the hospital sector. First, however, let us extend the analysis to consider a two-output firm (hospital), and consider the implications for the choice of output if uncertainty is present for one of the goods. Drawing on the above framework, let us assume that the hospital produces two goods; elective and emergency, and that uncertainty only affects the latter.

As earlier, the objective function of the hospital need not be any more sophisticated than specifying that the hospital's utility is positive in both goods. However, to make comparison with the earlier specification easier we will consider an expected surplus maximising hospital from a private perspective. We assume that price is determined exogenously for the two services, and that each service attracts a different price[2], and has a distinct cost associated with production

Let us set up the problem as above, where expected Eπ are given by:

$$E\pi = p_{el} \cdot q_{el} + p_{em} \cdot \int_0^{q_{em}} xf(x)dx + p_{em} \cdot \int_{q_{em}}^{\infty} q_{em} f(x)dx - c(q_{em}) - c(q_{el}) \qquad (3.9)$$

where p_{el} and p_{em} are the elective and emergency prices, respectively, (and $p_{em} > p_{el}$ for the hospital to allocate any beds to the emergency sector[3]), and q_{el} and q_{em} are the quantities of electives and emergency cases, respectively. And where the second term in equation 3.9 represents the expected level of emergency cases treated and the third term represents that probability that the last bed will be filled, given the chosen level of emergency capacity. Costs of production are represented by $c(q_{el})$ and $c(q_{em})$ for elective and emergency cases and we will assume, as above, that production levels are determined ex ante, before demand is realised, and are fixed ex post. The costs of

[2] This is consistent with, for example, prices being set via a bargaining process based on the previous period's average costs.
[3] If $p_{el} > p_{em}$ then this would lead to a corner solution where all beds where allocated to the elective sector.

production are, therefore, not represented as expectations, since the costs are related to output levels produced (rather than actually sold), whereas the revenue depends on the amount actually sold, which may of course be less than or equal to the amount produced. That is, a number of emergency beds may be staffed for emergency cases, but the demand for emergency beds may fall below this production level. This is picked up below when the capacity constraint is also introduced.

Differentiating with respect to q_{em} and q_{el} gives:

$$\frac{\partial E\pi}{\partial q_{el}} = p_{el} - c'(q_{el}) \tag{3.10a}$$

and:

$$\frac{\partial E\pi}{\partial q_{em}} = p_{em} - c'(q_{em}) + p_{em} \cdot F(q_{em}) \tag{3.10b}$$

Where, as earlier, the third term in equation 3.9 is differentiated using Leibniz's rule.

To consider the capacity issue we introduce a fixed capacity constraint determined by the number of hospital beds, B, such that the maximum number of emergency or elective cases treated is determined by the number of beds available and the length of stay. Let us further assume, for simplicity, that the length of stay is equal and normalised to one for both emergency and elective cases. This allows the number of beds to represent the number of cases treated for both outputs but is by no means necessary for the results to hold generally; rather it allows beds and cases to be used interchangeably in the subsequent expositions of the problem.

The capacity constraint introduced determines the maximum number of elective and emergency cases that can be treated such that:

$$B = q_{em} + q_{el} \tag{3.11}$$

Where B represents the total number of available hospital beds and q_{em} and q_{el} are the quantities of elective and emergency cases produced (i.e. rather than demanded). This allows the problem to be set up with reference to the constraint, B. It is apparent that once the capacity constraint is introduced .the output choice becomes a joint process. such that. if the level of emergency output is chosen this determines the remaining capacity to produce elective output, and vice versa. The capacity constraint reminds us that the production of emergency and elective cases depends on the allocation of capacity. and that capacity is allocated to either elective or emergency beds and this determines the level of each produced. While this split is based on output produced (e.g. an emergency or elective bed). it is once again the case that emergency demand (expressed as an expectation reflecting its stochastic nature) may fall below the level produced. It is the difference between realised demand and produced output that gives rise to the notion of excess (or reserve) capacity. This unused capacity is, nevertheless, an efficient response to demand uncertainty.

Expected surplus is now given by the equation:

$$E\pi = p_{el} \cdot q_{el} + p_{em} \cdot \int_0^{q_{em}} xf(x)dx + p_{em} \cdot \int_{q_{em}}^{\infty} q_{em} f(x)dx - c(q_{em}) - c(q_{el}) \quad (3.12)$$
$$s.t. B = q_{em} + q_{el}$$

This can be solved employing a Lagrangian multiplier. However, since B = q_{em} + q_{el} it is possible to substitute q_{el} (or q_{em}) for the expression (B - q_{el}), giving:

$$E\pi = p_{el} \cdot (B - q_{em}) + p_{em} \cdot \int_0^{q_{em}} xf(x)dx + p_{em} \cdot \int_{q_{em}}^{\infty} q_{em} f(x)dx - c(q_{em}) - c(B - q_{em})$$

(3.13)

This allows us to differentiate with respect to q_{em} in order to derive the first order condition which, on simplification, and once more employing Leibniz's rule, is:

$$-p_{el} + p_{em} \cdot \int_{q_{em}}^{\infty} q_{em} f(x)dx - c'(q_{em}) + c'(B - q_{em}) = 0 \quad (3.14a)$$

or:

$$p_{em}[1 - F(q_{em}] - c'(q_{em}) = p_{el} - c'(B - q_{em}) \quad (3.14b)$$

where $c'(q_{em})$ and $c'(q_{el})$ represent the marginal costs of producing emergency and elective output, respectively. The optimal allocation of capacity, under conditions of certainty, is achieved where the ratio of price, less marginal costs, are equal. Since return to the emergency output is uncertain, the impact of uncertainty is seen to be the reduction in quantity of emergency output relative to an environment of certainty, as the return to allocation of beds to the emergency sector is weighted by the probability of emergency demand being present, $[1-F(q_{em})]$, (which is < 1 for all q_{em}).

This can also be given as the following formally identical interpretations:

$$p_{em} = p_{el} - c'(B-q_{em}) + c'(q_{em}) + p_{em} F(q_{em}) \qquad (3.15)$$

$$[1-F(q_{em})] = p_{el} - c'(B-q_{em}) + c'(q_{el}) / p_{em} \qquad (3.16)$$

$$F(q_{em}) = [p_{em} - c'(q_{em}) / p_{em}] - [p_{el} - c'(B-q_{em}) / p_{em}] \qquad (3.17)$$

Where (3.15) replicates the Aiginger result for a two sector output model showing that there is a cost of uncertainty in that the marginal cost of an emergency bed is extended to include the lost marginal revenue related to the probability that the bed cannot be filled, $F(q_{em})$. Equations (3.16) and (3.17) show the probability that the last emergency bed should be needed and is equal to the ratio of marginal cost to price, where part of the marginal cost is the surplus that could have been earned in the elective sector, and the probability that the last emergency bed will not be filled, respectively. With the latter having to equal the mark-up of the price over marginal cost in the emergency sector and the markup, in terms of marginal surplus in the elective sector to the emergency price, with regards the elective sector. It is the value placed on these probabilities which reflect the hospital's decisions over the reservation capacity to hold in trading off the probability of unused capacity or turnaway against surplus. The setting of reservation capacity is, therefore, consistent with economic efficiency, even allowing for a cost of uncertainty as based on the lost marginal revenue attached to the possibility of having an unfilled emergency bed.

It is straightforward enough to reconcile this approach with the earlier representation. Recalling equation 3.3; the identity that specifies the efficient allocation of resources across the two outputs y_1 and y_2, we have:

$$-\frac{dy_1}{dy_2} = MRPT_{y_2,y_1} = \frac{p_{y_2}}{p_{y_1}}$$

Where the MRPT is equal to the slope of the PPF, and the ratio of P_{y2}/P_{y1} represents the slope of the iso-revenue curve. Since we have assumed that the lengths of stay are equal across the two treatments, the marginal rate of product transformation between emergency and elective care is equal to one[4].

From above we know that, taking into account uncertainty, the hospital will allocate

[4] Strictly, this requires the further simplification that the emergency and elective treatments also require the same labour input as well as the same capital input since the slope of the PPF, i.e. the MRPT is:

$$-\frac{\partial q_{em}}{\partial q_{el}} = \frac{MP_{L,q_{em}}}{MP_{L,q_{el}}} = \frac{MP_{K,q_{em}}}{MPK_{L,q_{el}}}$$

capacity to the emergency (elective) sector up to the point where:

$$p_{em}[1-F(q_{em}]-c'(q_{em})=p_{el}-c'(K-q_{em}) \qquad (3.18a)$$

or:

$$\frac{p_{em}\cdot[1-F(q_{em})]-c'(q_{em})}{p_{el}-c'(K-q_{em})}=1 \qquad (3.18b)$$

Since we have assumed that the marginal costs are equal (i.e. MRPT = 1) then profit maximisation is achieved by maximising revenue such that:

$$MRPT_{el,em}=\frac{p_{em}\cdot[1-F(q_{em})]}{p_{el}}=1 \qquad (3.19a)$$

or:

$$MRPT_{el,em}=\frac{p_{em}}{p_{el}}=\frac{1}{[1-F(q_{em})]} \qquad (3.19b)$$

This can be more generally represented as:

$$MRPT_{el,em}=M=\frac{C}{[1-F(q_{em})]} \qquad (3.19c)$$

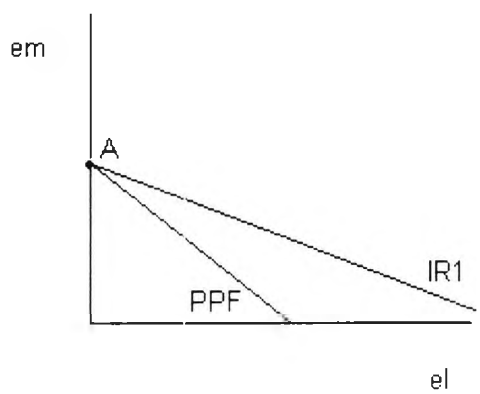
where, C is the rate of product transformation and M is the price ratio.

That is, the optimal allocation is at the point of tangency between the MRPT and the price ratio, however, the MRPT is determined by the allocation of capacity to the emergency sector (i.e. as more capacity is allocated to emergencies the probability of demand being present, [1-F(q_{em})], falls). Therefore, uncertainty alters the MRPT and, thus, the optimal allocation.

For exposition purposes, let us consider the impact that demand uncertainty has on the determination of efficient capacity utilisation, which is the capacity definition that embodies reserve capacity, through the use of a geometric representation of the problem. In this way we can concentrate on the manner in which demand uncertainty is consistent with reservation capacity.

This can be considered through the use of production possibility frontier and indifference curve analysis. If we assume, as above, that the length of stay is equal across emergency and elective treatments then the production possibility frontier, which represent the transformation of between emergency and elective cases as presented earlier, becomes a straight line. For consistency, let us assume that the hospital is an expected surplus maximiser where the indifference curves are represented by iso-revenue curves (such as IR1 in figure 3.6), which show the combination of emergency and elective cases that give rise to the same level of revenue. Let us further assume, purely for expositional purposes, that the price received for treating emergency cases is twice that for elective cases and that prices are constant across the whole output. This gives the situation presented in Figure 3.6 where IR1 is the iso-revenue curve, and PPF is, once more, the production possibility frontier. In these circumstances and under conditions of certainty the hospital would choose a point such as A, where all available capacity is given to emergency treatment.

Figure 3.6: surplus maximising hospital under conditions of certainty

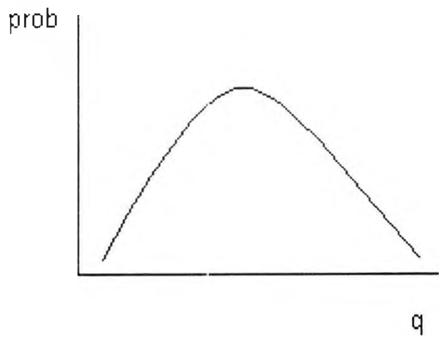


Given that the hospital receives twice the revenue for treating emergency cases as for electives, and that the marginal rate of transformation is one, it should not be surprising that a surplus maximising hospital would allocate all capacity to treating emergency cases. If, however, we introduce uncertainty then the situation changes.

In order to introduce uncertainty we will assume, as earlier, that the demand for emergency care is randomly distributed with a known probability density function. Further, we will assume that the hospital must make an ex ante decision regarding capacity utilisation based on the expectations of the level of emergency demand, and recognising that, ex post, these expectations may not be realised. In order to produce at any level of output the hospital must commit resources. Uncertainty results in the hospital being faced with a situation whereby there will be a probability of unused capacity (or a probability of insufficient capacity, but let us concentrate on the former). In the event of ex post unused capacity standard analysis, i.e. not taking account of the production response to demand uncertainty, would suggest that the hospital is not operating efficiently. That is, the firm will be operating within its production possibility frontier and not on its efficiency locus. This result can be considered rational, however, once uncertainty is introduced, explaining the existence of reserve capacity in the hospital sector.

Recognition of the production reaction to demand uncertainty in an analogous manner to the result above, can be shown by constructing a 'shadow PPF', which lies within the PPF facing the hospital under conditions of certainty. Let us, for the sake of argument, assume that emergency demand is randomly and normally distributed such that we have a bell-shaped distribution as in Figure 3.7. As more provision is made for emergency services the probability of emergency cases arriving falls. That is, as we move towards the right hand tail of the distribution in Figure 3.7 the probability of an additional emergency arriving falls from one and tends towards zero. (That is, in equation 3.12, as q_{em} tends towards ∞ the value of the third term, i.e. the integral between q_{em} and ∞, tends towards zero. Therefore, the expected marginal increase in the level of emergency output falls as q_{em} increases).

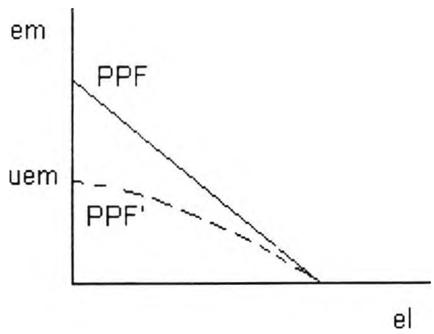
Figure 3.7: Normal distribution



For each allocation of capacity and, therefore, each point on the PPF, there is a probability of emergency demand being present, and this probability declines as the hospital allocates more capacity to treating emergency cases and moves along the x-axis away from the origin in Figure 3.7.

The hospital must choose, ex ante, how much capacity to allocate to emergency and elective services based on its expectation of emergency demand. The expectation of emergency demand is determined by the probability distribution and, when the hospital makes its choice of allocation, the expected number of emergency cases treated is determined by the expected mean to the left of the truncated distribution (which is the represented by terms 2 and 3 in equation (3.12)). Each allocation to the emergency sector represents a point on the distribution and, therefore, it is possible to calculate the expected number of emergency cases treated for every possible allocation. This allows us to construct a shadow PPF such as PPF′ in figure 3.8, below.

Figure 3.8: The shadow PPF

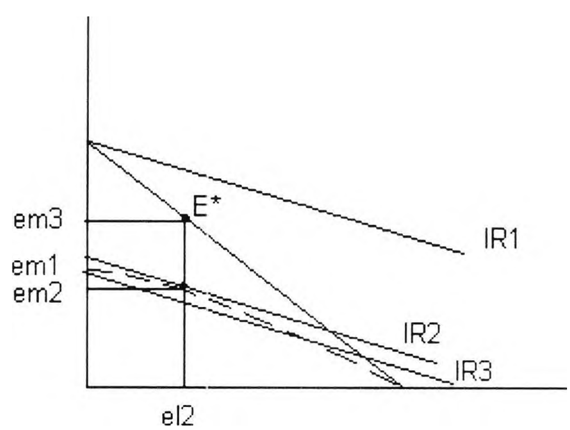


The shadow PPF (PPF′ in Figure 3.8) reflects the assumptions made above, where as more emergency services are catered for the probability of filling all the beds falls, consequently the slope of PPF′ becomes more shallow as emergency capacity moves away from the origin and as the probability density function tends towards zero. (The exact shape depends on the actual probability density function (p.d.f.) of emergency arrivals). We do know, however, that if the hospital had the capacity to treat all possible emergency arrivals, i.e. the full range of the distribution, that the expected number of cases treated would be equal to the mean of the normal (i.e. untruncated) distribution. Therefore, this sets the upper limit of PPF′ at u_{em} in Figure 3.8. Using this framework, it is now possible to consider the impact this has on the allocation decision of the hospital.

Let us reconsider Figure 3.6, where the hospital, under conditions of certainty where the price of emergency care was twice that of elective care, chose a corner solution such that it allocated all available capacity to treating emergency cases, attaining IR1.

However, when uncertainty is introduced, if the hospital were to allocate at this point it would expect to attain IR3 (on Figure 3.9) (where the IR curves represent 'certain equivalent' levels for emergency services), producing at [em1, 0]. If the hospital takes account of uncertainty, however, they would allocate capacity to achieve the highest *expected* utility; that is on IR2 [em2, el2]. To achieve this point, however, the hospital must allocate capacity to produce at E* [em3, el2], even though it only expects to produce at em3. Therefore, if the hospital takes account of uncertainty, it will allocate capacity in order to achieve the highest expected iso-revenue curve.

Figure 3.9: Optimal allocation under conditions of uncertainty using the shadow PPF



This produces a result where, as earlier, under conditions of uncertainty, the hospital produces less than under conditions of certainty. Furthermore, the hospital allocates capacity to generate reservation capacity as a reaction to demand uncertainty. Simply stated this shows it is the allocation of capacity that forms one of the cost components resulting from production responses to demand uncertainty. The precise cost of this response will depend on the marginal costs of production associated with this reservation capacity as shown in equation 3.17, which relates the probability of unused capacity to the markups, defined in terms of emergency sector revenue, achievable in the emergency and elective sectors.

Furthermore, this is provides a geometric representation of the optimal allocation (under conditions of uncertainty) identified earlier, where the MRPT under conditions of certainty equals one, and the price ratio equals two (or, more generally, that the MRPT equals C and the price ration equals M, where M > C). The hospital will, taking into account uncertainty, adjust the ex ante MRPT and allocate capacity based on expectations of emergency demand. The optimal allocation occurs where the price ratio of emergencies to electives equals the MRPT divided by the probability of emergency demand being present. The latter determines the slope of the shadow PPF, which is determined by the p.d.f. of emergency demand.

The specification of the problem outlined above has, however, employed a standard analysis of the problem in so much as it assumes that the aim of the hospital (firm) is surplus maximisation from a private perspective. In the next section we will consider a more general specification of the objective function of the hospital from a social perspective.

Section 3.1: Specification of a hospital objective function

The previous section considered two issues of importance when formally modelling hospital behaviour; output heterogeneity and demand uncertainty. However, the model specified above considered a surplus maximising hospital which only considered private costs. At this point it is worth re-considering the objective of the non-profit hospital by opening up the discussion to incorporate a social perspective.

The NHS is widely characterised by a capacity constrained system i.e. bed supply is limited, such that there exists excess demand for care. The situation is more complex, however, due to the existence of heterogeneous demand and demand uncertainty, such that there are conflicting demands for the available capacity and some of these demands are uncertain. In section 2, above, we have, as others have done, separated demand into two types of care; elective and emergency. In the UK demand for elective care is characterised by the existence of waiting lists, reflecting the excess demand for care in this sector. However, hospitals quite regularly operate at less than full capacity, which as we have shown, can be viewed as a rational response to demand uncertainty. However, here we want to move away from a model of behaviour based on assumptions of private surplus maximisation. It is clear from observing hospitals in the UK that there are times when the hospital will operate at full capacity and times when unused capacity exists. The question this raises is why, if demand for elective care is greater than capacity, do hospitals operate at less than full capacity? The answer to this question provides the general specification of the hospital objective function.

There are three plausible explanations why hospitals would keep reserve capacity, each dependent on the objective function. The first explanation can be illustrated by employing a simple quantity maximand. If the aim of the hospital was to maximise output, then the hospital may keep reserve capacity if the length of stay for emergency cases was less than the length of stay for electives. The hospital would maximise output by allocating capacity to treat emergency cases as long as the expected length of time the bed was empty was less than the difference in the length of stay between electives and emergency cases. The main problem with this explanation is that empirical evidence suggests that the length of stay of elective cases is less than the length of stay of emergency cases.

Therefore, a more sophisticated maximand must be employed. The second explanation why hospitals might keep reserve capacity relies on a weighted output maximand. If emergency cases spend longer in hospital than elective cases then one reason that hospital may keep reserve capacity to treat emergency cases might be that they attach greater weight to treating them. (It is immaterial where this weighting comes from at this stage; it could be physician, manager or consumer weighting). This being the case, hospitals would allocate bed capacity to emergency cases, even if the length of stay was longer than for elective cases and demand was uncertain. They would allocate beds to emergencies as long as the weight attached to output, adjusted for length of stay, and the expectation of demand arriving, was greater than that weight attached to treating elective cases, adjusted by length of stay. That is, as long as the expected return per day was greater from allocating beds to emergency care than elective care. The question this raises is why should emergency output receive greater weight than elective output; is there any inherent difference between the two such that the hospital would value one more than the other? The intuitive answer is yes. Emergency cases require treatment immediately or the consequences for the individual may be serious. This, however, makes no reference to the inherent benefit or returns to the hospital from treating emergency cases compared to elective cases. Rather it relies on the hospitals suffering in some way if emergency patients are turned away. Therefore, the explanation for hospitals keeping reserve capacity may have to extend beyond the private returns to the hospital.

The most plausible reason for hospitals keeping reserve capacity to treat emergency cases when there exists, simultaneously, an excess demand for elective care, is that they do not want to turn emergency patients away. That is, the hospital attaches some sort of weight to not turning away emergency patients, or, put another way, the hospital attaches some sort of cost to turning an emergency patient away. The source of this cost may be financial (e.g. penalties for turning patients away or loss of future business), however, there may also exist a social cost where financial penalties do not exist, which may be related to the cost imposed on the patient who does not receive treatment, or the adverse political attention received when such an event occurs. Given the non-profit nature of NHS hospitals, it is quite plausible that this may indeed be the source of such 'cost'.

Therefore, in specifying the objective function of the NHS hospital we should take account of at least two factors which might explain the existence of reserve capacity; the weight or utility attached to treatment, and the cost of turning patients away.

There are, however, two other factors that should be taken into account. If we assume that the hospital's objectives consider factors beyond the purely private costs and benefits, then there may also be a social cost attached to leaving patients on the waiting list. On the basis of political attention waiting lists clearly are important. Therefore, we may also want to specify an objective function that not only attaches a social cost or disutility to turning emergency patients away, but also attaches a similar cost to waiting lists for elective treatment.

Therefore, the complete objective function should consider six factors: the utility attached to treating elective and emergency cases: the disutility attached to turning away emergency cases and queuing elective cases: and the costs of treating elective and emergency cases. Therefore, the most general specification of the hospital objective function can be made with reference to these six factors such that the hospital utility function, U, is:

$$U = f(q_{em}, q_{el}, em_t, w_{el}, c_{em}, c_{el})$$

Where q_{em} and q_{el} are emergency and elective outputs, respectively, em_t is emergencies turned away and w_{el} is elective waiting, and c_{em} and c_{el} are the costs of treating emergency and elective cases, respectively.

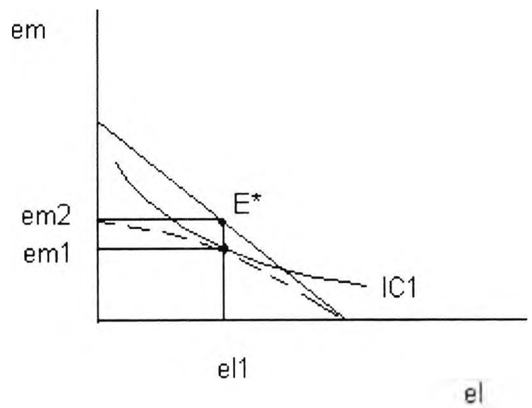
Let us, therefore, consider capacity utilitisation for an expected net social welfare maximising hospital with demand uncertainty for emergency care and excess demand for elective care and fixed bed capacity.

In the next section, we apply the production possibility/indifference curve approach developed above to consider utility and disutility conceptually, and attempt to draw together these two elements into a single geometric representation of the allocation problem. We will then proceed to formulate a mathematical model to derive the first order condition for optimal resource allocation.

Section 3.2: Iso-utility curves

The utility function of the hospital can be represented by standard indifference curves with elective and emergency cases on the two axes, such that the hospital decision-maker is indifferent between the combinations of elective and emergency cases on the same indifference curve. Introducing the PPF enables us to determine the optimal allocation point which, as before, is at the point of tangency between the PPF and the highest indifference curve. Introducing uncertainty for emergency cases, as earlier, gives us the situation in Figure 3.10, where an expected utility maximising hospital allocates capacity at E* [em2, el1] to achieve the ex ante maximum indifference curve, IC1, based on the expected demand for emergency services.

Figure 3.10: Optimal capacity allocation under conditions of demand uncertainty



The iso-utility curves are concave suggesting that the marginal utilities attached to treating emergency and elective cases decline with quantity (i.e. diminishing marginal utility sets in). This is consistent with general economic theory.

Section 3.3.1: Iso-disutility curves

In addition to the utility derived from the provision of elective and emergency services we also assume that there is disutility associated with every choice of elective and emergency service provision, based on the size of the waiting list and the number of emergency cases turned away, as a result of the choice of capacity given to each sector.

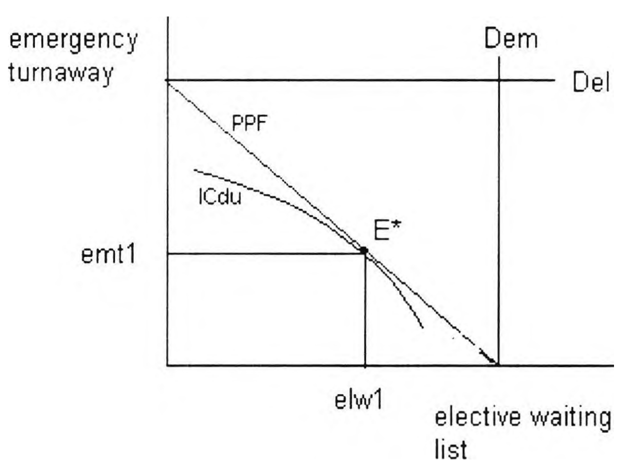
Consequently, it is possible to represent this situation in a similar framework, where instead of indifference curves, which represent iso-utility points, we can represent points of iso-disutility, where the production possibility frontier represents the feasible combinations of elective waiting list size and number of emergencies turned away. The aim of a disutility minimising hospital would be to operate on the lowest possible indifference curve, which is the point of tangency between the PPF and the lowest iso-disutility curve (as the hospital moves towards the origin the disutility falls).

Since the situation is less familiar than consideration of utility maximisation, it is worth considering the situation under conditions of certainty before introducing demand uncertainty into the problem.

The slope of the disutility curve in Figure 3.11 represents the marginal rate of substitution between emergency turnaway and elective waiting list. The shallower the slope of the curve, the more disutility the hospital derives from turning emergency cases away relative to placing elective patients on the waiting list.

In order to determine the size of the waiting list and number of patients turned away it is necessary to impose needs (or demand) constraints. The limits of the PPF are determined by the capacity constraints for treating emergencies and electives. The level of demand, in conjunction with the capacity constraints, will determine the number of patients on the waiting list and the number of patients turned away. (This will also determine whether the PPF intersects the x and y axes – in Figure 3.11, for example, we have assumed that there is sufficient capacity to treat either all the emergencies or all the electives. If capacity was insufficient to treat either of these then the PPF for disutility would not intersect either of the axes, i.e. the hospital could neither produce a zero waiting list nor zero numbers turned away). As the PPF moves towards the left hand corner of Figure 3.11 the closer the hospital is to being able to treat all demand (at the origin there is no capacity constraint, the hospital can treat all demand and the problem becomes uninteresting).

Figure 3.11: Disutility and the PPF



In Figure 3.11 the optimal allocation to minimise disutility is at E* [elw1,emt1] where

PPF and ICdu are at a point of tangency. The disutility PPF can be thought of as emanating from the top right hand corner of Figure 3.11, where the two demand curve intersect. At this point waiting list and turnaway are at their maximum, the hospital is producing zero output, and the turnaway and waiting lists are determined by size of total demand.

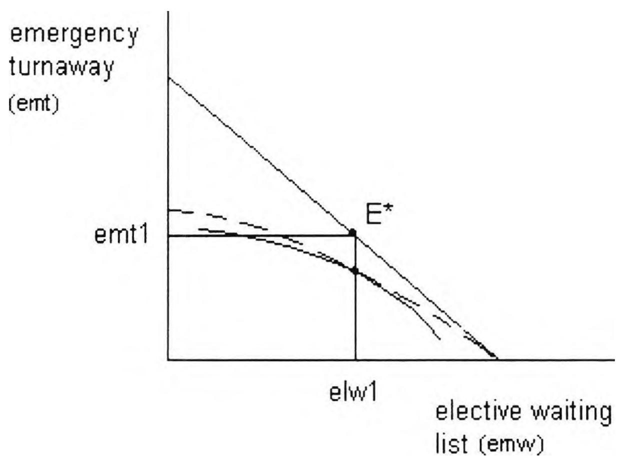
The disutility curves are convex if we assume that the marginal disutility associated with waiting list size and turnaway rate increases with quantity. This may make intuitive sense, as hospitals reduce the numbers of emergency cases turned away, they may become more concerned with the size of the elective waiting list and vice versa. However, this is an empirical question since we do not have general economic theory to guide us here (this issue will be considered in more detail in Chapter 5).

Section 3.3.2: Disutility and demand uncertainty

The optimal allocation of capacity under conditions of demand uncertainty can be determined for disutility in the same way it was for utility. Once more it is possible to construct a 'shadow PPF'. There will, in a similar way to utility, exist a shadow PPF that represents ex ante expectations, where the shadow PPF will have the same slope as for utility since it will be dependent on the p.d.f. of emergency demand.

The situation is represented below in Figure 3.12, where as more capacity is allocated to reducing emergency turnaway the lower the probability of patients actually arriving and being turned away becomes. This is analogous to the case for emergency treatments, however, the expected number of emergency cases turned away is equal to the expected mean to the right of the truncated distribution (whereas the expected number treated is equal to the expected mean to the left hand side of the truncated distribution).

Figure 3.12: Disutility and demand uncertainty



By taking account of uncertainty the hospital, ex ante, will attempt to attain the lowest possible disutility curve at E* [emt1, elw1]. Therefore, under conditions of demand uncertainty the hospital will be able to operate <u>outside</u> its PPF, i.e. at a lower level of disutility than implied by the PPF with certainty. This is analogous to the determination of utility under similar circumstances.

In the next section, we will bring together the four arguments in the objective function in a single representation of the problem, in order to attempt to identify a single optimal allocation point taking into account both utility and disutility.

Section 4.1: A four segment geometric model of hospital capacity allocation

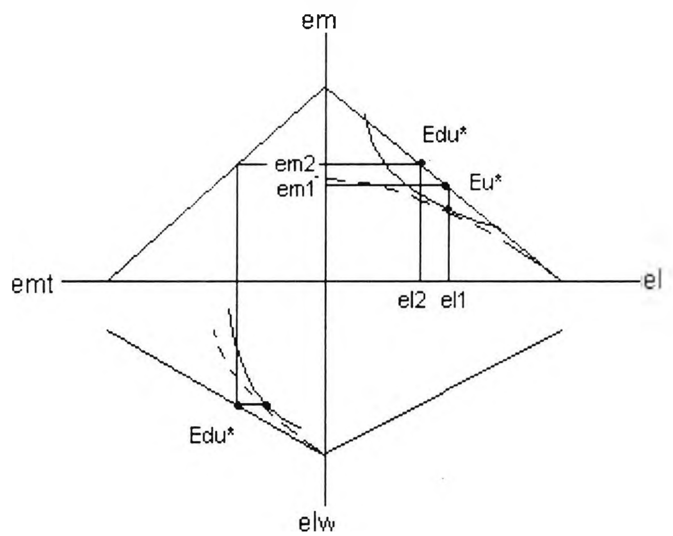
From the exposition of the problem above it is clear that the production of utility and disutility represents a joint production process, where the choice of emergency and elective cases treated determines the number of emergency cases that will be turned away, and the number of elective cases that will be queued. Furthermore, under a fixed capacity constraint, the production decision in each sector is also jointly determined, i.e. when the hospital decides on the capacity to allocate to treating emergency cases, this also determines the capacity available to treat elective cases, and consequently the size of the waiting list and number of emergency case turned away.

Consequently, the hospitals decision process can be represented on a single, four segment diagram incorporating all four arguments in the objective function (as in Figure 3.13). The main objective of this, is to bring together all the factors influencing the hospital's decision-making process, to attempt to identify a single optimal allocation point.

In Figure 3.13 below utility is represented in the top right hand segment of the four sector diagram and the bottom left hand segment represents the production of disutility.

It is now possible to identify the optimal allocation points for utility and disutility, independently.

Figure 3.13: Joint production process



Eu* represents the optimum point of production if hospitals consider only utility, where the hospital produces [em1, el1] and Edu* is the optimal point if they only consider disutility, producing at [em2, el2].

Section 4.2: Identifying a unique optimal allocation point

The optimal allocation points identified in the four segment diagram are considered in isolation. In order to identify a unique optimal allocation point it is necessary to bring the two segments together, as they represent a joint production process.

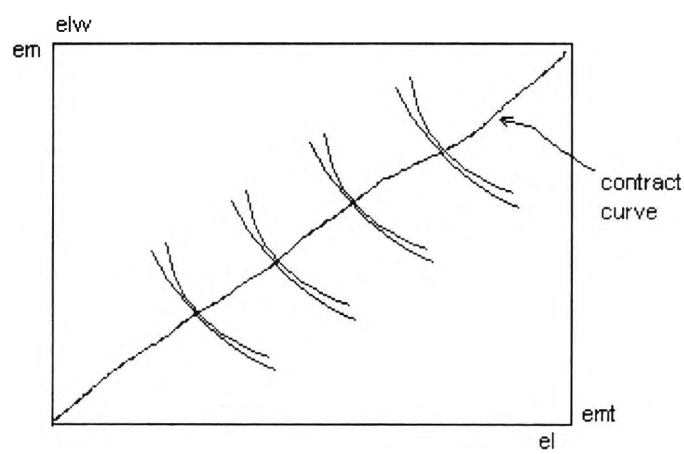
Bringing the two segments together allows us to consider the problem using a two-dimensional diagram, with elective services and emergency services on each of the axes, as originally presented. However, in this presentation of the problem the disutility associated with the indifference curves declines as we move away from the origin, (since more services are being provided, therefore, the turnaway and waiting lists are lower).

The result of incorporating the disutility curves in the analysis is that the hospital may move away from the original optimal allocation point, based solely on utility. The direction in which the hospital moves depends on the shape of the iso-disutility curves. If, for example, the disutility indifference curves are steep, suggesting that there is greater disutility attached to turning emergencies away than putting electives on waiting lists, then the hospital will shift its allocation upwards to the left (if the slope is shallow then the hospital may move downwards to the right). It is, of course, possible that the optimal allocation implied by the two arguments lead to the same allocation point.

The explanation of this result is that any move away from the optimal allocation point on the original curve will reduce utility, but if the allocation moves downwards to the right then this will also increase disutility. If, however, the hospital moves upwards to the left then disutility is reduced as well, therefore, there is a trade-off between a simultaneous reduction in utility and disutility.

The single optimal allocation point is identified by the point of tangency between the indifference curves, (this is similar to optima identified in an Edgeworth-Bowley box), where the hospital is simultaneously operating on its highest utility curve and its lowest disutility curve.

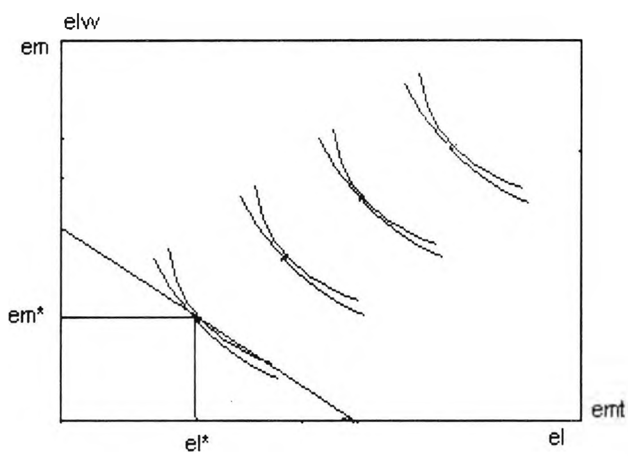
Figure 3.14: Efficiency locus for iso-utility/iso-disutility



The contract curve identified in Figure 3.14 represents the efficient combinations of output, since any points off this locus can be improved on, as with the Edgeworth-Bowley box, since it is possible to stay on the same utility (disutility) curve but operate on a lower disutility (higher utility) curve.

We can now superimpose the PPF on this locus to identify the unique optimal allocation point given available capacity as in Figure 3.15, below.

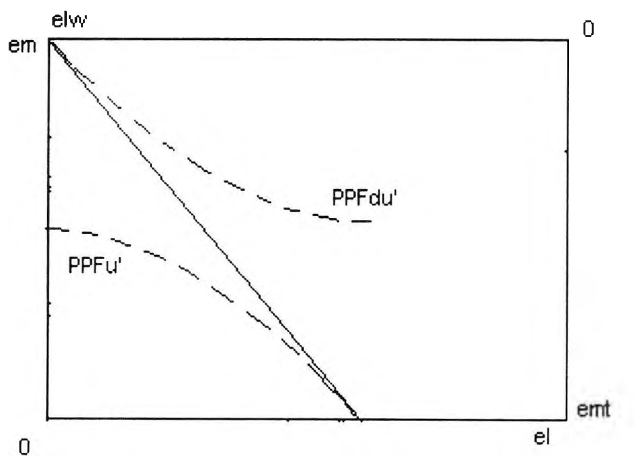
Figure 3.15: Optimal allocation given PPF



However, this identifies the optimal allocation under conditions of certainty. We need to introduce the shadow PPFs in order to identify the optimal allocation under uncertainty. This is not straightforward, since we know that for any given allocation of emergency capacity, the expected number of emergencies treated, and turned away, will be less than the under conditions of certainty. Consequently imposing the two shadow PPFs on a single diagram leads to the situation in Figure 3.16.

First, let us set out the problem where the hospital has capacity to treat all emergency cases (i.e. to cover the right hand tail of the distribution, such that it can choose the number turned away to be zero). This sets the limit of the shadow PPF at the expected mean of the distribution. Similarly, it is clear that the expected number of patients turned away at the limit, when emergency capacity equals zero, will be equal to this expected mean. This gives us a better intuitive feel for the situation.

Figure 3.16: Optimal allocation under uncertainty



The problem now is that the hospital may not operate at a point of tangency between the two isoutility curves, since there are effectively two production possibility frontiers. It is possible that an optimal allocation point cannot be identified. If this were the case, all we can say is that the hospital will attempt to operate at a point where, ex ante, the difference between the two isoquants is at a minimum, since at this point they will simultaneously be operating on the highest utility curve and lowest disutility curve. That is, where the sum of the social benefits less the sum of the social costs is greatest.

To formally identify the optimal solution we consider a more formal mathematical specification.

Section 5: Mathematical model of resource allocation decisions

The mathematical model will follow the same specification of the problem as above; output is split into emergency and elective care. The total number of beds represents the fixed capacity constraint. The decision variable is assumed to be the number of beds allocated to each sector. We assume that a hospital bed is a homogeneous unit and can be allocated to either sector, but once allocated to a sector they cannot be re-allocated in time period under consideration. Output is the number of cases treated, where the total number of cases treated per bed, per day, is equal to the number of beds allocated to the sector divided by length of stay.

Demand, or arrivals, are assumed to be random and exogenously determined in the emergency sector, according to an, as yet, undefined process. When the emergency beds are full the additional arrivals are assumed to be 'turned away' and lost to the system. There is assumed to be excess demand in the elective sector, such that demand is always greater than capacity, B, for all B. Therefore, the queue is never empty (and the choice for hospital is between treating an expected emergency arrival or taking one extra patient off the waiting list). Since we assume there is excess demand in the elective sector, such that demand is always greater than the available capacity to treat cases for all levels of hospital capacity, then the hospital knows that each bed allocated to the elective sector will be filled with certainty and that can choose the number of elective cases they treat. Furthermore, we assume that for any given decision period the total size of elective demand is known.

The objective function of the hospital is assumed to be utility maximisation, where utility is derived from treatments provided in both sectors, and disutility is associated with turning emergency patients away and placing, (or leaving), elective patients on a waiting list. The most general from of the utility function can be specified as:

$$U = U[q_{el}] - DU[el_w] + U[q_{em}] - DU[em_t] \qquad (3.17)$$

q_{el} = number of cases treated in the elective sector

el_w = size of waiting list in the elective sector

q_{em} = number of emergency cases treated

em_t = number of emergency cases turned away

In the emergency sector, however, we assume that patients arrive stochastically. For any chosen bed allocation level in the emergency sector there will be a probability that the all the beds are not filled, which, as outlined earlier, is determined by the probability distribution around the demand function. Conversely, there will be a probability that the number of arrivals will exceed capacity, if this happens patients are turned away. The expected numbers of patients treated and turned away are both determined by the number of beds allocated to the emergency sector.

The arguments in equation (3.17), therefore, become:

$q_{el} = B_{el}/LOS_{el}$

$el_w = f(B_{el}/LOS_{el})$

$q_{em} = f(B_{em}/LOS_{em}, f(x))$

$em_t = f(B_{em}/LOS_{em}, f(x)$

where,

$f(x)$ = the probability distribution around emergency demand

B_{em} = number of beds allocated to the emergency sector

B_{el} = number of beds allocated to the elective sector

B_T = total number of beds available

Where we know that,

$$B_{em} + B_{el} = B_T$$

where, for simplicity, let us also, as earlier, assume that the length of stay for electives and emergencies are equal to one; this provides a normalisation and allows us use beds and cases interchangeably but is by no means necessary. Furthermore, it allows us to assume that the marginal rate of transformation between emergency and elective cases is equal to one such that the costs of treating emergency and elective cases are equal as in the PPF/IC representation outlined earlier[5].

Furthermore let W represent the total number of elective patients waiting for treatment, such that:

$$el_w = (W - B_{el}) \qquad (3.18)$$

This allows us to set up the problem where the hospital aims to maximise expected utility, EU, where:

$$EU = U[B_{el}] - U[W-B_{el}] + EU[B_{em}, f(x)] - EDU[B_{em}, f(x)]$$

$$\text{s.t. } B_{em} + B_{el} = B_T \qquad (3.19)$$

This can be solved using a Lagrangian, but, as earlier, since $B_T = B_{em} + B_{el}$ we can substituting for B_{el} (or B_{em}) such that:

$$EU = U[B_T-B_{em}] - U[W-(B_T-B_{el}] + EU[B_{em}, f(x)] - EDU[B_{em}, f(x)] \qquad (3.20)$$

[5] Including marginal costs in the specification is trivial if we assume that output, and hence cost, is determined ex ante and fixed ex post, since, as earlier, the surplus maximising optimal condition coincides with the utility maximising solution. That is, we can simply add marginal costs to each side of the first order condition.

(3.20)

This tells us that expected utility is determined by the choice of beds allocated to treat emergencies (electives) given the total bed constraint.

First, let us consider the number of case treated in the emergency sector. Let us, as above, assume an unspecified distribution function f(x), where x is the number of arrivals (or demand) in the emergency sector and f(x) is the probability density function for any x. We know that once the hospital chooses the number of beds, B_{em}, allocated to the emergency sector, this determines the number of cases that can be treated. Let us consider the general representation of the situation where we will assume a non-linear utility function, and, as earlier, we will assume risk neutrality. The expected level of utility associated with any bed level, B_{em} can thus be represented as:

$$EU = \int_0^{B_{em}} U[x]f(x)dx + \int_{B_{em}}^{\infty} U[B_{em}]f(x)dx \qquad (3.21)$$

where the first term represents the expected level of emergency output, given the truncation at B_{em}, and the second the probability that the last bed will be filled. Similarly the number of emergency cases turned away by the hospital is also be determined by B_{em}, where the expected number of cases turned away is determined by the probability distribution around x, such that:

$$EDU = DU[\int_{B_{em}}^{\infty} \frac{(x).f(x)dx}{(1-F(x))} - B_{em}][1-F(x)] \qquad (3.22a)$$

which represents the expected number of cases greater than B_{em}. This can be simplified to:

$$EDU = \int_{B_{em}}^{\infty} DU[(x - B_{em})].f(x)dx \tag{3.22b}$$

The fully specified model becomes:

$$EU = U[(B_T - B_{em})] - DU[(W - (B_T - B_{em})] + \int_0^{B_{em}} U[x]f(x)dx + \int_{B_{em}}^{\infty} U[B_{em}]f(x)dx$$
$$- \int_{B_{em}}^{\infty} DU[(x - B_{em})]f(x)dx \tag{3.23}$$

Differentiating with respect to the choice variable, B_{em}, and once more using Leibniz's rule for terms 4 and 5 in equation 3.23, gives the optimal allocation (or general first order condition), which, in full, is:

$$-U_{el}'[B_T - B_{em}] - DU_w'[W - (B_T - B_{em})] + U_{em}[B_{em}]f(B_{em})(1) - U_{em}[\infty]f(\infty)(0)$$
$$-U_{em}[B_{em}]f(B_{em})(1) + \int_{B_{em}}^{\infty} U'_{em}[B_{em}]f(x).dx - DU[\infty - B_{em}](0) - DU[B_{em} - B_{em}](1)$$
$$+ \int_{B_{em}}^{\infty} DU_t'[(x - B_{em})].f(x).dx = 0 \tag{3.24}$$

where U'_{em} and U'_{el} represent the marginal utilities of emergency and elective cases, respectively and DU'_{em} and DU'_{el} are the marginal disutilities of emergency and elective cases, respectively.

Which is:

$$-U_{el}'[B_T - B_{em}] - DU_w'[W - (B_T - B_{em})] + \int_{B_{em}}^{\infty} U_{em}'[B_{em}]f(x).dx$$
$$+ \int_{B_{em}}^{\infty} DU_t'[x - B_{em}].f(x).dx = 0$$

(3.25)

or:

$$U_{el}'[B_T - B_{em}] - \int_{B_{em}}^{\infty} DU_t'[(x - B_{em})].f(x).dx = \int_{B_{em}}^{\infty} U_{em}'[B_{em}]f(x).dx - DU_w'[W - (B_T - B_{em})]$$

(3.26)

The second order condition of which is:

$$U_{el}''[B_T - B_{em}] + DU_w''[W - (B_T - B_{em})] - U_{em}'[B_{em}]f(B_{em}) + \int_{B_{em}}^{\infty} U_{em}''[B_{em}]f(x).dx$$
$$- \int_{B_{em}}^{\infty} DU_t''[x - B_{em}].f(x).dx = 0$$

where ´´ represents the second derivative. The sign of which is ambiguous and depends on the relative magnitude of each component of the solution.

Equation 3.24 can be more simply represented as:

$$MU_{em}.[1\text{-}F(B_{em})] - MDU_w = MU_{el} - MDU_t[1\text{-}F(B_{em})] \qquad (3.27)$$

where

MU_{em} = marginal utility of treating an emergency case (as a function of the number of cases treated)

MDU_w = marginal utility of treating an elective case (as a function of the number of cases treated)

MU_{el} = marginal disutility of turning away an emergency case (as a function of the number of cases turned away)

MDU_t = marginal disutility of placing an elective on the waiting list case (as a function of the number of cases on the waiting list)

$[1\text{-}F(B_{em})]$ = probability of all emergency case beds being full

The intuitive explanation is that the optimal allocation is where the increase in expected utility derived from allocating one more bed to the emergency sector, less the increase in disutility in the elective associated with the allocation of a bed to the emergency sector, is exactly equal to the increase in utility derived from allocating one more bed to the elective sector, less the increase in expected disutility in the emergency sector as a result of this. The hospital, therefore, takes account of uncertainty when allocating capacity such that it shifts off the PPF and the optimal allocation adjusts that relative utilities by $[1\text{-}F(B_{em})]$. Giving:

$$\frac{MU_{em}[1-F(B_{em})]+MDU_{t}[1-F(B_{em})]}{MU_{el}+MDU_{w}}=MRPT[1-F(B_{em})] \qquad (3.28)$$

This represents, mathematically, the result indicated by the geometric analysis, such that the optimal allocation is where the hospital maximises the difference between utility and disutility, i.e. allocates such that they operate on the highest expected utility curve and lowest expected disutility curve.

Section 6: Conclusions

In this chapter we have extended current theory by considering hospital bed allocation issues using a standard production possibility/indifference curve framework. We have introduced demand uncertainty and output heterogeneity into the representation; two aspects that are fundamental to the specification of the production process. Through this analysis we have shown that there is an inherent cost of uncertainty, and that this alters the resource allocation decisions of the rational hospital. We have also shown that that excess capacity can be consistent with an efficient allocation of resources when viewed as a response to demand uncertainty, and have identified a shadow production possibility frontier, which represents this response. The existence of planned reserve capacity has been shown to be rational from a private perspective. However, hospitals in the NHS implicitly have some kind of social perspective, and the existence of reserve capacity has been shown to be entirely consistent with this wider objective, although identification of an optimal allocation depends on the magnitude of the social elements. To show the response to uncertainty more formally we constructed a mathematical model that allowed the optimal allocation decision to be identified. This model showed that when a social perspective is introduced reserve capacity is a rational response, and that this response depends on the arguments that enter the hospital's utility function, and the relative magnitude of each argument, which is consistent with the geometric representation. the model provides a conceptual framework for an empirical investigation into the hospital's cost function. It is to this that we now turn.

Chapter 4: Estimating a hospital cost function

Section 1: Introduction

Interest in specifying and estimating cost functions for hospitals has grown in the last three decades largely reflecting a greater availability of data. Much of this work, however, lacks theoretical content. The literature has generally failed to address some of the key issues that may affect the structure of hospital costs, most importantly the heterogeneity of output and demand uncertainty. The aim of this chapter is to build on recent advances in the area to estimate a hospital cost function that incorporates production responses to demand uncertainty, and introduces output heterogeneity in an explicit manner.

In Chapter 2 we reviewed a number of hospital cost function analyses and highlighted some of the issues dealt with within this literature. We identified the underlying theories relating to cost function analysis. In particular, we highlighted the underlying neo-classical theory of the firm and the implications for cost function analysis, and the problem of assuming that the cost function describes the minimum cost of producing a given output. Reliance on neo-classical theory has important consequences for the specification of a regression equation if observed costs are interpreted as minimum costs. Independent variables may only comprise output quantities and input prices. Other potential independent variables, such as capacity utilisation, cannot be included in this type of specification because they do not determine the minimum cost; rather they explain deviations from the theoretical minimum. The importance of this was further highlighted in Chapter 3 where we showed how reserve capacity can be viewed as an efficient response to demand uncertainty, but that it leads hospitals to operate within their production possibility frontier. Therefore, we suggest, analysis of hospital costs should take this into account.

In the review chapter we discussed other theories that have arisen out of the discussion of hospital behaviour, but concluded that most work has relied on conventional assumptions in the estimation of cost functions. Econometric, rather than economic, considerations have dominated with cost minimisation and, hence, duality assumed.

It is not necessary to duplicate the reviews of hospital cost function estimation techniques here; a summary of which was provided in Chapter 2. Rather, we concentrate on building on recent developments in the literature by Friedman and Pauly (1983), Pauly and Wilson (1986), Gaynor and Anderson (1995) and Carey (1996), which incorporate amendments for demand uncertainty and highlight the impact that this has on the duality between cost and production conditions. In the following sections we will draw on previous work, where applicable, and formalise the relationship between demand uncertainty, output heterogeneity, and costs with the aim of developing and empirical estimate that is consistent with the theoretical specification of Chapter 3.

The chapter is structured as follows: Section 2 will develop a theoretical background to the estimation process. Section 3 will consider the specification of the cost function and will present the results. Section 4 will draw some conclusions.

Section 2: The theoretical background

Building on some of the most recent work in estimating hospital cost functions, we note that one common dimension of hospital output is the planned excess, or reservation, capacity that is provided to ensure treatment is responsive to demand uncertainty. If no adjustment is made for this reservation capacity when estimating hospital costs, the hospital will appear to be operating inefficiently. It is, therefore, important to adjust the cost function to take account of this factor.

The majority of previous studies have considered hospital demand as homogeneous. This leads to a number of implications; all demand is either endogenous or exogenous, reservation capacity exists to service all demand, and, building on the theoretical specification in Chapter 3, that the same social cost is attached to turning away all types of demand. A significant improvement is to recognise that demand is heterogeneous, a simple classification would be, as in Chapter 3, to specify output as elective and emergency cases, where it is possible to queue some kinds of demand (elective), whereas other demands must be treated immediately or turned away (emergency). We assume throughout that electives are endogenous, and that emergencies are exogenous and arrive with some uncertainty attached. As highlighted in the theoretical specification, it is due to this latter characteristic that hospitals hold reservation capacity.

The model we will develop here will specify the production process by referring to three categories of input with two levels of variability; those that are related to output and those that are related to capacity. The three inputs are capital, which is fixed, quasi-fixed, which are staffed beds, and variable, such as materials, drugs and dressings. We will assume, as previous authors have done, that the hospital chooses fixed and quasi-fixed inputs on the basis of ex ante demand expectations and that once chosen there is no ex post adjustment. We will also assume that variable costs are chosen on the basis of demand expectations but that ex post these inputs adjust to the realised demand level.

Following the discussion of the theoretical model framework in chapter 3 we will specify the objective function of the hospital as one of expected cost minimisation, however, we will also adjust for demand uncertainty as the hospital will operate within the standard PPF. In addition, we will introduce two further elements of cost, such that costs comprise the costs of production and the social costs attached to turning patients away, and placing, (or leaving), elective patients on a waiting list. As in the previous chapter, the most general specification of the optimisation problem is therefore one adopting a social perspective:

$$\begin{aligned} E(C) = K + sc_w[el_w] + \int_{q_{em}}^{\infty} sc_t[(x - q_{em})]f(x)dx + c_{Bem}[q_{em}] + c_{Bel}[q_{el}] \\ + \int_0^{q_{em}} c_{v_{em}}[x] \cdot f(x)dx + \int_{q_{em}}^{\infty} c_{v_{em}}[q_{em}]f(x)dx + c_{v_{el}}[q_{el}] \end{aligned} \tag{4.1}$$

where,

E(C) = expected costs

c_{vem} = variable cost of treating emergency cases

c_{vel} = variable cost of treating elective cases

c_{Bem} = quasi-fixed cost of treating emergency cases

c_{Bel} = quasi-fixed cost of treating elective cases

sc_t = social cost of turning away emergency cases

sc_w = social cost of putting patients on the waiting list

q_{em} = quantity of emergency cases

q_{el} = quantity of elective cases

el_w = elective waiting list size

K = Capital

Output is the number of cases treated, where the total number of cases treated per bed, per day, is equal to the number of beds allocated to the sector divided by length of stay.

Demand in the emergency sector is assumed to be random and exogenously determined according to an, as yet, undefined process. For any chosen bed allocation level in the emergency sector there will be a probability that the all the beds are not filled, which is determined by the probability distribution around the demand function. Conversely, there will be a probability that the number of arrivals will exceed capacity, if this happens patients are turned away. The number of beds allocated to the emergency sector also determines the number of patients turned away.

There is assumed to be excess demand in the elective sector, such that demand is always greater than capacity, *B,* for all *B*. Therefore the queue is never empty. Consequently, the hospital knows that each bed allocated to the elective sector will be filled with certainty and can choose the number of elective cases they treat. Furthermore, we will assume that for each decision period the hospital is aware of the total number of patients seeking treatment, such that the number of electives treated determines the waiting list size[1]. When all the elective beds are occupied any additional arrivals are assumed to join a waiting list.

Therefore, the arguments in equation (4.1) become:

$q_{el} = B_{el}/LOS_{el})$

$el_w = f(B_{el}/LOS_{el})$

$q_{em} = f(B_{em}/LOS_{em}, f(x))$

where,

$f(x)$ = the probability distribution around emergency demand

B_{em} = number of beds allocated to the emergency sector

B_{el} = number of beds allocated to the elective sector

LOS_{em} = length of stay in the emergency sector

LOS_{el} = length of stay in the elective sector

[1] More complex models of waiting lists exist, see for example Worthington (1987), however these models rely on arbitrary contructs such as 'discouragement factors' to enable stationarity to be induced in the models. Furthermore they require assumptions about intertemporal bed allocation decisions that are inconsistent with our model. We simply assume that elective demand is exogenously determined but that the total size of demand is known within any decision period. This does not affect the first order condition but does allow a straightforward exposition of the waiting list size.

Furthermore, let us assume that within any given decision period the hospital knows the number of elective patients requiring treatment. Let W represent the total size of elective demand such that:

$$el_w = (W - B_{el}) \qquad (4.2)$$

(This simply assumes that the waiting list is predetermined for any given period, this assumption does not affect the first order condition as long as $W > B_{el}$, i.e. the waiting list is non-zero.)

Furthermore, we know that,

$$B_{em} + B_{el} = B_T \qquad (4.3)$$

where B_T = total number of beds available

As in Chapter 3, let us also assume that the length of stay for electives and emergencies are equal and normalised to one, this allows us to use beds and cases interchangeably but is by no means necessary to what follows. By substituting B_{el} for $(B_T - B_{em})$ we get the following expression, which allows us to maximise with respect to B_{em}.

$$E(C) = K + sc_w[W - (B_T - B_{em})] + \int_{B_{em}}^{\infty} sc_t[(x - B_{em})]f(x)dx + c_{B_{em}}[B_{em}] + c_{B_{el}}[B_T - B_{em}]$$
$$+ \int_0^{B_{em}} c_{v_{em}}[x] \cdot f(x)dx + \int_{B_{em}}^{\infty} c_{v_{em}}[B_{em}]f(x)dx + c_{v_{el}}[B_T - B_{em}]$$

(4.4)

Where the expected costs and social costs are determined in the emergency sector by the number of beds allocated to that sector. The expected variable costs of production are determined by the expected mean of the probability density function truncated at B_{em} (i.e. terms 6 and 7 in equation 4.4, above). The quasi-fixed costs are determined with certainty, ex ante, based on the choice of B_{em}. The expected social costs are determined by the mean to the right of the distribution truncated at B_{em} (i.e. term 3 in equation 4.4).

This gives us the following first order condition (as in Chapter 3 we employ Leibniz's rule where B_{em} enters as a limit and a parameter within in the integral):

$$\begin{aligned} & sc_w'[W-(B_T-B_{em})]-\int_{B_{em}}^{\infty} sc_t'[x-B_{em}]f(x).dx \\ & +c'_{Bem}[B_{em}]-c'_{Bel}[B_T-B_{em}]+\int_{B_{em}}^{\infty} c'_{vem}[B_{em}]f(x)dx-c'_{vel}[B_T-B_{em}]=0 \end{aligned} \quad (4.5a)$$

The second order condition of which is:

$$\begin{aligned} & sc_w''[W-(B_T-B_{em})]+\int_{B_{em}}^{\infty} sc_t''[x-B_{em}]f(x).dx \\ & +c''_{Bem}[B_{em}]+c''_{Bel}[B_T-B_{em}]+\int_{B_{em}}^{\infty} c''_{vem}[B_{em}]f(x)dx+c''_{vel}[B_T-B_{em}] \end{aligned} \quad (4.5b)$$

which is positive as long as the first term is less than the sum of the remaining five terms, which is intuitively plausible as the social costs of turning an emergency patient away are likely to be greater than the social cost of placing a patient on the waiting list even without taking account of the private marginal costs of production. Under these conditions a minimum is identified.

Equation (4.5a) can be represented more simply as:

$$MSC_t[1\text{-}F(B_{em})] + MC_{Bel} + MC_{vel} = MSC_w + MC_{Bem} + MC_{vem}[1\text{-}F(B_{em})] \quad (4.6)$$

where,

MSC_t = marginal social cost of turning away emergency cases

MSC_w = marginal social cost of putting patients on the waiting list

MC_{Bem} = marginal cost of a staffed emergency bed

MC_{Bel} = marginal cost of a staffed elective bed

MC_{vem} = marginal variable cost of emergency case

MC_{vel} = marginal variable cost of elective case

$[1\text{-}F(B_{em})]$ = probability of a turning a patient away

That is, an expected cost minimising that takes a social perspective will allocate up to the point where the total, (i.e. private and social), expected marginal costs of allocating beds to the elective sector equals the total expected marginal costs of allocating beds to the emergency sector.

or:

$$MSC_t[1\text{-}F(B_{em})]\text{-} MSC_w = MC_{Bem} + MC_{vem}[1\text{-}F(B_{em})]\text{-} MC_{Bel} - MC_{vel} \quad (4.7)$$

Such that the difference in expected social cost between the two sectors equals the difference between the expected marginal costs of allocating an extra bed to the emergency sector, less the costs of allocating an extra bed to the elective sector. The existence of reserve capacity is consistent with social cost minimisation.[2]

Let us briefly reconsider this issue in such a way that the costs of uncertainty become intuitively more appealing, that is, by separating out expected costs and costs of uncertainty due to ex post rigidity. This extension of the general specification is useful in considering the various dimensions of uncertainty. Essentially there are three aspects.

Since we are assuming that there is ex post flexibility in the choice of variable inputs, but that hospitals must choose the number of staffed beds ex ante (with no ex post adjustment), equation (4.5) allows us to identify one element of the cost that is introduced due to this ex post rigidity, that is $c'_{Bem}[B_{em}]$. If there were ex post flexibility of bed allocation the expected costs of staffed beds would be:

$$c'_{Bem}[B_{em}]\int_{B_{em}}^{\infty} f(x)dx \qquad (4.8)$$

Therefore, the rigidity introduces an extra cost which is equal to the difference between the expected costs, given flexibility, and the actual costs with rigidity, such that:

[2] Note that the fixed capital costs drop out of the specification when the first order condition is derived so we move to a short-run variable cost function.

$$c'_{Bem}[B_{em}] - c'_{Bem}[B_{em}]\int_{B_{em}}^{\infty} f(x)dx \quad (4.9)$$

which is the marginal cost of an emergency bed less the marginal cost of an empty bed weighted by the probability of being filled. As the probability of turning a patient away tends to zero this term will tend towards the marginal cost of a bed. If we assume that the cost of a bed varies linearly with the number of beds this becomes:

$$c'_{Bem}[B_{em}][1 - \int_{B_{em}}^{\infty} f(x)dx] \quad (4.10)$$

which is:

$$c'_{Bem}[B_{em}](F(B_{em})) \quad (4.11)$$

The marginal cost of a bed multiplied by the probability of the bed being empty. So as the probability of turning a patient away falls the cost of uncertainty rises, therefore the hospital must trade off this increasing expected cost against the cost of turning a patient away. The extra cost associated with uncertainty means that the hospital operates inside its production possibility frontier (PPF), and appears to be operating inefficiently. However, the efficient level of capacity utilisation under conditions of uncertainty must take account of the cost of uncertainty. The hospital will attempt to minimise expected costs, and this is consistent with operating with reserve capacity. Any deviation from the *expected* cost minimising position will, however, result in the hospital operating off its shadow PPF, (PPF' as outlined in Chapter 3), and now the hospital will be operating inefficiently. It is the extent of this deviation that must be taken into account when estimating the cost function.

The cost function has all the usual properties except duality and independence of demand and cost, since the firm is constrained to have the capacity to meet randomly fluctuating demand with some probability, but will not generally be producing on the production possibility frontier. In addition to this cost, the hospital must also enter into the spot market to adjust variables inputs to satisfy the ex post realised level of demand. The extent of this adjustment corresponds to the difference between ex ante expected demand and ex post realised demand:

$$\dot{c}'_{vem}(q_{em} - [\int_0^{B_{em}} xf(x)dx) + \int_{B_{em}}^{\infty} B_{em} f(x)dx]) \tag{4.12}$$

where q_{em} is realised output and $\dot{c}'_{vem}$ is the spot market price for variable inputs (this can be negative or positive depending on whether more or less inputs are required, where the cost of uncertainty are represented by the adjustments costs). The total impact uncertainty has on costs therefore depends on the level of ex post adjustment, and the extent to which demand differs from ex ante expectations. That is, how far off the revised, ex ante, production possibility frontier (PPF' in the geometrical exposition in Chapter 3) the hospital is operating ex post.

This gives three potential elements of the impact of uncertainty. The first refers to the hospital operating with the PPF, i.e. such that duality is broken. The second refers to the hospital operating off its shadow PPF', this occurs when the actual admission level differs from the expected admissions such that the hospital once more is operating inefficiently, i.e. off the ex ante cost minimising level of production. The final element refers to the cost of entering the spot market to purchase the variable inputs required, (or reduce the level of variable inputs), to meet ex post demand. Using these notions of demand uncertainty we can develop an empirical model of hospital costs.

Section 3.1: The cost function

While the previous section developed a general social welfare cost function, we will now proceed to an empirically testable hypothesis of whether or not demand uncertainty and product heterogeneity affect hospital costs. This empirical specification will be developed using only data on private costs, since these are the only data reported and observed and by focusing on private, rather than social costs we have a 'cleaner' test of our hypothesis. As will become clear in Chapter 5, we need to estimate private marginal social costs before we can estimate social marginal costs, and we need to undertake further modelling in order to derive estimates of social costs. Nonetheless, the estimation of costs still relies on the underlying theory developed above.

From the discussion of the theoretical background, it is clear that if demand uncertainty is ignored the cost function will be misspecified. Therefore, in order to provide estimates that are consistent with our theoretical specification, we introduce demand uncertainty into our cost function estimate.

To allow an empirical specification of our hypothesis that demand uncertainty and product heterogeneity affects costs let us specify output as:

$$Q_T = Q_{em} + Q_{el}$$

Where the subscripts T, em and el, represent total, emergency and elective output, respectively. Let us define the production process such that:

$$Q_{em} \text{ and } Q_{el} = f(k, v_f, v_v)$$

This allows, from equation 4.5, the production function under conditions of demand uncertainty to be generally specified as:

$$Q_T = f(k, v_f, v_v, [1\text{-}F(Q_{em})])$$

Where $[1\text{-}F(Q_{em})]$ represents the probability that the hospital will be full, or alternatively that an emergency patient will be turned away. The implication is that there is an optimal level for this probability determined by the relative magnitudes of the social and private costs of production. Therefore, the hospital's aim is to minimise costs subject to demand uncertainty, and the maintenance of reserve capacity consistent with achieving its optimal level of $[1\text{-}F(Q_{em})]$.

This cost function has the normal properties except duality and independence of demand and cost. With the existence of demand uncertainty, it is no longer tenable to assume that that inputs have been hired to produce the observed output at least cost. Rather we must assume that inputs have been hired to minimise ex ante expected costs. Therefore, costs will depend on expected levels of admissions, as well as actual admissions, and the extent of divergence between the two.

From equations 4.8-4.12, above, we need to pick up three components of demand uncertainty. First, we need to adjust for the deviation of actual from expected output to consider the cost of ex post adjustment of variable inputs. Secondly, we need to consider the impact of the hospital operating off its ex ante cost minimising position (i.e. the shadow PPF). Finally, we need to consider the impact of operating within the PPF efficiency locus. One way of incorporating the first two elements is to include estimates of demand expectations and ex post errors of demand in the cost specification. In addition, if we also want to consider the impact that operating within the PPF has on hospital costs, we need to include a variable to pick up reserve capacity. We will now consider these factors in the specification of a cost function, including demand uncertainty and output heterogeneity.

3.2: Specification of the cost function

We estimate a total variable cost function to reflect production responses to demand uncertainty. Although in principle there are a number of different measures we could have used, such as average cost or total costs, we used total variable costs to enable us to pick up the responses to demand uncertainty, and their impact on short-run cost structures. Since we model for output heterogeneity, the use of an average cost is precluded, and total costs including capital would not focus on the short-run capacity utilisation issues in which we are interested.

We assume a monthly planning period within which hospitals allocate available bed capacity between elective and emergency cases and we estimate an annual cost function. In the cost specification we include estimates of actual emergency and elective output, the inverse of the occupancy rate, which measures the extent of excess capacity and also controls for fixed factors and length of stay, an estimate of the casemix, and a number of dummy variables. A variable is included to capture the impact of demand uncertainty. This variable is included to test empirically whether or not uncertainty impacts hospital costs. It is hypothesised that if the coefficient on this variable is positive and significant demand uncertainty imposes a real cost on hospital production.

As an initiation to our investigations we assume a total variable cost function of the following form:

$$TVC = f(ELADM, EMADM, RES, BEDSEL, BEDSEM, CASEMIX, INVOCC, OPV, DAYATT, AEATT, WI, DV_i) \quad (4.13)$$

where TVC is total variable costs. ELADM is total inpatient elective admissions. EMADM is total emergency inpatient admissions. These output measures pick up the two main outputs of concern here. BEDSEL are the total number of staffed beds allocated to the elective sector. BEDSEM are the total number of staffed beds allocated to the emergency sector. In the short-run the overall capacity is fixed. There is, however, still a choice over the level of different outputs. Therefore, to maintain consistency with the theoretical specification, we have separated beds into those allocated to the elective sector and those to the emergency sector. This will enable us to determine the cost of staffed beds allocated to each sector. CASEMIX is an estimate of casemix based on HRG weights. Since output is more complex than our two-category definition, this variable is necessary to pick up case complexity. INVOCC is the inverse of the occupancy rate (BEDS.365/LOS.ADM), which measures the extent to which reserve capacity exists. The main focus of the estimation is on inpatient admissions, however, hospital produce other outputs; namely outpatient visits and day attendances. OPV are outpatient visits. DAYATT are day attendances. AEATT are accident and emergency outpatient visits. WI is wage index based on the average wage rate for each hospital, and is a rather imperfect measure of the factor prices faced by each hospital. Although we do recognise that within the NHS factor prices are set through central bargaining processes and the discrepancy between providers should not be large, therefore, a wage index may well capture most of the variation. RES is the variable that captures demand uncertainty (discussed in section

3.3 below). DV_t represents a range of group dummy variables (discussed in more detail 3.5).

The above discussion has highlighted the importance of demand uncertainty, and the construction of an appropriate variable to pick up the influence of this uncertainty on costs. Before considering the estimation of the cost function we will consider, in some detail, the different ways of incorporating demand uncertainty, and the different ways that demand forecasts have been estimated.

3.3.1: Incorporating estimates of demand into a cost function

An important question is how should uncertainty be included in the cost function? A number of issues need to be addressed. These can best be assessed with reference to prior studies, albeit small in number.

Friedman and Pauly (1983) and Pauly and Wilson (1986) both employed a measure of the ratio of expected to actual demand to pick up the impact of uncertainty. The main problem with this approach is that on average, if we hypothesise that demand is truly random but with a known distribution, we may expect hospitals to be able to predict the aggregate demand reasonably well within a given time period. Furthermore, a ratio measure of this kind will not pick up the magnitude of the uncertainty. Therefore, the ratio of forecasted to actual demand may be a poor proxy for the level of uncertainty in the system. Indeed Friedman and Pauly reported that the mean of their forecast/actual admissions variable was 1.008. Therefore use of this variable may not be particularly informative.

Gaynor and Anderson (1995) and Carey (1996) both employed a measure of the standard error of the forecast. Gaynor and Anderson used the standard error of actual admissions whereas Carey used the standard error of the average daily census (ADC) as a proxy for uncertainty. This allowed the researchers to examine the impact of fluctuations around the expected mean of demand, although it is questionable whether use of the standard error of ADC will pick up the magnitude of the uncertainty, since it adjusts admissions by length of stay and the time period. Furthermore, it should be noted that both these authors used the standard error of the total actual admissions rather than the separating forecasted from unpredicted admissions. It is clear that demand fluctuations per se are not a problem, rather demand uncertainty is a problem because, by definition, it is unpredictable. Therefore, we need to take care to separate out the predicted from unpredicted demand (as Friedman and Pauly did). Consequently, it is likely that the estimates used by Gaynor and Anderson, and Carey, will significantly over estimate the uncertainty in the system. Since, however, they both used annual data this will temper the extent of this overestimation. Although a further compounding factor is that all the previous attempts to incorporate demand uncertainty have assumed that all demand is unplanned and this in itself is likely to lead to overestimates of the extent of demand uncertainty.

Nonetheless, these previous attempts assist in highlighting the options that are available when incorporating demand uncertainty. Consequently it is possible to identify, at least, four approaches:

i) Include the standard error of the unpredicted demand from the ADC forecast equation. This will give estimate of divergence from expected ADC. The main problem is that it does not give a feel for the size of the uncertainty in absolute terms.

ii) Include a ratio of actual to estimated demand. The problem is that these tend to smooth out such that this variable will tend towards one.

iii) Include the standard error of the unpredicted demand from the forecast of total admissions. This will give estimate of divergence from expected total admissions and will give a better feel for the size of the uncertainty in absolute terms.

iv) Include the sum of actual errors from forecasted equation. These might represent the best feel for the impact of uncertainty and the impact of divergence from the ex ante PPF'.

This gives a number of options. However, since we are trying to model the impact of the divergence from the ex ante PPF and estimate the impact this has on costs, perhaps the best of these options are represented by (iii) and (iv) above. These provide a clearer measure of the extent to which the hospital is operating off its ex ante cost minimising level of production and are the two approaches we will adopt in the estimation in section 3.6.

3.3.2: Previous estimates of demand

Friedman and Pauly (1983) estimated the demand for hospital services using data from monthly reports to the Hospital Administration Services in the US. The data comprised 72 monthly reports for 870 hospitals covering the years 1973 to 1978. However, in order to overcome some problems in the data set, (i.e. missing entries, negative numbers and large jumps that were inexplicable), they aggregated to a quarterly base. Clearly this raises issues regarding the reliability of their data, however these are not issues we wish to deal with here. The main problem with aggregating the data set is that they lose valuable information regarding fluctuations in demand, which may, as other authors have recognised, (see Joskow, 1980), exhibit strong monthly, weekly, or even daily components. Disaggregate data reveals the most information in this context.

Friedman and Pauly applied a simple structural model to forecast demand, where, it was assumed, the only information hospitals will use is demand experience. This led the authors to employ a method of minimum average prediction variance where S_t was defined as a seasonal factor for a quarter t, so that:

$$S_t = \Sigma f_j D_{jt} \qquad (4.14)$$

Each dummy variable, D_{jt}, corresponds to a season, where j = 1,..,4. For the stationary process,

$$q_t = qS_t + U_t \qquad (4.15)$$

Where q_t is quantity demanded and q is constant over time. U_t is the normally distributed random variable. A first-order autocorrelation in the error term was assumed such that:

$$U_t = \rho U_{t-1} + e_t \tag{4.16}$$

The autocorrelation process acts as a way of capturing trends or drift in the series over time. Perhaps unsurprisingly, given the inclusion of this process, the authors found that a time trend had little effect on the estimation of demand.

Using this simple model, Friedman and Pauly found that the size and parameter values of the seasonal factors were unstable, and the autoregressive parameter varied widely across the sample. They suggested that this diversity among the estimates meant they could not justify pooling the sample. Consequently, they estimated a different demand function for each hospital. The differences in goodness of fit between the hospitals were quite startling ranging from 0.16 to 0.93 as based on the R^2 values.

However, they noted, importantly, that the method used by each hospital for its own planning purposes is not observed. Since we are interested in how demand uncertainty affects the structure of hospital costs, this is perhaps the key to the whole exercise. The impact of uncertainty on the structure of hospital costs will depend on the extent to which the hospitals themselves experience uncertain demand, and this will depend crucially on their expectations of demand. Therefore, whilst the approach adopted by Friedman and Pauly is relatively crude, it may best resemble the expectations of hospital themselves.

Following this general approach, Gaynor and Anderson (1995) used data from the American Hospital Association's Annual Survey of Hospitals to estimate a demand function to incorporate in their cost function. Data from 1980-1987 were used to forecast admissions. Gaynor and Anderson's data, however, were annual, and therefore, they could not use the time series method adopted by the original authors, and had to 'exploit the cross-sectional as well as the time series variation in the data', or what might more contemporarily be termed panel data techniques. The use of annual data also meant that valuable information regarding the within year fluctuations was lost.

In order to facilitate the approach chosen, the authors grouped the hospitals by geographic area (i.e. those in the same urban, rural, or urban area were grouped together). The implicit assumption being that demand would be influenced in some systematic way by the geographic location within which hospitals were located.

The authors employed a forecasting equation including a three-lag dependent variable, a time trend, and hospital specific dummy variables, allowing the authors to calculate forecasts for each hospital, even though they were pooled into geographic areas. It is not clear how the grouping affected the lagged dependent variables. Since hospitals were grouped by geographic area, it is also not immediately obvious how lags could be included in such a model.

The authors reported excellent fits for the forecasting equations; R^2 of between 0.97 and 0.99. These values are indeed very high and somewhat surprising given the stochastic nature of demand hypothesised. In fact, if hospitals were able to predict demand with such accuracy based primarily on previous years, then it would suggest that demand uncertainty may not be a problem since they appear to be able to explain almost all of the demand fluctuation. Herein lies the problem with using annual data; the fluctuations in demand between years are likely to be quite small relative to the fluctuations within years. That is, it may in fact be surprising if the hospital could not forecast with reasonable accuracy the level of demand on an annual basis[3]. However it is likely to be the week to week, or month to month fluctuations that are likely to impact on the hospital's cost structure. Therefore, whilst the approach is similar to that employed by Friedman and Pauly, the use of annual data causes problems in assessing the level of uncertainty. This is reflected in the size of R^2 reported.

Carey (1996) employed a very similar approach to Gaynor and Anderson, once more using annual data based on nine years of observations employing a three-year lag structure. Dummy variables were included to control for the impact of teaching status and ownership along with variables for casemix, and the level of competition. Carey's

[3] If Gaynor and Anderson had adopted the same approach as Friedman and Pauly, and Pauly and Wilson, i.e. ratio of forecasted to actual admissions then the variable would, on average, have been 1.000034 (reflecting the high R^2 for the forecasting model).

results were very similar to Gaynor and Anderson in terms of R^2 values, ranging from 0.98 to 0.99.

Perhaps the greatest problem with these two approaches, however, lies in the application of the demand function estimates. In estimating the impact of uncertainty on the cost structure both Carey and Gaynor and Anderson used the standard error of the forecasted admissions. These measures of demand uncertainty will only pick up the expected fluctuations in demand – i.e. the ones the hospitals can predict. Therefore the theory they are testing is essentially different to that originally tested by Friedman and Pauly, i.e. that unexpected demand fluctuations affect hospital cost structures. If hospitals can accurately predict the fluctuations then there is no reason to expect this to impact on costs. The measure of fluctuations of actual demand around expectations is picked up through the standard error of the noise component, which will be minimal with R^2s of 0.99. (This is borne out by the estimates of the probability of turnaway calculated by Carey which appear to be too high given the R^2 estimated in her equation – see Chapter 5 for a full discussion).

To be consistent with the approach adopted by Friedman and Pauly, and to test the same hypothesis, Gaynor and Anderson and Carey should have taken the forecasted admissions away from actual admissions. This would have enabled them to obtain the size of the unexpected admissions for each hospital, then examine the impact this has on the cost structure.

Several issues come out of the above discussions that are pertinent to the estimation of a demand function in the NHS. First, the issue of demand heterogeneity, which has not been dealt with for various reasons. If hospitals do have an objective function that differentiates between demand that can be queued and that which requires immediate treatment, and they take account of the social cost imposed by non-treatment of cases turned away, then it becomes apparent that we should be attempting to disaggregate the data on admissions to reflect this. If we fail to adequately take account of this factor, problems in the estimation process arise. If part of the demand is planned it is, by definition, endogenous. Any forecasting of this element of demand will be meaningless, since hospitals will themselves determine the level of admissions. This also has implications for the estimation of cost functions. If some admissions are endogenous, whilst others are exogenous, this should be reflected in the cost function estimation procedure. Secondly, if it is hypothesised that hospitals are concerned with the social costs of turning patients away, and that this is one of the reasons why reserve capacity is held, then failure to recognise the heterogeneity of demand may lead to underestimates of the reserve capacity held, and over-estimates of the probability of turning patients away, (and consequently underestimates of the implied social cost of turning patients away). If all demand were planned and could be queued, then there would be no need to hold reserve capacity. It is, therefore, our intention to separate demand into two broad categories; planned and unplanned admissions to enable us to examine the implication of stochastic demand for the behaviour and cost structure of UK NHS hospital trusts.

3.4: Data

Monthly data on admissions were obtained on a sample of 148 NHS hospital trusts from CHKS Ltd over a three-year period (1993-1995). CHKS routinely collects data from NHS hospital trusts and holds the largest patient-based data set in Europe. It is currently involved with the Casemix Office in devising health related groups (HRGs) in the UK. and publishes annual reports listing data on the top 200 HRGs in the UK. Based on returns to CHKS by NHS trusts each hospital admission was categorised as either emergency, elective, or obstetric (including transfers from other NHS providers), corresponding to CMD codes. Data were also supplied on lengths of stay, bed numbers and a number of other variables listed in Table 4.1. Data were also collected on the catchment demand and estimated percentage share of demand for each provider (taken from Department of Health statistics and supplied by Ivan Csaba) and cost data (taken from CIPFA Healthcare databases). These data were tied together by CHKS since the data they provided were anonymised[4].

Table 4.1: Activity and cost variables

Variable name	Definition
ACTIVITY	
ADMS	Total number of admissions
EMERGADM	Number of admissions categorised as emergency (by provider)
ELADM	Number of admissions categorised as elective (by provider)
OTHADM	maternity care and transfers from other providers
BEDS	Total number of staffed beds
ALOS	Average length of stay (days) for all admissions
ELALOS	Average length of stay (days) for elective admissions
EMERGALOS	Average length of stay (days) for emergency admissions
OTALOS	Average length of stay (days) for admissions categorised as other
DAYATT	Total number of day attendances
AEATT	Total number of non-inpatient Accident and Emergency attendances
OPATT	Total number of non-A&E outpatient attendances
CATCHSH	% share of beds in total area (DoH (1989))
CATCHD	Catchment population of provider in 1989 based on DP41 (DoH (1989))
VALDWAIT	Total number of elective patients waiting for admission
AVCMIX	Casemix weight based on DRG weights
OPERATING EXPENSES	
BOARD	Board members' fees
SAL	Staff costs
SUPCLIN	Supplies and services: Clinical
SUPGEN	Supplies and services: General
ESTAB	Establishment
TRANSEXP	Transport
PREM	Premises
OTHER	Other expenses including miscellaneous services from other NHS providers
TOTAL VARIABLE COSTS	Sum of operating expenses

[4] See Appendix 3 for a full list of data and descriptive statistics.

The data were collected for financial years 1992/93, 1993/94 and 1994/95. However, they were not balanced and there were more observations for the later time periods. Furthermore, in attempting to bring together data from different sources we had missing observations, since some of the sample provided by CHKS were outside England, whereas the data on costs only refer to England. Since the ultimate aim of the work is to estimate a cost function for NHS providers we are limited by the lowest common sample.

It is possible from the available data to disaggregate admissions into planned and unplanned demand. Elective cases were defined as planned admissions and unplanned cases were defined as emergency and obstetric admissions and transfers from other providers. This was done for practical as well as theoretical reasons. Emergency admissions are obviously unplanned. Regarding obstetric admissions, although women are encouraged to register with their local hospitals, and are sometimes induced if they do not deliver within a pre-determined period (often two weeks) of the due date, they are still largely unplanned in terms of the monthly planning timeframe within which the provider units are operating. Transfers from other providers also tend to be unplanned within the monthly planning period. Emergency, obstetric and transfers from other providers will collectively be referred to as 'emergency' admissions from this point (labelled as EMADM).

The next section outlines the demand estimation and specification of the RES variable (in equation 4.13) that is included in the cost function in order to estimate the impact of demand uncertainty.

3.5.1: Demand forecasting

The separation of admissions into 'elective' and 'emergency' allows us to estimate demand for unplanned emergency care and cost the hospital production reaction to demand uncertainty. Since we are grouping together cross-sectional and time series data we will use panel techniques to estimate the demand-forecast equation. However, as recognised by Friedman and Pauly (1983), we may be less concerned with the statistical rigours of the estimation process, and more concerned with whether the estimation process is a good approximation to the hospital's own ability to forecast demand. In other words, as with all forecasting equations, the performance criteria rest on their ability to forecast rather than explain behavioural relationships. Demand estimation can take a number of forms; in general there is a trade-off between statistical criteria and forecasting ability. However, there is no correct method for specifying demand-forecast equations. Previous authors have used different approaches, as we reviewed in section 3.3.2. We will utilise available data to estimate three of the most intuitively appealing methods.

If we assume that we are dealing with a random stationary process, which is probably justifiable given the relatively short time period we are dealing with, then it is possible to apply a fairly simple estimation procedure. Notwithstanding this conclusion, it is still possible to estimate future demand through a number of different demand specifications. We will adopt three approaches; the first will be a departure from what has been done before, in that, we will estimate demand on the basis of the demand characteristics faced by the hospital utilising available data listed in Table 4.1. This is a fuller specification of the structure of demand faced by the hospital. The first model to be estimated is of the form:

$$D_t = f(CATCHD, CATCHSH, BEDS, DV_t) \qquad (4.17)$$

Where D_t is demand in period t, CATCHD and CATCHSH are defined as above in Table 4.1 and DV_t represents the monthly dummy variables.

The other two approaches follow the lines of previous studies. They assume that hospitals estimate demand solely and directly on the basis of previous experience. Two distinct estimation processes are followed; a lagged dependent variable model, along the lines of Gaynor and Anderson (1995) and Carey (1996) giving rise to the following specification:

$$D_t = f(D_{t-1}, DV_t) \qquad (4.18)$$

where D_t and DV_t are as above and D_{t-1} is demand in period t-1.

A more rigorously specified impact of previous demand is captured by an autoregressive model, following Friedman and Pauly (1983), of the form:

$$D_t = D.DV_t + \rho(D_{t-1} - D.DV_{t-1}) \qquad (4.19)$$

Where D_t and DV_t are as and D is constant over time and ρ represents the autocorrelation coefficient between periods[5].

3.5.2: Demand estimation methods

Since we have cross sectional and time series data we will employ panel data estimation techniques, where, essentially there are two techniques to choose from: fixed effects and random effects. Where the general model can be represented as:

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it} \qquad (4.20)$$

where α_i in the fixed effects model represent a separate constant term for each specified unit, ε_{it} is a classical disturbance term. The fixed effects model is a classical regression model where the main complication arises if the number of units, i, is large such that it becomes computationally cumbersome.

The random effects model can be represented by the a similar equation:

$$y_{it} = \alpha + \beta' x_{it} + \mu_i + \varepsilon_{it} \qquad (4.21)$$

[5] Friedman and Pauly estimated ρ to be, on average, 0.62,where ρ represents the autocorrelation coefficient between periods.

where μ_i is an individual specific disturbance term. The random effects model is a generalised least squares model where for a given i the disturbance terms are correlated because of their common component μ_i.

The random and fixed effects models can be extended to incorporate time-specific effects as well (this model is referred to as the two factor or two-way model). The general specification of this model is, for fixed effects:

$$y_{it} = \alpha_i + \gamma_t + \beta' x_{it} + \varepsilon_{it} \quad (4.22)$$

The model therefore has a group effect for each group, α_i, and a time effect for each period, γ_t. For the random effects model a similar equation can be represented of the form:

$$y_{it} = \alpha + \beta' x_{it} + \mu_i + w_t + \varepsilon_{it} \quad (4.23)$$

where, μ_i is as above in equation (4.21) and w_t is the time specific disturbance term.

The question this raises is; which is best in terms of testing our hypothesis, random or fixed effects? The choice of approach depends on a number of factors, some theoretical, others practical. The fixed effects model allows us to capture systematic differences across units that represent parametric shifts in the regression function. These shifts are captured in the constant term where each unit has a different α_i to estimate. Hence the model is often referred to as the least squares dummy variable model. The fixed effects model does, however, have a number of restrictions. First, the results of the analysis can only be applied to observations within the sample – i.e. it is not possible to extrapolate outside the sample taken. Secondly, it cannot deal with independent variables that are time invariant. Finally, and more practically, the dummy variable approach takes up a large number of degrees of freedom, which many panel data series will not be able to cope with, i.e. those with a relatively short time period and large number of units of observation. Under these circumstances it may be more appropriate to use the random effects model, which assumes that the individual specific effects are randomly distributed across the cross-sectional units. However, the random effects model also has problems. Unlike the fixed effects model it assumes that the individual effects are uncorrelated with the other regressors, this may not be justifiable. Furthermore, the random effects model may suffer from inconsistency due to omitted variables (see, for example, Hsiao (1986) for a fuller discussion of issues relating to panel data estimation techniques).

The general approach we will adopt follows that suggested by Greene (1993) where the equation is first specified as a classical ordinary least squares (OLS) regression model, we will then estimate one way and two fixed and random effects models. We will determine the 'best' model on the basis of two statistical tests devised specifically for panel data estimation techniques. The first is a test designed by Breusch and Pagan (1980) to test for the random effects model based on the OLS residuals. Essentially this is a Lagrange multiplier (LM) test which tests for the restriction $Corr[w_{it},w_{is}] = 0$, that is testing for whether the individual error terms are serially correlated. If they are then this suggests that there are indeed individual effects within the selected groups (which may be random or fixed). Where:

$$H_0: \sigma_u^2 = 0 \qquad (\text{or } Corr[w_{it}, w_{is}] = 0) \qquad (4.24)$$

$$H_0: \sigma_u^2 \neq 0$$

(The LM test is distributed as chi squared with one degree of freedom., where the test statistic can be found in standard texts, e.g. Greene, 1993).

The second is a test devised by Hausman (1978) which enables comparisons to be made between the fixed and random effects models on the basis of a Chi squared test and tests whether the random effects are correlated with the regressors. Under the null hypothesis of no correlation, OLS estimates should be consistent with other methods, in other words $\beta_{OLS}-\beta_{MLE} \rightarrow 0$. Hausman devised a formula to estimate the variance of the difference, $V(\beta)$, such that a Wald statistic can be calculated:

$$[\beta_{OLS}-\beta_{MLE}]'[V(\beta_{OLS})- V(\beta_{MLE})]^{-1}[[\beta_{OLS}-\beta_{MLE}] \qquad (4.25)$$

which is asymptotically distributed as chi-squared with 4 degrees of freedom. A significant value indicates that the regressors are correlated with the random effects and, thus, the random effects model should be rejected.

We highlighted a number of problems relating to the panel data estimation techniques, one of which we have to deal with immediately. The first specification of the demand equation (equation 4.17) includes time invariant factors (catchment demand and catchment share). Since time invariant independent variables cannot be included in the fixed effects approach, it is not possible to estimate the first specification of the model including hospital specific fixed effects. Consequently, if we are to estimate a fixed effects model we have to group providers in some other way. However, in grouping the data the implication is that there is some systematic effect in operation between the groups that leads to parametric shifts in the regression function. Previous authors faced with similar problems have suggested grouping hospitals by geographic location (Gaynor and Anderson, 1995), others have suggested grouping by bed size might be appropriate (Carey, 1996). Essentially any grouping of this kind is arbitrary unless we have strong a priori beliefs about why such groupings should systematically affect demand. For example, we could argue that the larger the demand faced by hospital the more emergencies hospitals will treat. We may hypothesise that there may be a size effect, where individuals are more likely to go to larger hospitals in an emergency, (possibly due to quality issues or expectations regarding bed availability). We may expect that the type of hospital (teaching, children's, etc) may influence the number of emergency admissions.

In order to cover these issues we will group the hospitals using three methods[6]:

i) Group by site type.

ii) Group by demand size.

iii) Group by bed size (using two groupings).

Since we have time invariant factors in the specifications of the demand equation above, it was also worth considering estimating demand using provider specific effects and time period effects alone. To this end we estimated a simple demand equation using only group and period effects.

In addition to the group fixed effects we will also consider two way fixed and random effects model, i.e. including time effects which we will model as monthly effects.

Section 3.5.3: Demand equation results

The following equations were all estimated using LIMDEP PANEL data estimation techniques using one-way and two-way fixed and random effects models. The one-way models allow group effects to be incorporated into the estimation process; two-way models allow group and time period effects to be included. When estimating two way fixed and random effects models using LIMDEP, as part of the output, the package produces one-way and OLS regressions, thus, allowing a full comparison of the different estimation techniques. LIMDEP also reports the LM and Hausman test statistics thus allowing direct comparison of the random and fixed effects models.

[6] See Appendix A1.1 for group sizes.

The LM and Hausman tests statistics are reported for all the estimated equations as well as the log-likelihood values, the autocorrelation coefficient (which is [1-1/2DW] or Durbin's h-test where the DW is not applicable) and a test for heteroskedasticity (White's test, see White, 1980). The LM test tests the general hypothesis of OLS versus individual effects (i.e. random or fixed), where a significant value favours the individual effects models over OLS. The Hausman tests for random versus fixed effects, where a significant value favours the fixed effects model over the random effects model. For both these statistics the significance level is reported in square parentheses [].

The results for the four grouping approaches are summarised below for the model described in equation 4.17. The actual groupings and group sizes in each of the following regressions are set out in the four tables in Tables A1.1.1-A1.1.4 in Appendix A1.1. The results below will present summary tables, the fixed effects (period and group), coefficients will not be presented here but are presented in Tables A1.2.1-A1.2.7 in Appendix A1.2.

Table 4.2: Demand estimation: grouping by site type

	OLS	Random Effects (One way)	Fixed effects (One way)	Random Effects (Two way)	Fixed effects (Two way)
Variable	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
CATCHD	1.6043** (0.15136)	1.5541** (0.15315)	1.557** (0.1537)	1.5540** (0.1523)	1.5581** (0.15291)
CATCHSH	1235.3** (110.07)	1023.1** (116.26)	1016.9** (116.53)	1023.6** (115.67)	1017.7** (115.94)
BEDS	1.7243** (0.06427)	1.8557** (0.07591)	1.8563** (0.07721)	1.8561** (0.0755)	1.8555** (0.07682)
FILT1993	-173.77** (34.802)	-159.55** (34.559)	-159.22** (34.567)	-159.52** (34.383)	-159.22** (34.391)
FILT1994	-44.833 (29.346)	-47.807 (29.168)	-47.386 (29.182)	-47.819 (29.02)	-47.382 (29.033)
CONSTANT	-878.71** (112.97)	-887.74** (147.01)	--	-892.33** (152.56)	-792.22** (113.26)
R^2	0.64439	0.626787	0.65322	0.62577	0.65927
Log-Likelihood	-11539.048	--	-11520.031	--	-11506.233
LM	--	29.19 [0.00]	--	37.10 [0.00]	--
Hausman	--	--	1.56 [1.00]	--	1.40 [0.924]
Autocorr.	0.88540	0.140485	0.14091	0.143024	0.143370
Hetero.	0.00	22.80	0.00	22.78	0.00

* significant at 10% level

** significant at 5% level

Grouping by site type (Table 4.2) appears to add little information to the demand estimation process. The magnitudes of the coefficients differ little when estimating using random or fixed effects models. The LM test suggests that the individual effects models should be favoured over the OLS estimates. The Hausman tests suggest that neither of the fixed effects models should be favoured over the random effects model.

Table 4.3: Demand estimation: grouping by demand size

Variable	OLS	Random Effects (One way)	Fixed effects (One way)	Random Effects (Two way)	Fixed effects (Two way)
	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
CATCHD	1.6043** (0.15136)	-0.18707 (0.22703)	-0.47397* (0.23731)	-0.24194 (0.22751)	-0.47343* (0.23572)
CATCHSH	1235.3** (110.07)	-14.509 (162.21)	-209.60 (169.02)	-51.834 (162.45)	-209.22 (167.88)
BEDS	1.7243** (0.06427)	1.4376** (0.0660)	1.4011** (0.0665)	1.4299** (0.06567)	1.4004** (0.06615)
FILT1993	-173.77** (34.802)	-184.92** (32.446)	-186.09** (32.484)	-185.18** (32.235)	-186.09** (32.266)
FILT1994	-44.833 (29.346)	-49.219 (27.025)	-49.054 (27.049)	-49.212 (26.848)	-49.052 (26.867)
CONSTANT	-878.71** (112.97)	1082.5** (222.99)	--	1140.0** (228.85)	1327.5** (216.02)
R^2	0.64439	0.5178	0.70618	0.5102	0.71225
Log-Likelihood	-11539.048	--	-- 11394.7432	--	-- 11378.4526
LM	--	817.60 [0.00]	--	825.52 [0.00]	--
Hausman	--	--	19.53 [0.0015]		16.03 [0.0068]
Autocorr.	0.88540	0.178905	0.183962	0.184372	0.188542
Hetero.	0.00	395.63**	0.00	418.16**	0.00

* significant at 10% level
** significant at 5% level

Grouping by demand size (Table 4.3) provides mixed results. The LM test suggests that the individual effects model should be favoured over the OLS estimates. The Hausman tests suggest that both the fixed effects models should be favoured over the random effects model. However, the magnitude and significance of the coefficients on the independent variables cause some concern. Particularly the counter-intuitive negative signs on both catchment demand and catchment share. This is probably explained by the impact of the grouping effects. These will be interrelated with the independent variables since the groupings are based on these variables. There are also problems of heterogeneity in this model.

Table 4.4: Demand estimation: grouping by bed size (ten groups – see Appendix A1.2.1 for four groups results)

Variable	OLS	Random Effects (One way)	Fixed effects (One way)	Random Effects (Two way)	Fixed effects (Two way)
	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
CATCHD	1.6043** (0.15136)	1.501** (0.15479)	1.4722** (0.15625)	1.4923** (0.15446)	1.4727** (0.15526)
CATCHSH	1235.3** (110.07)	1049.8** (109.69)	1018.9** (110.37)	1038.1** (109.28)	1019.2** (109.67)
BEDS	1.7243** (0.06427)	1.5870** (0.15307)	1.2158** (0.21795)	1.4580** (0.17704)	1.2122** (0.21659)
FILT1993	-173.77** (34.802)	-189.86** (33.164)	-191.61** (33.204)	-190.67** (32.975)	-191.61** (32.995)
FILT1994	-44.833 (29.346)	-102.47** (28.078)	-111.71** (28.255)	-106.05** (27.965)	-111.77** (28.077)
CONSTANT	-878.71** (112.97)	-588.48** (178.71)		-471.12** (209.27)	-206.38 (219.07)
R^2	0.64439	0.63680	0.69212	0.62867	0.69823
Log-Likelihood	-11539.048		-11430.086		-11414.425
LM	--	1196.84 [0.00]		1204.75 [0.00]	
Hausman	--		10.85 [0.054447]		6.33 [0.27527]
Autocorr.	0.88540	0.026777	0.025238	0.027523	0.026588
Hetero.	0.00	0.00	26.13*	0.00	73.50**

* significant at 10% level
** significant at 5% level

Grouping by bed size (Table 4.4) appears to provide some evidence for fixed effects. The LM test suggests, one more, that the individual effects model should be favoured over the OLS estimates and the Hausman tests suggests, at the 10% significance level, that the one way fixed effects model, should be favoured over the random effects model. However the test rejects the two way fixed effects model. Furthermore grouping by bed size appears to reduce the, albeit small, problem of autocorrelation, although there does appear to be potential problems of heteroskedasticity associated with the fixed effects models.

The above approaches required what was essentially an arbitrary grouping of the provider units. Perhaps the best way of grouping would by provider unit; this would allow us to pick up the provider specific effects. However, as mentioned earlier, this requires time invariant factors to be dropped. This required both the catchment variables to be dropped, which might be expected to significantly weaken the explanatory power of the equation. However, since these variables were essentially representing provider specific effects, this may not necessarily reduce the predictive power of the equation, and may in fact improve it since the provider specific dummies will be allowed to pick up all the individual effects. A summary of the results of this approach are presented in Table 4.5 below (full results are presented in Table A1.2.5 in Appendix A1.2).

Table 4.5: Demand estimation: grouping by provider type

Variable	OLS	Random Effects (One way)	Fixed effects (One way)	Random Effects (Two way)	Fixed effects (Two way)
	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
BEDS	2.2195** (0.04539)	0.80694** (0.10389)	0.19766 (0.12229)	0.75089** (0.10039)	0.18543 (0.11645)
FILT1993	-168.12** (36.185)	-110.28** (17.063)	-118.34** (17.119)	-111.01** (16.243)	-118.50** (16.292)
FILT1994	-58.867* (30.477)	-29.704** (13.866)	-35.761** (13.908)	-30.272** (13.200)	-35.897** (13.236)
CONSTANT	310.42** (41.854)	1477.8** (103.98)		1524.1** (104.69)	1969.8** (96.988)
R^2	0.61440	0.36475	0.93923	0.34465	0.94538
Log-Likelihood	-11600.265		-10203.4202		-10122.249
LM		13602.35 [0.00]		13608.01 [0.00]	
Hausman			89.86 [0.00]		92.55 [0.00]
Autocorr.	0.89410	0.557103	0.548860	0.665013	0.660235
Hetero.	0.00	563.73**	0.00	616.83**	0.00

* significant at 10% level
** significant at 5% level

Grouping by provider type (Table 4.5) appears to provide the best estimation of demand in many ways. The LM tests provide strong evidence in favour of the groupings. The Hausman test indicates that the fixed effects models are both favoured over the random effects, and this provides strong evidence for systematic monthly effects, which is in line with a priori expectations. The R^2 in this model is also very high (0.94). Furthermore, the inclusion of fixed effects overcomes the problems of heterogeneity exhibited in the random effects model.

There does, however, appear to be a problem of autocorrelation in this model. This may be due to omitted variables, since we have had to drop two of the main explanatory variables. This will lead to potentially inefficient but unbiased results and will inflate the goodness of fit.

The next two models employ the two different approaches set out in section 3.5.1. The first model, based on equation 4.18, employs a simple naïve forecasting model, which assumes that hospitals base future forecasts on previous demand. As above, provider specific dummy variables were included to pick up any provider specific effects. Since the aim of estimating these equations is also to replicate the approaches used by other authors, we will not subject these equations to the LM and Hausman tests. (Although it is clear from the results of Table 4.5 above that there is strong evidence to support individual provider specific fixed effects).

The lagged dependent variable approach is more easily estimated using SPSS as the first observations for each provider unit had to be dropped. The full results of this approach are also presented in Table A1.2.6 in Appendix A1.2.

Table 4.6: Demand estimation: lagged dependent variable

Variable	Lagged dependent variable
	Coeff. (SE)
FILT1993	-32.279** (12.000)
FILT1994	-4.049 (9.348)
LAGEM	0.703** (0.018)
CONSTANT	320.975** (32.541)
R^2	0.97348
Durbin's h-test	34.42**
Hetero.	199.00**

* significant at 10% level
** significant at 5% level

The results indicate that including a lagged dependent variable improves the fit of the equation. This gives an R^2 of 0.97, which is of a similar magnitude to that found by other authors who have employed this technique. The main problem with the lagged dependent variable approach is that it almost inevitably induces autocorrelation. Since we had previously encountered problems of autocorrelation using a similar approach, it is not surprising that the inclusion of a lagged dependent variable does not ease these problems (Durbin's h-test strongly indicates the presence of autocorrelation in this equation)[7].

When autocorrelation is present the least squares estimators are usually unbiased but inefficient, although the sampling variances will be biased and sometimes seriously

[7] The DW test in invalid when a lagged dependent variable is included as a regressor. Durbin's h-test provides an alternative in such circumstances.

underestimated. Therefore, the R^2, as well as the t-statistics, tend to be exaggerated. When lagged dependent variables are included in the equation generally least squares estimators are consistent but biased. However, if serial correlation is present in a lagged dependent variable model then the least squares estimates become inconsistent and biased. Therefore, the problems are potentially serious in a forecasting model. Furthermore, there appears to be problems of heteroskedasticity in this model.

The final approach, based on equation 4.19 adopts a simple autoregressive (AR1) process to estimate demand. Once more this model assumes that the hospital bases its forecasts on demand realisation in the previous period. The AR1 process has dominated the empirical literature and can provide a reasonably good model where the underlying processes are often very complex. It is an estimation process that explicitly allows for autocorrelation. However, it does rely on the premise that the entire correlation structure can be explained by a single adjustment.

Equation 4.19 was estimated using LIMDEP and employed the AR1 procedure within this statistical package. Estimation is done in two steps, the first estimates the equation ignoring the autocorrelation to obtain an estimate of ρ. The second stage employs a generalised least squares procedure. The summary results are presented in Table 4.7 below (see Table A.1.2.7 in Appendix A1.2 for full results).

Table 4.7: Demand estimation: AR(1)

Variable	AR(1)
	Coeff. (SE)
FILT1993	-107.33** (34.540)
FILT1994	-28.613 (21.567)
RHO	0.660691
R^2	0.82276
Log-Likelihood	-9176.5572
Autocorr.	-0.080802
White χ^2(78 dof)	86.18

* significant at 10% level
** significant at 5% level

The equation performs quite well with a relatively high R^2 and appears to deal with the problem of autocorrelation. The White test is insignificant (with 78 dof) indicating that heteroskedasticity is not a problem. The value of ρ, 0.66, is similar to the average value calculated by Friedman and Pauly (1983) of 0.62.

Overall the three approaches give a range of forecasts with goodness of fit ranging from 0.64 to 0.97, although we have noted that the existence of autocorrelation in some of the models may inflate the R^2s. Thus the model with the worst fit still represents a reasonably good forecasting tool. Prediction errors from the three approaches were around 16% for models based on equation 4.17, 4% for the model based on equation 4.18 and 6% for the model based on equation 4.19 (see Table A1.3.1 in Appendix A1.3 for full results of the prediction errors and a comparison of the three estimation techniques). Friedman and Pauly found that, on average, their estimates suggested estimation error of around 2 to 3%, but the data were more aggregate (i.e. quarterly). Pauly and Wilson

(1986), using actual hospital forecasts returned to Blue Cross and Blue Shield, found a variation of 4%.

The choice of demand model is in many ways arbitrary, since we are concerned with trying to proxy the hospitals own forecasting abilities and techniques. One criterion we could have used was the goodness of fit. However, since one of the problems of autocorrelation is that the R^2 tends to be exaggerated, this is ruled out.

It could be argued that the simplest model best represents hospital forecasts, as hospitals are not likely to develop sophisticated forecasting techniques, this would favour a simple OLS model. Alternatively, it could be argued that we should use the statistically most rigorous model, as the others are likely to incorporate problems in estimating the error and variance accurately, and since this is one of the main reasons for our estimation we should follow this line. As noted earlier, with any forecasting equation there is a trade-off between statistical rigour and forecasting ability.

The lagged dependent variable approach, although intuitively appealing, has statistical problems that may limit it usefulness; in particular it will underestimate the standard errors of the forecasted equation (one of the main reasons for our estimation). The panel data models provide a reasonable estimate of demand. In terms of intuitive appeal, however, they may be more limited since they rely on fixed estimates of demand and fail to incorporate any systematic time effects that are indicated strongly by the other two models. Furthermore, the other models appear to suffer from problems of heteroskedasticity, this may indicate that some of the group effects are being missed, or, more generally, that there are problems of omitted variables.

Therefore, perhaps the best estimate of demand is found using the simple AR(1) model that allow complex relationships to be modelled using a relatively simple model. This model incorporates previous demand realisations and also includes the fixed effects associated with provider specific influences and therefore combines statistical rigour with intuitive appeal. Consequently, when estimating the cost function we will use the estimates of demand taken from the AR(1) estimates based on equation 4.19.

3.6: Cost function estimation

The estimation of the cost function will employ the same data set identified in section 3.4 and listed in Table 4.1. The activity data will be aggregated to fit in with the annual data (see Appendix 3). Since panel data are used it is necessary to investigate the data for evidence of fixed or random effects. One of the problems with the analysis is that in order to determine the appropriate grouping and test for the appropriate functional form requires the equation to be specified in advance. In order to specify the appropriate grouping we must specify the functional form, but in order to investigate the functional form we must first specify the appropriate grouping. In order to break into the analysis we must therefore choice the starting point. We will first consider the same groupings as considered for the demand equations (i.e. site type, demand size and bed numbers) on the basis of a linear specification. We will then choose the appropriate grouping, if any, on the basis of the LM and Hausman tests, as earlier.

The results of the fixed and random effects models are presented in Tables A1.4.1-A1.4.4 in Appendix A1.4. The results indicate that grouping the data by bed numbers, demand size, or site type has little statistical justification. That is the LM test indicates that the OLS model performs best. Consequently, all further analysis was undertaken without group dummy variables (the only dummy variable included is for teaching hospitals as this was significant in the OLS estimate). The general form of the cost function will be as set out below:

$$TVC = f(ELADM, EMADM, RES, BEDSEL, BEDSEM, CASEMIX, INVOCC, OPV, DAYATT, AEATT, WI, DVTEACH) \quad (4.26)$$

As there is no theoretically accepted 'true' functional form for hospital cost functions we will adopt two alternative approaches to estimation. First we will attempt a transcendental logarithmic (translog) function. This allows considerable flexibility in estimation, allowing for interactions between variables, and is based on a second series Taylor expansion, such that the most general form of the translog function with multiple outputs is represented as:

$$LnC = \alpha + \beta_i lnX_i + \Sigma_i\theta_i lnp_i + 1/2\Sigma_i\delta_i(lnX_i)^2 + 1/2\Sigma_i\Sigma_j\delta_{ij}lnX_ilnX_j + 1/2\Sigma_i\Sigma_j\gamma_{ij}lnp_ilnp_j + \Sigma_i\phi_i lnp_ilnX_i + \varepsilon \quad (4.27)$$

Where C is cost, p is input prices, and X is a vector of outputs. The function is assumed to be a representation of the minimum cost function. It has to be positive and homogeneous of degree one in prices.

However, the increased flexibility that the translog function allows comes at a cost, not least with respect to the number of independent variables, with consequent loss in degrees of freedom and potential problems of multicollinearity. Including seven output variables[8] would require 65 parameter values to be estimated. In order to keep the number of parameters down to a reasonable size we imposed restrictions on the number of interactive terms we will estimate the following form[9]:

$$\begin{aligned} LNTVC = {} & \alpha + \beta_1 LNEMADM + 1/2\beta_2 LNEMADM^2 + \beta_3 LNELADM + 1/2\beta_4 LNELADM^2 \\ & + \beta_5 LNBEDEM + 1/2\beta_6 LNBEDEM^2 + \beta_7 LNOPATT + \beta_8 LNOPATT^2 + \beta_9 WI \\ & + 1/2\beta_{10} WI^2 + \beta_{11} LNELADM.LNEMADM + \beta_{12} LNELADM.LNBEDEM \\ & + \beta_{13} LNEMADM.LNBEDEM + \beta_{14} LNELADM.LNOPATT \\ & + \beta_{15} LNEMADM.LNOPATT + \beta_{16} LNELADM.LNWI + \beta_{17} LNEMADM.LNWI \\ & + \beta_{18} LNDAYATT + \beta_{19} LNAEATT + \beta_{20} LNCASEMX + \beta_{21} LNRES \\ & + \beta_{22} LNIVOCC + \beta_{23} DVTEACH + \varepsilon \end{aligned} \tag{4.28}$$

where the variables are defined as earlier in section 3.2.

---

[8] These are ELADM, EMADM, BEDSEL, BEDSEM, OPV, DAYATT, AEATT.

[9] Preliminary analysis revealed that elective admissions and elective beds are highly correlated (correlation coefficient = 0.94), which is to be expected since BEDSEL = ELADM*LOS/365, therefore the analysis uses elective admissions only to pick up the cost of an elective admission. Since the theoretical model assumes that elective beds are all occupied this does not cause problems of interpretation, as the elective admission variable will pick up both the variable and quasi-fixed element of cost.

The form of this equation can be simplified further by imposing restrictions on the higher order terms such that the restricted equation becomes:

$$LnC = \alpha + \beta_i lnX_i + \theta lnp_i + \varepsilon \qquad (4.29)$$

which is the Cobb-Douglas form[10].

As an alternative specification we will perform a Box-Cox transformation. This approach has been suggested by Greene (1993) as an appropriate alternative to the translog when variable values equal zero and/or when the functional form of the equation is unknown, and allows flexibility in the estimation process including linear and log-linear models as special cases of the transform.

The Box-Cox model has appeared in a number of recent studies and is based on the transformation:

$$g^{(\lambda)}(x) = (x^{(\lambda)}-1)/\lambda \qquad (4.30)$$

The Box-Cox model is a useful formulation that embodies many models, and is particularly useful if the functional form is unknown, or we have no a priori information to guide us. The most general form of the Box-Cox transformation allows the dependent and independent variables to be transformed by different values such that:

$$y^{(\theta)} = \alpha + \Sigma\beta_k x_k^{(\lambda)} + \varepsilon \qquad (4.31)$$

Where the linear model and log-linear models are special cases, which results if $\lambda = \theta = 1$ and $\lambda = \theta = 0$, respectively.

3.7: Cost function results

The results of the translog model are poor with counterintuitive signs on the coefficients; most notably, a negative sign on the coefficient on emergency admissions variable, and insignificant t-statistics on almost all the independent variables. (The results of the translog approach are presented in Table A1.4.5 in Appendix A1.4). These problems are not unique; others have reported similar findings, (see, for example, Vita, 1990, and Dor and Farley, 1996). The relatively high R^2 and lack of significant coefficients on the independent variables might indicate problems of multicollinearity. Further investigation revealed the probable existence of multicollinearity with 12 of the correlation coefficients higher than 0.9. Since the focus of our modelling is to elicit marginal cost estimates the translog function would be too unreliable. The collinearity diagnostic tests, including eigen values of condition indices, are presented in Tables A1.4.6 and Table A1.4.7 in Appendix A1.4.

---

[10] Other restrictions could be imposed that would lead to other specifications.

To examine the model specification we can test the unrestricted translog model (equation 4.27) against the restricted, Cobb-Douglas, model (equation 4.29) by undertaking an F-test to test whether the higher order terms are jointly significant, i.e. whether $\delta_{ij} = \gamma_{ij} = \phi_i = 0$ in equation 4.27. The F-value is $((RSS_R-RSS_U)/q)/(RSS_U/(n-k))$, where RSS_R is the residual sum of squares in the restricted model, RSS_U is the residual sum of squares in the unrestricted model, q is the number of restrictions, n is the number of observations and k is the number of regressors in the unrestricted model. The F-test statistic was 0.94, therefore, accepting the restrictions placed on the higher order terms and choosing the Cobb-Douglas in favour of the translog form. However the Cobb-Douglas model performed equally badly in terms of counter-intuitive signs on the independent variables, in particular the negative signs on day attendances and emergency admissions, and lack of significance of coefficients (only four of the eleven independent variables are significant – see Table A1.4.8 in Appendix A1.4).

The Box-Cox model was therefore employed as an alternative. We adopted the most general estimation procedure possible allowing the most flexibility of functional form and imposing as few restrictions as possible. This involved allowing the right hand side and left-hand side variables to be transformed by different parameter values. The transformation parameters were chosen on purely statistical grounds.

If we assume that λ and θ are unknown parameters, then the choice of transform values can be identified by employing a grid search. Since the least squares values of λ and θ are usually found between –2 and 2 typically the search takes place within these values. It is possible, using LIMDEP, to specify a MLE technique that allows the statistical package to identify the optimal values of the two parameters, however the technique used is sensitive to the starting values specified. To overcome any potential problems we employed a manual grid search using increments of 0.25 over the range –2 to 2 for λ and θ to identify the appropriate starting values for the MLE search. The estimates were then fine-tuned these using the MLE technique within LIMDEP.

The full log-likelihood values from the grid search employing different values of lambda and theta are presented in Table A1.4.9 in Appendix A1.4 (where lambda is the transform applied to the right-hand side variables and theta the transform applied to the left-hand side variable). From Table A1.4.9 it is clear that the maximum log-likelihood lies somewhere between θ(0.25-0.5), λ(1.0-1.50). This suggests that the starting values of θ = 0.20 λ = 0.95 should enable us to locate the minimum using MLE. The optimal values of λ and θ were identified as 1.2112 and 0.53812, where the log-likelihood value is –325.71[11].

As outlined above, the linear and log-linear models represent special cases of the Box-Cox transform and as such it is possible to test for these two restrictions, i.e. that λ and θ both equal either 0 or 1. To test these restrictions we employ the likelihood ratio test, which is a Chi squared test where the test statistic is:

[11] To double-check these values for optimality we also used different starting values for λ and θ and this verified these values.

$$-2[Ln(R) - Ln(max)] \qquad (4.32)$$

where Ln(R) is the likelihood evaluated at the restricted estimate and Ln(max) is the optimal value of the likelihood. This test therefore is similar to the F-test. The test statistic was constructed using the values from the grid search:

Log-linear (0,0) = -336.94

Linear (1,1) = -335.40

Max (1.2112, 0.53812) = -325.71

The log-likelihood test with 95 observations is Chi squared (1) with 1 degree of freedom. The critical value is 3.84. The two test statistics are:

Log-linear = -2[-336.94 – (-325.71)] = 22.46**

Linear = -2[-335.40 – (-325.71)] = 19.38**

Therefore both the linear and log-linear models can be rejected at the 5% level in favour of the optimal transformed model.

At least two issues present themselves in the cost function estimation that are worthy of further investigation; heteroskedasticity and endogeneity.

We hypothesised that two variables were particularly susceptible to endogeneity; the elective admissions and the inverse of the occupancy rate. A Hausman (1978) test was performed on the Box-Cox transformed equation, indicating that both could be treated as exogenous[12]. The instruments used to test for endogeneity were catchment demand, catchment share, waiting list size, and average waiting time as defined in Table 4.1.

It is often suspected that the error terms in hospital cost functions are heteroskedastic, presumably through the effect of increasing hospital size. We tested for heteroskedasticity using the White (1980) test, and found that the test rejected the presence of heteroskedasticity at the 5% level.

The results of the transformed model are presented below (and in full in Table A1.4.10 in Appendix A1.4. (The results of the linear and log-linear models are presented in Tables A1.4.12 in Appendix A1.4 for comparison). The standard errors presented in the model assume that the values of lambda and theta are fixed, that is they ignore the variation in the transformation parameters, this provides t-statistics which are comparable with those produced by OLS.

---

[12] We did, in fact, run a regression treating electives as endogenous using an instrumental variables approach. In the event the results changed very little (see results in Table A1.4.13 in Appendix A1.4).

Table 4.8 Cost function: Box-Cox Non-linear Regression Model

Variable	Coefficient	Standard Error	t-stat
EMADM	0.17291E-05	0.47135E-05	0.367
ELADM	7.1242	2.8144	2.531**
BEDSEM	0.15571E-02	0.37484E-03	4.154**
CASEMIX	0.28872E-04	0.11057E-04	2.611**
INVOCC	-0.39541	0.29344	-1.348
RES	0.88775E-02	0.49121E-02	1.807*
AEATT	0.71681E-05	0.20272E-05	3.536**
DAYATT	0.86015E-05	0.32872E-05	2.617**
OPATT	0.26963E-02	0.10556E-02	2.554**
WAGE INDEX	0.41711E-02	0.50607E-01	0.082
DVTEACH	5.7278	0.90725	6.313**
CONSTANT	9.9175	1.6035	6.185**
LAMBDA	1.2112		
THETA	0.53812		

Number of observations	85
Log-likelihood	-325.71
F[12, 82]	3131.17**
White test for heteroskedasticity (χ^2):	0.43
Hausman test for endogeneity (t-test):	
ELADM	0.683
INVOCC	0.539
ELADM & INVOCC (F-test)	0.246

* significant at 10% level
** significant at 5% level

The coefficient on RES is both positive and significant as hypothesised. Thus, supporting the hypothesis that uncertain demand has an impact on hospital cost, and the higher the extent of the uncertainty the higher costs will be. The signs on all the other coefficients are as expected a priori. There are, however, three coefficients that are insignificant, those on INVOCC, WAGE INDEX, and EMADM. The insignificant signs on the first two variables may be explicable. The impact of reserve capacity may be picked up by the inclusion of emergency beds, some of which will remain empty, thus picking up some of the effects of unused capacity. However, since INVOCC picks up the impact of operating off the PPF, this suggests that the effect on the hospital costs due to operating within the PPF may be insignificant at standard levels.[13] This is consistent with the theory developed in section 2 above, as we suggested that hospitals will in fact attempt to minimise ex ante expected costs and operate on the shadow PPF rather than the PPF itself. The insignificance of the coefficient on the wage level may be due to the existence of national pay settlements in the NHS resulting in little variance across provider units. The insignificant coefficient on emergency admissions is, however, is a little surprising, although it is appears to be of a reasonable magnitude (see the marginal cost estimates in Table 4.9 below).

The coefficients on the variables in the results presented above are in themselves uninformative, since as in a non-linear model they will not represent the slopes with respect to the variables. (Although given the positive values of lambda and theta the signs of these coefficients can be interpreted the same way as in a linear model). In order to interpret the magnitudes of these coefficients, we need to calculate the implied marginal costs using the elasticities based on these coefficients. The elasticity is calculated at the mean values of the independent variables using the predicted value of total variable cost from the estimated equation (see Greene 1993), where the elasticity is:

[13] A quadratic term for inverse of occupancy rate was also included but this was also insignificant. The quadratic term was not included in the final specification as it added little to the analysis and reduced the significance of the linear term.

$$\delta TVC/\delta X_i * X_i/TVC \qquad (4.33)$$

where, X_i= independent variable and where:

$$\delta TVC/\delta X_i = [[\alpha\theta + 1 + \Sigma_i \beta_i \theta (Xi^{(\lambda-1)})/\lambda]^{(1/\theta-1)}]\beta Xi^{(\lambda-1)} \qquad (4.34)$$

The calculated elasticities based on the above coefficients are given in Table 4.9 below:

Table 4.9: Marginal cost estimates

Variable	Coeff.	Mean	Elasticity	Marginal cost
EMADM	0.17291E-05	25,026	0.0381	£110.97
CASEMIX	7.1242	0.6937	0.4726	£49,660,433.30
BEDSEM	0.15571E-02	645	0.4084	£45,154.55
RES	0.28872E-04	1,536	0.0217	£1,029.81
INVOCC	-0.39541	1.49	-0.0797	-£3,358,154.71
AEATT*	0.88775E-05	59,137	0.1290	£159.01
DAYATT	0.71681E-02	11,111	0.0591	£387.72
OPATT*	0.26963E-02	190,243	0.1612	£61.77
WAGE INDEX*	0.41711E-02	19,061	0.0154	£58.89
ELADM	0.86015E-05	24,672	0.1863	£550.42
TVC		72,893,450		

All calculations are re-adjusted by scaling factor for TVC of 1,000,000
* re-adjusted for further scaling factor of 1,000

The impact of unexpected demand translates into an average cost of £63 per emergency admission. The marginal costs of elective and emergency admission are presented in Table 4.10 below.

Table 4.10: Marginal cost estimates of elective and emergency admissions

Variable	Mean	Elasticity	Marginal cost
EMADM	25026	0.0381	£111
BEDSEM	645	0.4084	£760
ELADM	24672	0.1863	£550

Therefore the total marginal cost of an emergency case is estimated at £871 (i.e. the sum of the variable cost element, EMADM, and the quasi-fixed element, BEDSEM) and an elective case is estimated to be £550.[14] The marginal cost of a bed is estimated by calculating the cost per bed day and multiplying by the average length of stay of an emergency admission.[15]

Section 3.8: Sensitivity analysis

3.8.1: Uncertainty

Let us we now consider the impact of dropping RES and INVOCC and re-estimate marginal costs without taking into account uncertainty. The results are presented in Table A1.4.14 in Appendix A1.4. The calculated elasticities after dropping RESTOT and INVOCC are presented in the Table 4.11 below.

Table 4.11: Marginal cost estimates (dropping RES and INVOCC)

Variable	Mean	Elasticity 1	MC[1]	Elasticity 2	MC[2]
EMADM	25026	0.0882	£257	0.1357	£395
BEDSEM	645	0.3707	£691	0.3278	£611
ELADM	24672	0.2077	£614	0.2149	£635

1 = dropping RESTOT only
2 = dropping RESTOT and INVOCC

[14] These estimates are of a similar magnitude as those reported by Csaba (1996) who reported the cost of an inpatient episode as £585 (representing an aggregate of emergency and elective cases), although these estimates did not take account of uncertainty.

[15] This was the weighted average length of stay of emergency and other admission, which is 6.02.

These results appear to indicate that, if we do not take account of uncertainty, then we will overestimate the variable costs of emergency cases, and underestimate the costs of bed days for emergency cases. This is in line with a priori expectations, since estimating costs without taking account of uncertainty will fail to recognise that hospitals will aim to operate within the PPF efficiency locus. The total impact on the marginal costs of emergencies leads to an overestimate of around £100, or about 10%. The estimates of the marginal cost of electives appear to be overestimated by a similar amount.

3.8.2: Demand uncertainty variable

Section 3.5.2 discussed a number of different variables that could be used to model the impact of uncertainty. The two main variables suggested were the total unexpected demand and the standard error of the total unexpected demand. Replacing total unexpected (RES) with the standard error of this variable (SERES), appears to provide an equally good estimate, (the results are presented in Table A1.4.16 in Appendix A1.4). The estimated coefficients appear to be stable in sign, magnitude and significance. The estimates of marginal costs using standard error in place of total unexpected demand, gives estimates of £898 and £647 for emergency and elective cases respectively. The estimate of the marginal cost of emergency cases, therefore, remains fairly stable, although the estimate of elective marginal costs is higher than when employing the total residual.

Section 4: Conclusions:

In this chapter we have presented an empirical estimate of a total variable cost function consistent with the theoretical model developed in Chapter 3, where costs reflect and incorporate production responses to demand uncertainty. To do so, we have considered the forecasting error that hospitals make in seeking to predict demand. Because it is not obvious how hospitals in reality estimate demand, a number of different estimation procedures were considered. The chosen method employed a simple autoregressive process, as this seemed to perform as well as the others, and was both intuitively appealing and statistically robust.

In specifying the total variable cost model we have deliberately allowed the data to specify the functional form, after ruling out a number of commonly used functional forms. The estimation was undertaken through a Box-Cox procedure, and the resultant total variable cost function gave robust parameter estimates, in particular, our a priori expectation that demand uncertainty was important was verified.

The main results reported in this chapter provide reliable estimates of the marginal costs of provision of hospital care, based on a cost function with sound theoretical underpinnings, taking account of demand uncertainty and output heterogeneity.

Chapter 5: Activity Analysis

Section 1: Introduction

In Chapter 3 we developed a theoretical model of capacity utilisation decisions, and focused on two perspectives; private and social. We showed that reserve capacity could be viewed as an efficient response to demand uncertainty from either perspective. In Chapter 4 we built on this model and focused on the costs of production. The optimal allocation of capacity was identified and, once more, we showed that reserve capacity was an efficient response for a cost minimising hospital where demand uncertainty was present. The cost function analysis in Chapter 4 focused on the private costs of production, as these are the only costs that are observed. However, in the initial specification of the problem we considered not only the private perspective but also the social costs of production. The optimal allocation of capacity, when the hospital takes a wider social perspective, was identified (taken from equation 4.6 in Chapter 4) as:

$$MSC_t[1\text{-}F(B_{em})]\text{-}\ MSC_w = MC_{Bem} + MC_{vem}[1\text{-}F(B_{em})]\text{-}\ MC_{Bel} - MC_{vel} \qquad (5.1)$$

where,

B_{em} = number of beds allocated to the emergency sector

MSC_t = marginal social cost of turning away emergency cases

MSC_w = marginal social cost of leaving patients on the waiting list

MC_{Bem} = marginal cost of a staffed emergency bed

MC_{Bel} = marginal cost of a staffed elective bed

MC_{vem} = marginal variable cost of emergency case

MC_{vel} = marginal variable cost of elective case

$[1-F(B_{em})]$ = probability of turning away an emergency case

Where the social cost was derived from two sources; the size of the waiting list for elective care and the number of emergency cases turned away. This optimal allocation condition suggests that hospitals trade-off the private costs of production associated with treating elective and emergency cases and the social costs of turning away emergency cases and placing elective cases on the waiting list. Furthermore, it suggests that hospitals also take account of uncertainty such that the social and private costs associated with the uncertain demand would be weighted by the probability of demand being present.

Seven elements determine the optimal allocation of capacity when considered from a social perspective. Four of these elements, from equation 5.1, are private costs (i.e. MC_{Bem}, MC_{vem}, MC_{Bel}, and MC_{vel}), two are social costs (MSC_t and MSC_w), and the remaining factor is the probability of demand being present $[1-F(B_{em})]$.

The cost function in Chapter 4 provides estimates of the first four elements. However, as noted in Chapter 4, the social costs are unobservable. Nonetheless, we hypothesise that if hospitals take account of the social costs attached to production that they will trade-off the size of hospital waiting lists against the number of patients turned away. Furthermore, from the data identified in Table 4.1 in Chapter 4, we have monthly data on waiting lists, bed numbers and emergency demand. Consequently, using these variables, and constructing a measure for the probability of turning away an emergency patient, we can directly infer the relationship between the waiting list size and turnaway rates at the margin. This will enable us to estimate the monthly allocation response of hospitals to exogenous changes in the size of the waiting lists and emergency demand.

Using this implied value approach it is possible to estimate the social costs of turning an emergency patient away and placing an elective patient on the waiting list. This approach relies on estimates of the three outstanding components of equation 5.1. These are; the social costs of turning emergency cases away (MSC_t), the social cost of placing elective cases on a waiting list (MSC_w), and the probability that the hospital will be full [$1-F(B_{em})$]. This will allow a full empirical specification of the model to be considered.

The aim of this chapter then is to provide the empirical estimates necessary to calculate the implied relative marginal social costs of turning patients away and the marginal social cost of placing patients on the waiting list. This allows us to examine the social perspective. The chapter will be structured as follows: Section 2 will provide estimates of the probability of the hospital being full. Section 3 will investigate the relationship between waiting lists and turnaway rate to provide the implied values attached to waiting lists and turnaway rates. This will bring together all the elements necessary to enable the full empirical specification of the first order condition to be specified, giving estimates of the implied social costs. Section 4 will consider two of the key issues identified focused on throughout this thesis; heterogeneity and demand uncertainty and will consider how they are crucial in accurately estimating the marginal social costs.

Section 2.1: Estimating the probability of turnaway

The probability of turning an emergency patient away is highlighted above as one of the determinants of the optimal allocation of hospital capacity. Hospitals will take account of uncertainty and will allocate capacity on the basis of private and social costs weighted by the probability of demand being present. However, it is not possible to directly observe the probability of turnaway. Therefore, it is necessary to construct an estimate of this probability using the information we have available. The available data provide all the necessary information we need to construct this variable. That is, emergency demand, elective demand, and bed numbers. However, in order to estimate the individual monthly, provider specific, probabilities we need to estimate demand expectations and the standard error associated with this demand, since these will determine the expected probability of turnaway. Before outlining the approach adopted and the results of this analysis it is worth considering the current, but small, literature on this specific issue.

2.2: Previous estimates of the probability of turnaway

The literature on probability of turnaway estimation is, unsurprisingly, sparse. To our knowledge, there have only been two previous attempts to empirically estimate the probability of turnaway in the literature[1]. The first, by Mulligan (1985), extended the work by Joskow (1980) on reserve capacity. Mulligan re-examined the determinants of reserve capacity using the same data to calculate exact hospital turnaway rates, rather than the notion of a target reserve capacity 'k' employed by Joskow (see Chapter 2).

Mulligan assumed a Poisson distribution for arrivals such that the standard deviation was equal to the square root of the mean average daily census. The results were sparsely presented in the paper. The actual values for the probability of turnaway were not reported, although he did report that 70% of hospital appeared to operate with a probability of turnaway of less than 0.1%. Furthermore, since he assumed a Poisson distribution no attempt was made to estimate the actual standard error of the sample. Therefore the approach represented a mixture of observation and simulation.

More recently, Carey (1996) undertook empirical work to directly estimate the probability of turnaway. This estimate was necessary to provide empirical content to her model of the optimal total bed capacity within a region (see Chapter 2). She estimated the average daily census (ADC) for total aggregate demand using a lagged dependent variable approach outlined in the previous chapter. Where ADC = (Demand x length of stay)/number of days.[2] These estimates were used to calculate the expected average daily census, the standard error of the ADC and, subsequently, the probability of turnaway. Detail on the exact method employed to calculate the probability of turnaway is lacking

[1] Shonick (1970) provided simulations that enabled probabilities of turnaway to be calculated but they where not based on observed data.

[2] The estimation of average daily census provides a way of converting aggregate demand and capacity, represented by bed numbers, into common currency.

in her paper, particularly regarding the distribution used and the exact calculations used to estimate the probabilities. It would appear, however, that she assumed a normal distribution for ADC, and used the estimates of standard error from the forecasted equation to calculate these values using an equation of the form:

$$1\text{-}F[(B - E(x))/\sigma] \tag{5.2}$$

where B is the number of beds, E(x) is expected ADC, σ is the standard error of the distribution and F[.] is the probability density function representing demand. Carey estimated the average probability of turnaway across the whole sample period as 0.0062.

There are, however, a number of problems with the approach adopted by Carey that are worth discussing here. First, Carey estimated ADC using annual data. This is likely to significantly under estimate the variation in demand, as smoothing occurs, since there are likely to be significant seasonal, monthly, and even daily fluctuations in admissions. Secondly, she did not separate demand into planned and unplanned. This may have two effects: It is likely to underestimate the level of reserve capacity held to treat the stochastic element of demand. It may also lead to an underestimation of demand fluctuations, as planned admissions are likely to be increased when expected emergency admissions are low and vice versa, thus reducing the observed fluctuations in admissions. Related to this, and perhaps most importantly, she estimated her probability of turnaway using the standard error of the actual demand. That is, instead of considering the standard error of the level of unexpected demand (i.e. actual – forecast), she considered the standard error of total forecasted demand as the appropriate variation to use when estimating the probability of turnaway. This is likely to seriously over estimate the variation in demand and, consequently, the probability of turnaway, although, given the dual effect identified above, the total effect on turnaway rates is indeterminate.

It is, however, difficult to comment extensively on Carey's figures as she does not present all the data; most notably the standard error values are absent. If, however, we assume that the standard error of the distribution is equal to the square root of the mean, (as Joskow and Mulligan), and use Carey's data, it is possible to calculate the standard errors using the ADC figures presented in her paper.

If we use equation 5.2 and assume that the probability density function (p.d.f.) can be represented by a normal distribution, (where the normal approximates a Poisson when the mean is large). Then we can calculate the probability of turnaway using the following equation:

$$E(Pt) = 1 - \int_0^B \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx] \qquad (5.3)$$

where B is the number of beds, σ is the standard error of ADC and μ is the mean of ADC.

It is now possible to re-examine Carey's figures. For example, taking the first of Carey's results[3], the probability of turnaway is reported to be 0.0055, based on the reported data given in Table 5.1 below

Table 5.1: Carey's Probability of turnaway

Number of beds	Occupancy rate	Probability of turnaway	Implied ADC	S.D. (√ADC)
179	0.58	0.0055	104	10.19

Carey only reports the first three values given in Table 5.1. However, it is possible from the first two values to calculate the implied ADC (i.e. ADC = Beds x Occupancy rate). Using the properties of the Poisson distribution, the estimated standard error is equal to √ADC. Re-estimating the probability of turnaway using these data and equation 5.3 gives an estimated value of 7.99×10^{-14}, compared to 0.0055 calculated by Carey. To replicate Carey's figures the standard error of the distribution would have to be around 30. This represents a variation around the mean of nearly 30%, (that is, a factor of ten higher than Friedman and Pauly found in their estimates of demand fluctuations using less aggregated data). Furthermore, this does not fit well with the high R^2 found in her predicted equation, which suggests that hospital can predict demand very well over the year.[4] This highlights the problem of using annual data. Carey must rely on fluctuations of demand in previous years to generate the standard error in any period and this implicitly assumes that the demand level could reach any of the levels previously

[3] The estimated probability of turnaway for 1987, presented in Table 2a in her paper (Carey, 1996).

[4] Although a rather crude relationship, the goodness of fit of the predicted equation will increase as the standard error of the unpredicted element of demand falls (i.e. the standard error of the noise component will fall as the R^2 increases).

observed, consequently the standard errors used by her are very high. At the same time, however, the predictions of annual demand perform very well, suggesting that within any one year period hospitals have a very good idea of total aggregate demand.

Building on this literature we intend to address two issue here. First we discuss the use of total standard error of the forecasted equation to represent unexpected fluctuations in demand. Secondly, we disaggregate this demand into monthly forecasts to calculate the estimates of probability of turnaway.

2.3: Estimating the standard error of ADC

To obtain estimates of the standard error of unexpected demand, that is fluctuations around the expected demand, we first need to re-estimate the demand equations in the Chapter 4, replacing total demand with ADC. It is reasonable to assume that the process determining demand is the same for ADC as for total demand[5], therefore we used the same demand function as Chapter 4 and estimated demand using the same three approaches and considered the same groupings (i.e. site type, demand size, and bed numbers).

As with the previous demand estimation, the AR1 process appeared to perform best in terms of econometric and statistical considerations, and this is the equation we will

[5] Since ADC is simply admissions adjusted for length of stay and the time period.

continue to use in subsequent calculations.[6] The results of the AR1 estimate of ADC are presented in Table A2.1 in Appendix A2.1 and will not be presented in detail here.

The estimated standard errors for these equations were calculated by taking actual admissions away from forecasted admissions and then calculating the standard of this residual (these are presented in Table A2.2 in Appendix A2.2). These estimated standard errors now allow us to calculate the probability of turning a patient away.

Section 2.4: Probability of turnaway

The estimated standard errors from the above equations allow us to calculate the probability of turning a patient away, given the allocations of beds to the elective and emergency sectors. If we make the same assumptions as in Chapters 3 and 4, i.e. that all elective admissions are planned and that there is an excess demand for elective beds such that all beds allocated to the elective sector are filled, then we can assume that the remaining beds are available for, or allocated to, emergency admissions.[7] On this basis we can calculate the number of beds available for emergency cases, by subtracting the number of beds filled by electives from the total number of available beds.

The probability of turning an emergency patient away is estimated by calculating the probability that the demand for hospital services exceeds capacity such that:

[6] This is not surprising given that ADC is a construct based on total emergency demand. The results of the other estimation techniques are not presented as they are very similar to the total demand equations.
[7] This assumption does not strictly require that the hospital actually allocates the beds to emergency and elective cases simply that unfilled beds are available for emergency cases should they be required.

$$1\text{-}F[(B_{em} - E(D_{em}))/\sigma] \geq 0 \quad (5.4)$$

where, B_{em} is the number of beds allocated to treat emergency cases, $E(D_{em})$ is expected ADC for emergency services, and σ is the standard error of the unexpected element of ADC.

In order to operationalise this equation we need to define the distribution. As indicated above, most commentators have assumed that hospital demand is likely to be distributed Poisson, (see Blumberg, 1961; Shonick, 1970; and Joskow, 1980), which describes a process where each event is independent and unrelated to previous events.[8] The normal distribution may be used as a proxy for the Poisson since it approximates a Poisson for large N (i.e. N > 30). The assumption of a normally distributed error residual allows us to operationalise the above equation, employing the standard normal distribution, such that the probability of turnaway becomes:

$$E(Pt) = 1 - \int_0^{B_{em}} \frac{1}{\sigma_{em}\sqrt{2\pi}} e^{-\frac{(x-\mu_{em})^2}{2\sigma_{em}^2}} dx] \quad (5.5)$$

where B_{em} is the total number of beds allocated to treat emergencies, μ_{em} and σ_{em} are the mean of the expected ADC and standard error of the unexpected emergency demand, respectively.

Using this equation and the standard errors and means from the estimates of ADC, it is now possible to calculate provider specific estimates of the probability of turning an emergency patient away.

2.5: Empirical estimates of the probability of turnaway

The average probability of turning away a patient, for each provider unit, is presented in Table A2.3 in Appendix A2.2. Since we have converted aggregate demand into ADC, adjusting by length of stay and time period, these values represent the average daily probability of turnaway. These averages differ substantially between provider units, where some hospitals have a probability of turning away a patient of zero and others have a probability of turning a patient away of close to one.

There is however, another issue raised by these variations, that is; which is the appropriate value to use for the average probability of turnaway? If we take the average of the sum of monthly turnaway probabilities, we get a different value than if we take the average of the sum of the annual probabilities. This, once more, highlights the problem of using annual data. The estimate of the probability of turnaway based on monthly estimates is 0.040215 compared with the probability given above, i.e. 0.007124 (that is a factor of 5.6 times higher). The relative distributions are presented in Table 5.2 below. It

[8] It has been noted, as pointed out in Chapter 2, that this distribution should only really be applied when the facility is rarely full, and Blumberg identified emergency and obstetric treatments as the most likely to satisfy this assumption. Therefore, this fits our categorisation of 'emergency' demand perfectly.

is clear that using the monthly provider specific will pick up more of the fluctuations in demand, indicated by the higher standard deviation of the distribution in Table 5.2.

Table 5.2: Mean and SD of probability of turnaway

	Mean	S.D.
Annual average	0.007124	0.0309
Monthly average	0.040215	0.1759

These values provide the first part of the empirical content for equation 5.1. The next stage of the analysis is to attempt to provide an empirical estimate of MSC_w and MSC_t.

The estimates of probability of turnaway presented in Table A2.2 present the average probabilities across the all time periods. However, these averages mask a wide dispersion of probabilities within hospitals across monthly time periods. These differences potentially allow us to investigate the first two components of equation 5.1. That is, the implied social costs of turning away emergency cases and the social costs of placing elective cases on the waiting list. They potentially allow us to isolate the relationship between turnaway and waiting lists such that we can conduct a partial analysis necessary to define this relationship.

Section 3.1: Social cost of turnaway and waiting lists

The cost function in Chapter 4 provided estimates of the private marginal costs of production. The wider social perspective outlined in equation 5.1 highlighted two other costs that influence hospital capacity allocation decisions; the social cost of turnaway and waiting lists. Therefore, in order to provide the full empirical content it is necessary to augment the private cost estimates with estimates of these social costs.

Since we have information on the size of waiting lists, elective admissions, and emergency admissions, we have all the information necessary to determine the relationship between these variables. Examination of the monthly fluctuations in bed allocation will provide, through an implied value approach, the relative social costs attached to turning away emergency patients and placing elective patients on the waiting list. That is, since we assume that the total elective and emergency demands are exogenous, any shift in demand from one allocation period to the next will potentially alter the optimal allocation decision. Such fluctuations in total demand and bed allocation decisions will provide us with estimates of the implied social costs.

In order to provide such estimates we need to re-consider the optimal allocation condition identified in equation 5.1 and utilise the estimates of the probability of turnaway estimated in section 2.

3.2: Theoretical background

The first order condition presented in equation 5.1 is derived from the theoretical model, and represents a comparative static model that assumes that this condition holds for each decision period.

From the full comparative static identified in equation 4.5 in Chapter 4, we know that:

$$MSC_t(x-B_{em})[1-F(B_{em})] = MSC_w(W) + MC_{Bem} + MC_{vem}[1-F(B_{em})]- MC_{Bel} - MC_{vel} \quad (5.1a)$$

Where, MSC_t is a function of $(x-B_{em})$, which is the number of patients turned away and MSC_w is a function of W, which is the number on the waiting list. MSC_t and MSC_w are, therefore, potentially functions of the number of patients turned away and the size of the waiting list, respectively. From this relationship we know that, there are only three reasons that the optimal bed allocation decision and, hence, the probability of turnaway $[1-F(B_{em})]$ should change. These are if:

i) MSC_t changes

ii) MSC_w changes

iii) MCs changes

Since, from consideration of the monthly data, we observe monthly fluctuations in the probability of turnaway, this suggests that at least one of these components is changing in response to maintain the equality in equation 5.1a (assuming hospitals behave efficiently).

In the conceptual specification outlined in Chapter 3, the relationship between waiting lists and turnaway implied that the marginal social costs of waiting lists, and the marginal social costs of turnaway may be related to the size of waiting lists and numbers turned away. That is:

$$MSC_w = f(W) \qquad (5.6)$$

$$MSC_t = f(x-B_{em})$$

where diminishing (increasing) marginal disutility of waiting lists may exist, such that the first order derivatives $f'(W)$ and $f'(x-B_{em})$ are negative (positive) and the second order derivatives $f''(W)$ and $f''(x-B_{em})$ are negative (positive).

Furthermore, we know, once more from consideration of the monthly data, that waiting list size and emergency demand both alter between decision periods, and, as highlighted above, we know that the probability of turnaway, and hence bed allocation, alters between periods.

We assume that hospitals operate in a single decision period, one month, and therefore the waiting list size and the probability of turnaway are based on their bed allocation decision.[9] If the marginal social costs of turnaway and waiting lists are linked to the

[9] We assume a single decision period where the total size of elective demand is known by the hospital. There are more complex models of waiting list determination, see, for example, Worthington (1987), however these rely on arbitrary constructs such as 'discouragement factors', to enable stationarity to be induced in the models. Furthermore, they rely on static decisions regarding bed allocation between periods, clearly this would be inappropriate for our model.

absolute size of turnaway and waiting lists, then the fluctuations in demand for elective and emergency care, (leading to changes in the size of waiting lists and numbers turned away), will lead to changes in the marginal social costs and, hence, the probability of turnaway, if the equality in eqaution 5.1a is maintained.

Therefore, any analysis of the implied values attached to waiting lists and turnaway should focus on the three main determinants of the probability of turnaway identifed above: That is, MCs, MSC_w, and MSC_t.

Since the marginal costs are estimated on an annual basis, the only reasonable assumption to make is that it is these are fixed over monthly decision periods, (this assumption only requires that they be fixed within individual hospitals, not necessarily across provider units). These assumptions allow us to directly estimate the relationship between MSC_t and MSC_w by considering the relationship between probability of turnaway and the size of the waiting list, controlling for the size of emergency demand. The only reason the probability of turnaway will be altered, given the assumption above, is because either MSC_w or MSC_t has changed. That is, from equation 5.1, we can hold the marginal costs constant between decision periods such that:

$$MSC_t[1\text{-}F(B_{em})] = MSC_w \qquad (5.7a)$$

and where we can estimate the relationship between changes in MSC_w on the probability of turnaway, i.e. $\delta[1\text{-}F(B_{em})]/\delta MSC_w$, which is:

$$\delta[1\text{-}F(B_{em})]/\delta MSC_w = 1/MSC_t \qquad (5.7b)$$

If the probability of turnaway changes between periods then this suggests that either the marginal social cost of turning a patient away has changed, or the marginal social cost of leaving an elective on the waiting list has changed. The only reason the MSC_w will change is if, as stated above, it is related to the size of the waiting list[10] and the MSC_t will only change if the MSC_t is related to the number turned away. If, for example, the marginal disutility of waiting lists increase with waiting list size, then we would expect an exogenous increase in the size of the waiting list to lead to an increase in the probability of turnaway, holding everything else constant. Therefore the above relationship in equation 5.7b, from equation 5.1a, becomes:

$$MSC_t(x\text{-}B_{em})[1\text{-}F(B_{em})] = MSC_w(W) \qquad (5.8a)$$

and where $\delta[1\text{-}F(B_{em})]/\delta W$ is:

$$\delta[1\text{-}F(B_{em})]/\delta W = MSC_w/MSC_t(x\text{-}B_{em}) \qquad (5.8b)$$

[10] This is also consistent with waiting times since the expected wait for the marginal patient is equal to the number already on the waiting list/BEDS*LOS. To maintain consistency with the theoretical construct outlined in Chapter 3 we will focus on waiting lists. The problem of using waiting times is, however, that it moves away from a single period decision model and moves towards a more complex model that has to consider inter-temporal bed allocation decisions.

The implied social costs of turning patients away and placing patients on the waiting list can be calculated using this estimated partial relationship between the probability of turnaway and the waiting list size. Where the estimated coefficient from the partial relationship represents the left hand term in equation 5.8a and, therefore, the implied value of the marginal social cost of turnaway in relation to the marginal social cost of waiting lists.

Since we want to examine the partial relationship between MSC_w and MSC_t we also need to control for any other factors that may have an impact on MSC_t or may lead to the probability of turnaway changing. If we introduce the size of emergency demand then this can control for two important factors that may alter the probability of turnaway. First, this variable controls for the relationship between the number turned away and the MSC_t, i.e. $(x-B_{em})$ in 5.8b, since the total number of emergencies turned away will be determined by the size of demand and MSC_w might be related to the number of cases turned away. This variable also controls for the direct impact that an increase in demand has on the probability of turnaway. Therefore, when estimating the partial relationship, we will include an emergency demand variable in the equation.

Two further restrictions are necessary. First, we must assume that the relationship between MSC_t and MSC_w is constant over time, i.e. that the transformation between the two variables does not itself vary over time. Secondly, we need to control for the individual provider effects, since it may be that the relationship between the two variables varies between providers. Controlling for individual provider effects will also allow for any differences between marginal social costs across provider units, which could also be a potential source of variation in the probability of turnaway.

The calculation of the marginal relationship between turnaway and waiting list is, however, further complicated as, for the empirical calculation, we need to adjust this relationship to effect the impact that altering the probability of turnaway has on the expected marginal costs. Since, although we assume that marginal costs are constant across decision periods, a change in the probability of turnaway will have an impact on the expected marginal variable costs of providing emergency, such that the partial relationship in equation 5.7a becomes:

$$MSC_t[1-F(B_{em})] - MC_{vem}[1-F(B_{em})] = MSC_w \qquad (5.9)$$

Any estimation of the relationship between the probability of turnaway and the marginal social costs of leaving an elective patient on the waiting list must, therefore, take this extra element into account.

Since there may be many different influences on the probability of turnaway, (e.g. the hospital's ability to deal with overcrowding of facilities, their ability or willingness to cancel elective admissions, their ability to discharge patients early, or the availability of substitute facilities such as trolleys and corridors), one way to determine the specific relationship between waiting lists and turnaway is through a multivariate regression, which controls for these other factors. This allows us to estimate the coefficient on the waiting list variable with respect to changes in the probability of turnaway, which gives the left hand side term in equation 5.8b, taken from the regression of probability of turnaway on waiting list size.

3.3.1: Model specification

The use of regression analysis raises the same issues as have been dealt with in the empirical estimates in Chapter 4. Most notably issues of functional forms and panel data analysis. Since we have no a priori expectations regarding the functional relationship between waiting lists and turnaway, it seems clear that we should allow the data to determine the appropriate specification. We, therefore, once again, use a Box-Cox analysis. The issue of panel data analysis can be more easily resolved, since we have a priori expectations that individual providers will have different factors influencing the probability of turnaway, this implies the use of fixed effects models to attempt to model all the potential arguments that enter the individual provider's objective function.

3.3.2: Data

Any regression analysis is restricted by the availability of appropriate data. Table 5.3 indicates the activity data that were available.

Table 5.3: Activity data

Variable name	Definition
ACTIVITY	
ADMS	Total number of admissions
EMADM	Number of admissions categorised as emergency (by provider), maternity care, and transfers from other providers
ELADM	Number of admissions categorised as elective (by provider)
BEDS	Total number of staffed beds
ALOS	Average length of stay (days) for all admissions
ELALOS	Average length of stay (days) for elective admissions
EMERGALOS	Average length of stay (days) for emergency admissions
OTALOS	Average length of stay (days) for admissions categorised as other
DAYATT	Total number of day attendances
AEATT	Total number of non-inpatient Accident and Emergency attendances
OPATT	Total number of non-A&E outpatient attendances
CATCHSH	% share of demand (DoH (1989))
CATCHD	Estimated demand (DoH (1989))
WAIT	Total number of elective patients waiting for admission
AVCMIX	Casemix weight based on DRG weights

The aim of the regression analysis was to get an estimate of the partial relationship between waiting list size and probability of turnaway. Given the limited data availability there were only a few variables that we could include in this model with theoretical justification. Therefore, the use of provider specific dummy variables were thought to be the best way of picking up such effects, and they would allow greater flexibility in picking up any other provider specific influences, such as the ability to cancel elective admissions to create additional bed capacity, the impact of geographic location, and the availability of other hospital facilities to act as a safety net. The use of provider specific dummies does, however, preclude the inclusion of time invariant factors, such as catchment share, which may have an influence on the willingness of hospitals to turn patients away and hence the social costs attached to doing so. However, this trade-off was thought to be valid since the provider specific effects should also pick up any influences that are time invariant. We estimate a simple model outlined in section 3.2.3 below.

There are, however, other issues that need to be addressed, most notably potential problems of endogeneity. Since waiting lists are partially determined by the number of beds given to the elective sector, and hence related to the number of beds given to the emergency sector there is a problem of endogeneity with the waiting list variable. That is, the expected probability of turnaway, E(Pt) and the size of the waiting list, are both partially determined by the number of beds given to the emergency sector[11].

Therefore it is necessary to instrument for the waiting list. The most straightforward way of instrumentation is by using the identity:

$$WAIT = D_{el} - ELADM \qquad (5.10)$$

Rearranging gives:

$$D_{el} = WAIT + ELADM \qquad (5.10a)$$

[11] From equation 5.5 we have specified that $PT = f(BEDS_{em}, ADC_{em})$ and since $ADC = (D_{em} \times LOS_{em})/t$ (where D_{em} is total expected emergency demand and LOS_{em} is emergency length of stay), this gives have $E(Pt) = f(BEDS_{em}, D_{em}, LOS_{em})$. Furthermore, we know that $WAIT = D_{el} - ELADM$ (where D_{el} is total elective demand and ELADM is total number of electives treated) we also know that $ELADM = (BEDS_{el} \times t)/LOS_{el}$ so that $WAIT = f(BEDS_{el}, D_{el}, LOS_{el})$. Since the total number of beds available is fixed we know that $BEDS_{el} = BEDS_T - BEDS_{em}$, this means that $WAIT = f(BEDS_{em}, D_{el}, LOS_{el})$. Therefore the waiting list and probability of turnaway are both determined by the number of beds allocated to the emergency sector.

If we assume that for any given time period that the total elective demand is exogenous, which is reasonable, then we can simply replace WAIT with D_{el} in the estimated equation. Providing the waiting list is non-zero, we can interpret the coefficient on D_{el} list in the same way we would the coefficient on WAIT. This allows us to examine the impact of an exogenous change in the size of the waiting list on the probability of turnaway and, therefore, the relationship between the two elements of social cost.

Including a variable to adjust for numbers turned away has similar endogeneity problems, since the number of emergency patients turned away, $(x-B_{em})$, is determined by B_{em}, which is endogenous. Therefore we need to instrument in a similar manner as above, using emergency demand as a proxy for numbers turned away.

3.3.3: Specification

The basic specification of the model estimating the probability of turnaway will be of the form:

$$PT = f(D_{el}, D_{em}, DV_i) \qquad (5.11)$$

Where PT is the probability of turning a patient away, D_{el} is the total number of elective patients requiring treatment and D_{em} is the total emergency demand. DV_i represent the provider specific effects. D_{em} is represented by ADC which proxies for the number of emergency patients turned away (where total number of patients turned away = PT x ADC) and the impact direct impact of ADC on PT.

As with the cost function we will employ a Box-Cox analysis allowing the data to determine the appropriate functional form. As before we will use the most general form of the Box-Cox transformation, which allows the dependent and independent variables to be transformed by different values such that:

$$y^{(\theta)} = \alpha + \Sigma\beta_k x_k^{(\lambda)} + \varepsilon \qquad (5.12)$$

There are, however, a number of other issues that have to be dealt with before we move onto the estimation of the model. The main issue is that, as indicated in Table A2.2, there are a number of zero values for the dependent variable. The existence of zero values can cause a number of problems, both practical and theoretical.

Dealing with the practical issue first; the Box-Cox analysis can deal with non-zero values, however this requires that both θ and λ are non-negative and lie between zero and one. In addition, this requires that the Box-Cox transformations be performed manually, (i.e. not through LIMDEP using the MLE search). On a theoretical level, the existence of zero values can be dealt with in a number of different ways.

There are a number of options to overcome the problem of zero values. First, we could undertake the analysis on the whole sample using a manual Box-Cox analysis. This has the advantage of using the full sample, but, as identified above, has the disadvantage of restricting the values of θ and λ such that they are non-negative and lie between zero and one. Secondly, we could select-out the non-zero values and do the analysis on the restricted sample. This, however, may introduce bias. Thirdly, we could use a model that explicitly deals with unobserved values of the dependent variable, such as the censored TOBIT model that treats zero values as censored.

The TOBIT model is similar to the PROBIT and LOGIT models, which assume that the left-hand side variable either takes the value of one or zero, and therefore, takes the form of a dummy variable approach for the dependent variable[12]. The model assumes that the dependent variable is not observed, rather that it takes a zero or one value[13] (e.g. working or not working, taking out a loan or not). This gives the following:

$$\begin{array}{lll} y_i = & 1 & \text{if } y^* > 0 \\ y_i = & 0 & \text{if } y^* \leq 0 \end{array} \tag{5.13}$$

The approach calculates a probability of an event occurring based on the explanatory variables. The TOBIT model, however, allows the dependent variable to be observed such that y^* is observed if $y^* > 0$ and is not observed if $y^* \leq 0$. The observed relationship therefore becomes:

$$\begin{array}{lll} y_i = & y_i^* = \beta_i + \mu_i & \text{if } y^* > 0 \\ y_i = & 0 & \text{if } y^* \leq 0 \end{array} \tag{5.14}$$

[12] See, for example, Maddala (1993) for a full explanation of these models.

This model, however, still treats the zero values as unobserved and, in principle, these could take negative values. This is not the case in our analysis. Maddala (1992) warns against the mechanical use of the TOBIT model for these reasons, i.e. that these values are zero, not because they are unobserved, but due to the decisions of hospitals. In this situations it is advisable to model the process that leads to zero values.

The final option we could use is a Heckman two-stage procedure that allows us to treat the zero values as representing a decision (see Heckman, 1979). This model has applied to the participation in the labour supply market (Heckman, 1976) dealing with the issue of non-labour market participation, and the idea of a reservation wage below which workers do not enter the market. The situation may be may be similar in hospitals. If hospitals do not appear to be trading turnaway against waiting lists, in some periods, then this suggests that there may exist a cut-off point where the hospitals do not 'enter the market'. The Heckman procedure involves estimating a PROBIT model first to identify the selection mechanism and then estimation of a TOBIT model. This is designed specifically to deal with zero values, where the zero values are a decision variable rather than 'unobserved'.

The PROBIT equation is estimated using maximum likelihood to obtain estimates of γ. Where, for each observation, a value of λ is constructed to compute:

$$\hat{\lambda}_i = \frac{\phi(\hat{\gamma} w_i)}{\Phi(\hat{\gamma} w_i)} \qquad (5.15)$$

---

[13] Models such as this have been applied to the labour market where the dependent variable is zero or one representing working or not working, see Quester and Greene (1982) for an example of this approach.

using the PROBIT coefficients.

Then β and $\beta_\lambda = \rho\sigma_\varepsilon$ are calculated by least squares regression of y on x and λ.

As there is no a priori evidence to allow us to choose over the different models we considered four approaches:

1) A Box-Cox analysis considering all values of the dependent variable.
2) A Box-Cox analysis considering non-zero values of the dependent variable.
3) A censored TOBIT model which categorises treats zeros as unobserved or censored.
4) A two-stage Heckman procedure.

3.3.4: Results of waiting list-turnaway estimation

The results from the regression analysis allow us now to complete the empirical content of the first order condition, and identify the implied social costs associated with waiting lists and turnaway. From the regression equation we can determine the trade-off between waiting list and probability of turnaway. The calculation is based on the elasticity calculated from the coefficient on D_{el}. This coefficient allows us to calculate the marginal effect on the probability of turnaway of an increase in the waiting list, holding ADC constant.

The full regression results for all the approaches are presented in Appendix A2.3. The same grid search technique was used as in Chapter 4. A manual search was undertaken to identify the values of lambda and theta between which the optimal values lie. The results of the grid searches for the Box-Cox models are presented in Table A2.3.1 and A2.3.3.

The regressions performed relatively well; the coefficient on waiting list size was positive and significant indicating that there does appear to be a trade-off between waiting lists and probability of turnaway. The Box-Cox sub-sample model (PT > 0, see Table A5.3.1) indicated a transform close to semi-log. The full sample Box-Cox model indicated a linear relationship between the variables (see Table A2.3.3 in Appendix). Consequently the Tobit and Heckman models were estimated on untransformed data. There were, however, problems with the Tobit and Heckman models.

First, since the Tobit model basically assumes that zero values were unobserved this may induce bias. Furthermore, it represents a restricted version of the Heckman model, therefore it was considered that this model should be discounted early on.

The Heckman model, however also has problems, since the estimated equation has a large number of dummy variables, this causes problems for the first stage of the estimation process in the Heckman model. That is, the probit model cannot deal with left-hand and right-hand side variables that are both equal to one with no variation for a large number of observations. Consequently, the dummy variables had to be dropped from the first stage of the analysis; this led to an unstable estimate of the relationship (see Table A2.3.6). Overall the Heckman model performed reasonably well (see Table A2.3.7). The insignificant t-statistic on lambda, however, indicates that an estimate on just the sub-sample of data, excluding the zero values, may itself be unbiased. To test this we estimated a Box-Cox linear regression using only the sub-sample of PT (see Table A2.3.8). The results of this analysis confirmed that estimation using a sub-sample might indeed be unbiased, since the results were almost identical to the Heckman two-stage procedure.

On this basis, we estimated the partial relationship between waiting lists and turnaway probability using the sub-sample, which allowed lambda and theta to take any values (i.e they were not restricted to positive values between zero and one as in the full sample). We employed a Box-Cox MLE analysis to identify the appropriate transform, (see Table A2.3.1). This analysis indicated a (close to) semi-log relationship. The summarised results of this analysis are presented in Table 5.4. (The full results are in Table A2.3.2 in Appendix A2.3, i.e. including provider specific dummy variables).

Table 5.4 Box-Cox regression (sample = non-zero PT)

Variable	Coefficient	Standard Error	t-stat
DEL	0.34788E-02	0.40177E-03	8.659**
DEM	0.70107E-02	0.34797E-02	2.014*
Lambda	0.76969		
Theta	0.065794		

Log-likelihood = 8644.05
N = 814
White test for heteorskedasticity: 0.1599
Autocorrelation: 0.0006

* significant at 10% level
** significant at 5% level

3.3.5: Implied trade-off between turnaway and waiting list size

The figures presented in Table 5.5 are calculated from the coefficient taken from Table 5.4.

Table 5.5 Elasticity and trade-off

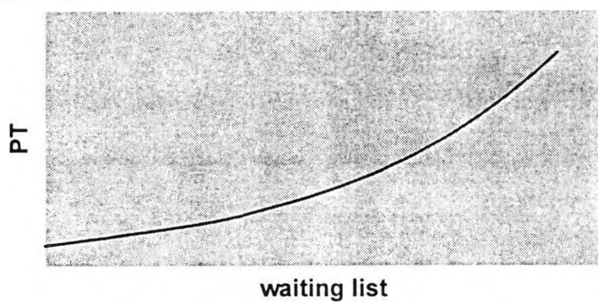
Variable	Mean	Elasticity	Trade-off
DEL[1]	3772	5.8167	0.0000619913
PT[2]	0.0402		

[1] The trade-off between waiting lists and PT must be estimated using the mean value of DEL (i.e. rather than WAIT) since the coefficient is estimated based on DEL.
[2] The trade-off is calculated at the mean of the full sample to allow comparison with the second stage of the analysis.

The elasticity of substitution between the probability of turning a patient away and the size of the waiting list, calculated at the mean, is estimated to be 5.8167. That is an exogenous 1% in the size of the demand (waiting list) will lead to a 5.8167% increase in the probability of turning a patient away. From the above estimation, on average, hospitals will increase probability of turnaway by 0.0000619913 when the waiting list increases by one. Therefore, if waiting list size increases by one, (representing the value of the MSC_{W}, from equation 5.8b), the probability of turnaway adjusts by a relatively small amount. Nonetheless, this suggests that as waiting list size increases that hospitals respond by increasing the probability of turnaway. This suggests that, holding all other things constant, the marginal social cost attached to leaving an elective patient on the waiting list does indeed increase as the size of the waiting list increases.

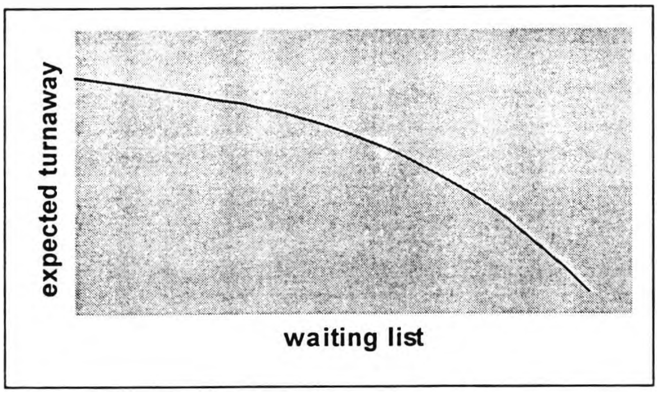
The Box-Cox transform implies that the trade-off between waiting lists and probability of turnaway is non-linear (and close to semi-log), such that as the size of the waiting list increases, the hospital reduces the rate at which it is prepared to trade-off the probability of turnaway against waiting list size. This gives the relationship in Figure 5.1 below.

Figure 5.1: Trade-off between probability of turnaway and waiting list size



This estimated relationship between an exogenous increase in waiting lists and the adjustment of bed allocation in response to this embodies the marginal rate of substitution of emergency turnaway and elective waiting lists and, therefore, we can derive the shape of the indifference curve from this relationship. The relationship implies an increasing marginal disutility of waiting lists such that the relationship between waiting lists and probability of turnaway represented on an indifference curve would be convex (i.e. as represented in Figure 5.2 below).

Figure 5.2: Indifference curve turnaway-waiting list



Therefore, the empirically estimated indifference curve is consistent with the shape represented in the theoretical model in Chapter 3. That is, the assumed marginal rate of substitution is validated.

3.3.6: Implied social costs

Estimation of the cost function enabled the marginal variable and quasi-fixed costs to be estimated and put back into the theoretical model. The marginal variable cost of an emergency case was estimated at £111, and the marginal quasi-fixed costs of staffed beds for an emergency case was estimated to be £760. The marginal cost of an elective case was estimated at £550.

From equation 5.1 we know that the optimal allocation is achieved where:

$$MSC_t[PT)] - MSC_w = MC_{Bem} + MC_{vem}[PT)] - MC_{Bel} - MC_{vel}$$

and we have estimated the terms on the right hand side of the equality from the cost function such that:

$$MSC_t[PT)]- MSC_w = £760 + £111_{em}[PT)]- £550 \qquad (5.16)$$

We do, however have to make two adjustments at this point. In the initial specification of the model we assumed that length of stay for emergency and elective cases was equal. Therefore we have to adjust for the difference in length of stay. Furthermore, we have to adjust for time period, however, since the probability of turnaway is a daily estimate we only need divide by length of stay to derive the estimated cost per day.

Since the full relationship should be:

$$MSC_t[PT)]- MSC_w(B.t/LOS_{el}) = MC_{Bem}(B.t/LOS_{em})+ MC_{vem}[PT)]$$

$$- MC_{Bel}(B.t/LOS_{el}) - MC_{vel}(B.t/LOS_{el}) \qquad (5.17)$$

Where B represents a bed, t is time period and LOS_{el} and LOS_{em} are elective and emergency lengths of stay, respectively. (Where PT is already adjusted for length of stay and time period therefore no further adjustment is necessary). Since we differentiate with respect to beds to derive the first order condition, all we need to do now is multiply through by (t/LOS). Furthermore, since the probability of an extra emergency case arriving is PT in any one day, the total expected marginal variable cost for emergency cases is MC_{vem} multiplied by PT, whereas the cost of emergency beds is adjusted for the expectation of being filled, as we assume that the bed numbers are decided ex ante, and fixed ex post. This is entirely consistent with the theoretical construct.

From the analysis in Chapter 4 we know that:

$MC_{vem} = £111$

$MC_{Bem} = £760$

$\Sigma(MC_{vel} + MC_{Bel}) = £550$

We also know that the average values for the length of stay in the two sectors are[14]:

$LOS_{el} = 4.76$

LOS_{em}[15] $= 6.02$

And that:

$t = 1$

We can now say that the optimal allocation occurs where:

$$MSC_t[PT)]- MSC_w/4.76 = 760/6.02 + 111[PT]/6.02 - 550/4.76 \quad (5.18)$$

where [PT] is the probability of turning a patient away.

[14] See Appendix 3 for data description.

[15] Where LOS_{em} represents an average of emergency and other admissions, weighted by the number of admissions.

The probability of turning a patient away has been estimated at 0.0402. Therefore the implied marginal social cost of turnaway relative to the marginal social cost of placing a patient on the waiting list would is estimated to be:

$$MSC_t[0.0402] = 0.21MSC_w + 10.69 + 18.44[0.0402] \qquad (5.19)$$

therefore:

$$MSC_t = 5.22MSC_w + £265.92 \qquad (5.20)$$

Consequently, the optimal allocation is a trade-off between the relative size of the social costs attached to non-treatment, and the difference between the marginal costs of treatment and staffed beds.

From the above estimation, on average, hospitals will increase the probability of turnaway by 0.0000619913 when the waiting list increases by one. From equation 5.7d we also know that:

$$MSC_t = 1/\beta MSC_w + MC_{em}/LOS_{em} \qquad (5.21)$$

Where β represents the coefficient from the estimated partial relationship.

However, we have to make a further adjustment here, since we need to convert the probability of turnaway into actual numbers turned away, in order to make the appropriate calculation. The expected number of patients turned away is average (expected) number of daily admissions x PT, therefore an increase in turnaway by

0.0000619913 will increase the number of emergency patients turned away by 0.00438573 per day (i.e. 70.75 x 0.0000619913). Consequently, the trade-off between marginal social costs of turnaway and waiting lists, calculated at the mean values of both is:

$$MSC_t = [228.01]MSC_w + £18.44 \quad (5.22)$$

and:

$$MSC_t = 5.22MSC_w + £265.92 \quad (5.23)$$

Solving simultaneously gives:

$$MSC_t = £290.60$$

$$MSC_W = £1.19$$

The marginal social cost of turning a patient away dominates this relationship, at nearly £300, whereas the marginal social cost of leaving an elective case on the waiting list is estimated to be a little over £1.

The marginal social costs of turning an emergency patient away are relatively small, especially given the potentially catastrophic implications if emergency patients are not treated. There are, however, a number of potential explanations for this: First, if the hospital is full it does not necessarily mean that the patient will be turnaway away. There are increasing reports in the media of emergency patients being placed on trolleys and in waiting rooms. The social cost may refer to placing patients in inappropriate facilities. Secondly, when the hospital is full, the hospital may have other options available rather than turning a patient away, that is, they can cancel elective treatments to make room for the extra patients, therefore, the social costs may also reflect this response. Nonetheless, these estimates appear to be reasonable in terms of the implied value attached to turning an emergency patient away relative to placing an elective patient on a waiting list (i.e. in the region of 200 times greater).

Section 4.1: Output heterogeneity and demand uncertainty

Let us now reconsider two issue that have consistently featured throughout this thesis; output heterogeneity and demand uncertainty. In this section we will consider the impact of each on the estimation of the implied social costs.

4.2: Heterogeneity

If, we failed to recognise the heterogeneity issue and concentrated solely on the social costs of the hospital being full, then the decision moves away from the use of existing capacity and focuses, as other authors have done, on the issue of how many beds should be supplied, in aggregate. In order to estimate the implied social costs of turnaway in aggregate, we need to first adjust the marginal cost estimates, since the costs of provision would be based on an average marginal cost of output in aggregate, rather than separating output into emergency and elective cases. Nonetheless, the basis of the estimation would be the equation:

$$MSC[PT] = MC_B/LOS + MC_v[PT] \qquad (5.24)$$

Where:

MSC = marginal social cost of turning away a patient

LOS = average length of stay

MC_v = marginal variable costs

MC_B = marginal costs of a bed

Let us now adjust the cost estimates to provide the average values necessary. If we assume that the same proportion of elective marginal costs are due to the quasi-fixed element (i.e. staffed beds) as emergencies, then this enables us to estimate the variable and quasi-fixed costs for elective cases. That is, the variable element of emergency costs represents 13% of total marginal cost (i.e.111/871). Applying this proportion to electives gives £70 variable cost and £480 staffed bed cost. Since approximately 50% of admissions are elective and 50% emergency, the average marginal cost of a staffed bed, and the average variable cost, are £620 and £90, respectively. The average length of stay can be calculated in the same way, i.e. (4.76+6.02)/2, which is 5.39. If we assume that the probability of turnaway is the same as that calculated for emergency cases, (although, as we have indicated earlier, this may not be the case but will suffice for illustrative purposes), then we have all the data necessary.

This gives the following equation:

$$MSC[0.0402] = 620/5.39 + (90/5.39)[0.0402]$$

Therefore, the marginal social cost, not accounting for heterogeneity, would be £2,878. To provide some comparison we can compare this with the average estimate by Carey, which was $24,710. Consequently, the implied social cost of an extra bed is much smaller in the UK, which may be expected as there is likely to be less reserve capacity in the UK and, thus, higher turnaway probability. This is illustrated by using Carey's own probability of turnaway (0.0062) to calculate the social costs. If we use this value the implied social costs are estimated to be £18,570. This compares with Carey's average estimate of $24,710 (1987 prices). Therefore these two estimates are very similar.

Clearly the probability of turnaway has an important influence on the estimated social costs. To estimate the impact of changing the probability of turnaway on our estimates, we will estimate a marginal social cost of turnaway elasticity with respect to the probability of turnaway

There is, however, another issue that will affect the social costs; the total bed availability. If there is a limited bed supply, this will inevitably impact the implied social costs since the greater the total bed availability, the lower the probability of turnaway, since more beds are available to allocate to both elective and emergency cases, therefore the implied social costs associated with both will fall. Clearly, with a chronic hospital bed shortage, the length of waiting lists may become so large that the hospital does not have sufficient capacity to enable it to meet the conflicting demands, and achieve an optimal level of both waiting list size and turnaway rate, (i.e. that reflects society's values). This will be reflected by the relatively low implied marginal social costs of turnaway. This is a public policy issue regarding funding and availability of hospital beds. If the implied social costs of turnaway are regarded as 'too small', in some subjective sense, then this suggests there are not enough hospital beds in total. The low implied marginal social costs associated with turning emergency patients away calculated here suggest that this may be the case in the UK.

4.3.1: Probability of turnaway: annual average

If we utilised the annual rather than monthly average probability of turnaway the implied marginal social costs would, from equations 5.19 and 5.20, become:

$$MSC_t[0.007124] = 0.21MSC_w + 10.69 + 18.44[0.007124]$$

$$MSC_t = 29.48MSC_w + £1500.56 + £18.44$$

Therefore:

$$MSC_t = 29.48MSC_w + £1500.56 + £18.44$$

And $$MSC_t = 1286.65MSC_w + £18.44$$

Solving simultaneously, the marginal social costs are estimated to be:

$$MSC_t = £1554$$

$$MSC_W = £1.19$$

The impact of the probability of turnaway is quite substantial in this model. If we significantly underestimate it, as using annual data may well do, then this could lead to serious overestimates of the implied social costs. (MSC_w remains the same in this example since we have imposed a new probability of turnaway and assumed that the marginal trade-off between waiting lists and turnaway probability remains the same as estimated earlier).

4.3.2: Cost elasticity with respect to probability of turnaway

By altering the probability of turnaway by a small amount, e.g. 1% we can estimate the impact this has on our social cost estimates by re-estimating equation 5.19 using the new values. Therefore, let us estimate the impact of a 1% reduction in PT, such that PT becomes 0.039798. The marginal social costs of turnaway now become £293.36, respectively. Therefore the implied cost elasticity is –0.95. This suggests that the impact on marginal social costs of altering PT is almost proportional. If, however, we consider the absolute size of the change, i.e. a 1% change in PT is only 0.000402, then it is apparent that the impact of altering the probability of turnaway is potentially large (for example using annual rather than monthly figures represents a 823% reduction in PT).

It is clear that the social costs depend crucially on the estimation of probability of turnaway and the estimates of the marginal costs of elective and emergency care, since the difference between these two is an important determinant of the implied values. If the marginal social costs of leaving an elective patient on the waiting list were zero, then the implied marginal social cost of turning an emergency patient away would be solely determined by the magnitude of this difference.

The size of the marginal social cost is determined jointly by the difference between the marginal costs of provision and the probability of turnaway. However, if the implied marginal social costs of turnaway appear to be small in some subjective sense, this may indicate that the hospital sector has too few beds. Increasing the number of beds will reduce the probability of turnaway, and increase the implied social costs of turning a patient away. The fact that governments take this action by ring-fencing extra money for emergency beds may indicate that they do indeed think that the implied social costs are too small.

Section 5: Conclusions

This chapter has considered the optimal use of hospital capacity from a wider social perspective and has calculated the implied social costs associated with turning away emergency patients and leaving elective patients on the waiting list. The underlying theoretical model employed in this chapter has maintained consistency with that developed in Chapter 3 and the empirical analysis undertaken has utilised the private marginal cost estimates derived from the cost function analysis in Chapter 4. Whilst the actual estimation procedure is complex the implied social costs calculated from the empirical analysis in this chapter seem reasonable.

Previous work in this area focused solely on the total number of beds required, rather than the optimal use of existing bed capacity and, as a result, failed to take account of the inherent trade-off between the use of hospital capacity. As indicated in section 4.2, this will lead to an over-estimate of the implied social value attached to turning a patient away, and fails to distinguish between the different types of demand for hospital care and the way it can be treated. As such, this analysis represents the first specification of the actual trade-off between the different hospital outputs. The results indicate that the hospital places greater emphasis on not turning emergency patients away, and the higher implied social costs of turning an emergency patient away relative to the social costs attached to placing an elective patient on a waiting list, reflect this. Furthermore, the empirical results presented in this chapter allowed us to estimate the actual shape of the indifference curve between turnaway and waiting list size. This indicated a convex indifference curve, such that there would be an increasing marginal disutility attached to turning patients away.

Chapter 6: Conclusions

This thesis has investigated the utilisation of hospital capacity under conditions of uncertainty and output heterogeneity. A model of hospital capacity utilisation decisions has been developed that is located firmly in economic theory. The specification of the theoretical model allows an empirical investigation of the hospital's behaviour to be developed. The analysis is rich enough to address the notion of efficient use of capacity using a private and social perspective.

Chapter 2 provided a review of the literature within hospital economics that deals with issues related to capacity utilisation. Two distinct strands were highlighted; those developed under conditions of certainty and those developed under uncertainty. The first set of theories considered the aims and objectives of the hospital as a non-profit institution and developed, in some cases, grand theories of hospital behaviour. The second set of theories considered hospital behaviour addressing issues of capacity use. These allowed some of the important issues regarding hospital behaviour to be identified, most notably uncertainty. There were, however, some problems with the empirical assessments associated with these models as they often failed to fully specify the problem, or were inadequately located in economic theory without specifying a formal objective function. Consequently, many of these more formal models were unable to provide optimal solutions to the problems they identified. The final section of Chapter 2 reviewed empirical estimates of hospital costs. These identified the theoretical foundations of the empirical studies to date, and highlighted the lack of focus on issues such as demand uncertainty, which has only recently been addressed.

Chapter 3 constructed a formal model of hospital behaviour, building on theoretical foundations, to enable the problem of hospital capacity allocation decisions to be viewed within standard economic theory. The basic model introduced output heterogeneity, separating output into two components, planned and unplanned (elective and emergency) and considered bed allocation decisions from a private perspective. This was then represented in a geometric representation using a standard production possibility framework. Uncertainty was then introduced into this model, allowing stochastic demand for one of the outputs. The impact of uncertainty was considered in a formal manner by drawing on current theoretical knowledge regarding the influence of demand uncertainty on the production responses of the firm. This represents, to our knowledge, the first formal specification of uncertainty using such a framework. Reserve capacity was shown to be consistent with the efficient allocation of capacity under these circumstances. The objective of the non-profit hospital within the NHS was then re-considered, to broaden the aims beyond the narrow private perspective of surplus maximisation. We considered a utility maximand where we assumed that hospitals derive utility from treating patients and disutility if patients have to be turned away, or queued. A four-sector model was outlined representing the two outputs and four arguments in the utility function, showing how the trade-offs could be interpreted. The impact of demand uncertainty was considered with reference to the four-sector model, allowing the optimal allocation decisions of the hospital to be compared under conditions of uncertainty and uncertainty.

Once more, reserve capacity was shown to be a rational response to demand uncertainty. The optimality condition highlighted the empirical content necessary to identify a fully specified model of hospital allocation decisions.

In Chapter 4 we built on this theoretical model and considered a restricted optimality condition that allowed some empirical content to be provided to the model. To this end, we considered a cost minimand, where costs included private and social costs, i.e. costs of production and costs associated with the disutility of turning away or queuing patients. This minimand identified six elements that needed to be identified to provide a fully specified optimality condition. These were the variable and quasi-fixed marginal costs of elective and emergency cases, and the marginal social costs of turning patients away and placing patients on waiting lists. However, since private costs are the only ones observed we focused on the estimation of a private cost function.

In Chapter 4 we estimated a cost function consistent with the theoretical model, which adjusted for demand uncertainty and output heterogeneity. In this chapter we outlined previous studies that have considered the issue of uncertainty and the impact on hospital costs and, building on this earlier work, we considered, in a formal way, the impact of uncertainty on hospital costs. The implications for the estimation of a hospital cost function were then outlined. The analysis of the theoretical underpinnings and the review of previous studies highlighted the need to adjust for demand uncertainty, and in particular that hospitals would attempt to minimise expected ex ante costs. Therefore, we considered a way of adjusting for demand uncertainty through estimating demand expectations. A number of different methods for estimating the demand for unplanned care were attempted. Various problems were highlighted with the approaches used. The best method, in terms of statistical and theoretical considerations, was a simple autoregressive process, one that had been applied by authors when estimating hospital demand in previous studies. The estimation of hospital demand allowed us to construct a variable that in theory would take account of demand uncertainty, and would fit in with the theory developed in the previous chapter.

Having estimated the impact of demand uncertainty, we used this to assess the private cost minimising position of the hospital. When estimating the cost function, we also considered a number of different approaches. The translog approach, which allowed considerable flexibility, the Cobb-Douglas model, which is essentially a restricted translog model, and the Box-Cox model, which allows the data to determine the appropriate functional form. The translog model performed badly with a large number of insignificant coefficients and counter-intuitive signs. Two potential problems were highlighted with this approach. First, it involves estimating a large number of independent variables and, therefore, the degrees of freedom are significantly reduced. Furthermore, due to the large number of independent variables that are constructs of each other, there was a significant problem of multicollinearity, so much so that the estimates of the coefficients on the independent variables may be biased. The Cobb-Douglas restricted model performed little better. The Box-Cox approach, however, led to a model that appeared to perform well, with a large number of significant coefficients with intuitive signs. Most notably, the demand variable included to pick up the influence of demand uncertainty was both significant and of the correct hypothesised sign.

Chapter 5 considered the relationship between waiting lists and turnaway rates to provide the further empirical estimates necessary to consider optimal allocation decision from a wider social perspective. The cost function, estimated in Chapter 4, allowed four elements of the empirical content to be identified; these were the variable and quasi-fixed marginal costs of elective and emergency care. Two further elements were therefore left to be identified; the marginal social costs of waiting lists and turnaway rates.

In Chapter 5 we were particularly interested in the impact of an exogenous shift in waiting list size on the probability of turnaway. This potentially allowed the partial relationship between waiting lists and turnaway probability to be identified. Before this analysis could be undertaken, we needed to calculate the probability of turnaway rates. To do this it was necessary to re-estimate demand, adjusting for time period and demand size. This allowed provider specific probabilities of turnaway to be calculated. The calculation of these probabilities maintained consistency with the theoretical model by assuming that all elective beds were occupied, and that the remaining beds were available for emergency cases, if required.

The implied marginal social costs of turnaway and waiting lists were derived by regressing the probability of turnaway on total elective demand, adjusting for the size of emergency demand, and using the estimated coefficient to represent the trade-off between the two. Since we observed a number of zero values for the probability of turnaway, and since this was the independent variable and may introduce bias if we used include all values, we considered a number of different estimation techniques, once more allowing the data to determine functional form. A Tobit analysis considered zero values as unobserved. A Heckman two-stage procedure considered the zero values as a choice. We also considered two Box-Cox analyses, where we used the entire sample, and only those non-zero values of probability of turnaway. The results of the analysis indicated that employing a restricted sample, (i.e. non-zero values of PT), and estimating using the Box-Cox method, may not in fact introduce any bias. Therefore, we considered the empirical content using the limited sample Box-Cox method, which allowed the functional form to be determined using the full sample data. The partial relationship between these variables was, therefore, derived based on the coefficient on waiting list size. This allowed the final empirical content to be identified.

Based on the fully specified empirical model, the implied marginal social costs of turning an emergency patient away were estimated to be around £300, and the implied marginal social costs of placing a patient on the waiting list were estimated to be just over £1. These represent indicative values, and are based on aggregate estimates across all hospitals, as such, they may mask considerable differences between hospitals. Nonetheless, they provide the first estimates of this kind adjusting for demand uncertainty and including output heterogeneity.

The analysis in this thesis has separated admissions in a way that has not been attempted before. If we do not recognise output heteogeneity, then the marginal social costs of turning away, or queuing, the different types of admissions, can only be estimated in aggregate. If we assume that reserve capacity is held in order to treat all patients, then this is likely to over-estimate the probability of turning away patients. This will have the effect of reducing the implied social costs attached to turning patients away. This may lead to an under-estimate of the social costs attached to turning emergency cases away, and an over-estimate of the social costs of queuing elective cases. This error may well be significant; the empirical estimates in Chapter 5 suggest that the marginal social costs of turning away emergency patients may be a factor of more than 200 greater than the marginal social costs of queuing an elective patient.

The estimates presented in this thesis suggest that cost analyses that fail to take account of uncertainty may produce biased estimates. If marginal cost estimates are used for the contracting process, this has potential implications for the pricing hospital services, if demand uncertainty is ignored. More generally, it suggests that economic evaluations should also take account of uncertainty when estimating hospital costs. Furthermore, the potential impact of demand uncertainty on costs has implications for the construction of efficiency indices. That is, the more uncertain demand is, the higher costs are likely to be, therefore, any ranking of hospitals on the basis of costs should take account of the extent of demand uncertainty, and the nature of demand for individual hospital services. If, for example, an efficiency index is constructed without adjusting for uncertainty, it may, incorrectly, attribute the cost residual solely to inefficiency, rather than taking into account uncertainty. The implied social values calculated here may allow us to construct social efficiency rankings. That is, they potentially allow hospitals to be ranked on the basis of the implied social costs attached to turnaway and waiting lists. Thus, allowing a wider social perspective to be included in efficiency tables.

Throughout this thesis we have attempted to maintain theoretical consistency and build a model of hospital behaviour located firmly in economic theory. Furthermore, we have attempted to apply a more rigorous theoretical, mathematical and empirical basis to the model. This has facilitated two things. First, we have been able to identify an optimal allocation decision based on the theoretical model. Secondly, we have been able to provide the empirical input necessary to derive the implied social values derived from the first order condition. There are, however, a number of issues that we have not been able to address in this thesis. First, we have not been able to consider, what may be termed the 'micro' response of hospitals to the problems of capacity utilisation, most notably, the potential adjustment of admissions policies, discharge policies and length of stay to the problem of capacity constraints. However, given the nature of the monthly data some of these responses may be picked up by these data (i.e. adjustments in length of stay will be taken into account). It is clear that hospitals may respond to capacity shortages by revising current scheduled admissions or discharges, however, it is equally clear that there is a limit to the extent to which hospitals can do this, limited by the number of scheduled admissions and the medical requirements of current patients. This, however, does not affect the theoretical foundation of the model developed here, although does potentially represent an interesting addition to the work carried out in this thesis.

Secondly, the availability of data restricted us to panel data analyses that can only provide results across the whole sample. Clearly differences between hospitals may be important, different hospitals may have different aims, or operate within different environments. Such differences might account for the apparently wide diversity of turnaway probabilities identified in Chapter 5. Thirdly, we have not considered the impact of competition and payment for hospital services, although these considerations may be secondary to our model since we have only assumed that elective demand is planned and that emergency demand is unplanned. We have not made any assumptions regarding the determination of demand, (our demand equations merely consider how hospitals predict future emergency demand); rather we have modelled the use of capacity and the implications for the likelihood of turning emergency patients away. Once elective output is determined, through contracts or otherwise the hospital must still determine the level of planned reserve capacity remaining to treat emergency, unplanned arrivals. This is still fundamental to the nature of hospital capacity utilisation decisions and is reinforced by the ability of hospitals to be able to sell unplanned reserve capacity at marginal costs; this reinforces the notion of hospitals planning reserve capacity levels. Finally, we have estimated a comparative static model, this, by definition, cannot deal with dynamic relationships, in particular we cannot deal with the relationship between decision periods. For example, hospitals may allow waiting lists to build in one period, when emergency demand is high, with the hope and expectation of being able treating more elective patients in subsequent periods.

The estimates of the probability of turnaway are clearly key to the whole analysis, and perhaps, represent the area most in need of advancement in the literature. It is clear that the monthly estimates presented here may not pick up the, potentially important, weekly and daily fluctuations. Consequently, any estimates of turnaway based on aggregate data can only represent an average over the period of analysis. It may well be the case that turnaway rates fluctuate daily, for example, if most elective admissions occur during the week, then this may lead to the turnaway rates for emergencies being higher in these periods. Similarly, if most admissions for emergencies occur during the daytime, this may condense arrivals into distinct times within the aggregate period of analysis. These factors may explain the identification of zero turnaway probabilities in our sample. Furthermore, if the hospital is full, this does not, in practice, always mean that the patient is turned away, rather there are other options that have been evident recently, for example, placing patients on trolleys. Nonetheless, this does often have serious implications for the patients well-being and thus, may have a significant social cost attached.

In this thesis, we attempted to identify the inherent trade-offs that hospitals face when allocating limited capacity. We have attempted to formalise this trade-off based on fundamental economic principles in order to identify a theoretical, geometric optimal solution and a mathematical solution to the problem. In addition to this, we have provided empirical content to the analysis, and shown empirically that hospital may indeed trade-off turnaway and waiting lists. As such, this represents a progression in current knowledge in this area.

APPENDIX 1: Appendix to Chapter 4

A1.1: GROUP CATEGORIES

Table A1.1.1: Site type group

Hospital type	Group number	Number of observations
--	1	60
C	2	24
N	3	1092
TL	4	120
TP	5	216

Table A1.1.2: Demand group

Demand	Group Number	Number of observations
0-140	1	156
140-170	2	216
170-200	3	192
200-230	4	180
230-260	5	108
260-290	6	252
290-320	7	108
320-350	8	36
350-380	9	72
380+	10	192

Table A1.1.3: Bed group (1)

Bed size	Group Number	Number of observations
0-500	1	60
500-600	2	84
600-700	3	264
700-800	4	216
800-900	5	96
900-1000	6	228
1000-1100	7	144
1100-1200	8	133
1200-1300	9	144
1300+	10	143

Table A1.1.4: Bed group (2)

Bed size	Group number	Number of observations
Less than 700	1	408
700-900	2	312
900-1100	3	372
1100+	4	420

A1.2: DEMAND EQUATION ESTIMATES

Table A1.2.1 Grouping by site type RE and FE

Variable	OLS		Random Effects (One way)		Fixed effects (One way)		Random Effects (Two way)		Fixed effects (Two way)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
CATCHD	1.6043**	0.1513	1.5541**	0.1531	1.5570**	0.15370	1.5540**	0.15236	1.5581**	0.15291
CATCHSH	1235.3**	110.07	1023.1**	116.26	1016.9**	116.53	1023.6**	115.67	1017.7**	115.94
BEDS	1.7243**	0.0642	1.8557**	0.0759	1.8563**	0.07721	1.8561**	0.07554	1.8555**	0.07682
FILT1993	-173.77**	34.802	-159.55**	34.559	-159.22**	34.567	-159.52**	34.383	-159.22**	34.391
FILT1994	-44.833	29.346	-47.807	29.168	-47.386	29.182	-47.819	29.020	-47.382	29.033
CONSTANT	-878.71**	112.97	-887.74**	147.01	--	--	-892.33**	152.56	-792.22**	113.26
GROUP EFFECTS										
1					-848.038**	124.565			-56.315	65.175
2					-988.543**	153.539			-197.057	115.589
3					-733.519**	115.298			58.296**	9.514
4					-870.175**	124.430			-78.345	43.559
5					-1005.718**	118.315			-213.67**	42.857
PERIOD EFFECTS										
APR									-46.289	42.174
MAY									-8.829	42.174
JUN									-2.058	42.174
JUL									25.419	42.174
AUG									-43.425	42.174
SEP									-38.826	42.174
OCT									27.347	42.174
NOV									2.562	42.174
DEC									61.304	42.174
JAN									34.615	42.174
FEB									-145.089**	42.174
MAR									133.268**	42.174
R^2	0.64439		0.626787		0.65322		0.62577		0.65927	
Log-L	-11539.0481				-11520.0315				-11506.2330	
LM			29.19[0.00]				37.10 [0.00]			
Hausman					1.56 [1.00]				1.40 [0.924]	
Autocorr	0.88540		0.140485		0.14091		0.143024		0.143370	
Hetero	0.00		22.80		0		22.78		0.00	

* significant at 10%

** significant at 5%

Table A1.2.2 Grouping by demand size RE and FE

Variable	OLS		Random Effects (One way)		Fixed effects (One way)		Random Effects (Two way)		Fixed effects (Two way)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
CATCHD	1.6043**	0.1513	-0.18707	0.22703	-0.47397*	0.23731	-0.24194	0.22751	-0.47343*	0.23572
CATCHSH	1235.3**	110.07	-14.509	162.21	-209.60	169.02	-51.834	162.45	-209.22	167.88
BEDS	1.7243**	0.0642	1.4376**	0.0660	1.4011**	0.0665	1.4299**	0.06567	1.4004**	0.06615
FILT1993	-173.77**	34.802	-184.92**	32.446	-186.09**	32.484	-185.18**	32.235	-186.09**	32.266
FILT1994	-44.833	29.346	-49.219	27.025	-49.054	27.049	-49.212	26.848	-49.052	26.867
CONSTANT	-878.71**	112.97	1082.5**	222.99	--	--	1140.6**	228.85	1327.5**	216.02
GROUP EFFECTS										
1					611.773**	177.208			-715.824**	57.764
2					1101.080**	199.625			-226.509**	39.279
3					1029.020**	201.689			-298.447**	36.705
4					1189.332**	213.903			-138.208**	33.935
5					1629.317**	209.685			301.935**	44.778
6					1334.460**	230.366			7.041	29.081
7					1922.961**	250.372			595.652**	54.721
8					1390.892**	247.142			63.414	78.348
9					1704.707**	258.485			377.470**	66.970
10					1924.222**	262.528			596.872**	53.996
PERIOD EFFECTS										
APR									-48.503	38.914
MAY									-8.628	38.912
JUN									-1.856	38.912
JUL									25.620	38.912
AUG									-43.224	38.912
SEP									-38.624	38.912
OCT									27.549	38.912
NOV									2.763	38.912
DEC									61.505	38.912
JAN									34.817	38.912
FEB									-144.888**	38.912
MAR									133.469**	38.912
R^2	0.64439		0.5178		0.70618		0.5102		0.71225	
Log-L	-11539.0481				-11394.7432				-11378.4526	
LM	--		817.60 [0.00]				825.52 [0.00]			
Hausman	--				19.53 [0.0015]				16.03[0.0068]	
Autocorr	0.88540		0.178905		0.183962		0.184372		0.188542	
Hetero	0.00		395.63**		0.00		418.16**		0.00	

* significant at 10%, ** significant at 5%

Table A1.2.3 Grouping by bed size (1) RE and FE

Variable	OLS		Random Effects (One way)		Fixed effects (One way)		Random Effects (Two way)		Fixed effects (Two way)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
CATCHD	1.6043**	0.1513	1.5010**	0.1547	1.4722**	0.15625	1.4923**	0.15446	1.4727**	0.15526
CATCHSH	1235.3**	110.07	1049.8**	109.69	1018.9**	110.37	1038.1**	109.28	1019.2**	109.67
BEDS	1.7243**	0.0642	1.5870**	0.1530	1.2158**	0.21795	1.4580**	0.17704	1.2122**	0.21659
FILT1993	-173.77**	34.802	-189.86**	33.164	-191.61**	33.204	-190.67**	32.975	-191.61**	32.995
FILT1994	-44.833	29.346	-102.47**	28.078	-111.71**	28.255	-106.05**	27.965	-111.77**	28.077
CONSTANT	-878.71**	112.97	-588.48**	178.71			-471.12**	209.27	-206.38	219.07
GROUP EFFECTS										
1					-917.142**	143.074			-710.244**	142.025
2					-475.241**	158.691			-267.507**	93.816
3					-221.043	172.272			-13.030	65.421
4					-277.115	195.435			-68.793	48.384
5					-284.941	203.641			-76.278	52.041
6					-252.406	225.682			-43.377	29.590
7					-395.276	249.507			-185.886**	47.301
8					155.097	268.496			364.780**	62.719
9					257.952	290.161			468.055**	82.538
10					-130.866	334.791			80.0677	129.320
PERIOD EFFECTS										
APR									-49.420	39.860
MAY									-8.545	39.860
JUN									-1.773	39.860
JUL									25.703	39.860
AUG									-43.141	39.860
SEP									-38.541	39.860
OCT									27.632	39.860
NOV									2.846	39.860
DEC									61.588	39.860
JAN									34.900	39.860
FEB									-144.805**	39.860
MAR									133.553**	39.860
R^2	0.64439		0.63680		0.69212		0.62867		0.69823	
Log-L	-11539.0481				-11430.0864				-11414.4259	
LM	--		1196.84				1204.75 [0.00]			
Hausman	--				10.85 [0.054447]				6.33 [0.27527]	
Autocorr	0.88540		0.026777		0.025238		0.027523		0.026588	
Hetero	0.00		26.13*		0.00		73.50**		0.00	

* significant at 10%, ** significant at 5%

Table A1.2.4 Grouping by bed size (2) RE and FE

Variable	OLS		Random Effects (One way)		Fixed effects (One way)		Random Effects (Two way)		Fixed effects (Two way)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
CATCHD	1.6043**	0.1513	1.3479**	0.15336	1.3123**	0.15418	1.3479**	0.15254	1.3127**	0.1533
CATCHSH	1235.3**	110.07	1176.8**	109.92	1163.6**	110.14	1177.0**	109.33	1164.0**	109.54
BEDS	1.7243**	0.0642	1.2379**	0.10828	1.1610**	0.11377	1.2359**	0.10775	1.1591**	0.1131
FILT1993	-173.77**	34.802	-164.02**	34.009	-163.78**	34.011	-164.00**	33.824	-163.76**	33.827
FILT1994	-44.833	29.346	-71.972**	28.921	-73.346**	28.931	-71.990**	28.764	-73.355**	28.774
CONSTANT	-878.71**	112.97	-341.59*	171.82			-340.14**	172.78	-242.94**	154.36
GROUP EFFECTS										
1					-440.087**	128.808			-196.645**	44.264
2					-354.173**	146.870			-110.439**	30.944
3					-347.718**	161.882			-103.608**	23.282
4					120.168	192.126			364.830**	49.633
PERIOD EFFECTS										
APR									-49.678	41.571
MAY									-8.521	41.571
JUN									-1.750	41.571
JUL									25.727	41.571
AUG									-43.117	41.571
SEP									-38.518	41.571
OCT									27.656	41.571
NOV									2.870	41.571
DEC									61.612	41.571
JAN									34.923	41.571
FEB									-144.781**	41.571
MAR									133.576**	41.571
R^2	0.64439		0.60196		0.66296		0.60167		0.66905	
Log-L	-11539.0481				-11498.5060				-11484.2224	
LM	--		237.82 [0.00]				245.74 [0.00]			
Hausman	--				1.16 [1.00]				7.87 [1.00]	
Autocorr	0.88540		0.102155		0.101053		0.104485		0.103393	
Hetero	0.00		163.37**		0.00		165.09**		0.00	

* significant at 10%

** significant at 5%

Table A1.2.5 Grouping by provider type RE and FE (dropping time invariant factors – i.e. CatchD and CatchSH)

Variable	OLS		Random Effects (One way)		Fixed effects (One way)		Random Effects (Two way)		Fixed effects (Two way)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
BEDS	2.2195**	0.0453	0.80694**	0.10389	0.19766	0.12229	0.75089**	0.10039	0.18543	0.1164
FILT1993	-168.12**	36.185	-110.28**	17.063	-118.34**	17.119	-111.01**	16.243	-118.50**	16.292
FILT1994	-58.867	30.477	-29.704**	13.866	-35.761**	13.908	-30.272**	13.200	-35.897**	13.236
CONSTANT	310.42**	41.854	1477.8**	103.98			1524.1**	104.69	1969.8**	96.988
GROUP EFFECTS										
1					1951.202**	110.474			-8.142	35.259
2					2997.113**	165.393			1043.468**	69.893
3					3688.554**	143.357			1732.418**	55.949
4					1319.581**	95.368			-642.862**	67.774
5					3198.514**	181.724			1245.852**	93.461
6					1693.595**	87.884			-268.179**	41.265
7					1124.599**	64.201			-839.882**	60.253
8					1314.412**	106.568			-646.625**	63.159
9					3717.735**	131.181			1759.570**	63.299
10					1927.460**	98.027			-34.641	66.401
11					1459.769**	89.984			-502.142**	48.962
12					2939.627**	158.328			984.458**	76.264
13					2094.096**	111.823			134.893**	35.406
14					770.345**	75.306			-1195.01**	84.400
15					1286.314**	84.950			-676.178**	51.976
16					832.921**	126.382			-1124.75**	40.311
17					1725.178**	94.151			-237.424**	68.461
18					1425.830**	95.368			-536.613**	67.774
19					2276.021**	91.726			314.666**	39.210
20					1686.519**	119.022			-271.924**	37.126
21					1712.006**	105.068			-249.215**	63.600
22					958.264**	106.456			-1001.81**	43.445
23					1385.254**	83.138			-577.536**	54.043
24					3184.415**	132.556			1227.150**	49.376
25					1298.377**	76.469			-665.124**	58.250
26					1789.405**	109.903			-171.228**	62.370
27					3145.204**	152.457			1189.988**	62.364
28					2200.212**	132.482			242.194**	63.662
29					3427.161**	153.338			1472.079**	63.050
30					2243.994**	114.835			283.946**	61.686

31					1404.380**	75.796			-558.744**	49.771
32					2731.859**	167.001			778.379**	71.335
33					3717.761**	210.778			1768.156**	118.435
34					1737.374**	108.090			-222.223**	35.142
35					2051.302**	135.532			93.663	64.632
36					1473.382**	83.486			-489.281**	52.955
37					3143.088**	143.927			1187.021**	56.231
38					1627.255**	78.033			-336.056**	56.985
39					1639.648**	76.513			-323.393**	49.192
40					3148.081**	135.745			1191.381**	45.935
41					1759.111**	89.708			-202.463**	40.245
42					2135.218**	106.790			175.183**	43.408
43					1987.601**	108.205			28.0149	35.144
44					2098.201**	126.616			140.551**	40.433
45					2096.455**	119.431			136.945**	61.555
46					-2.591	65.646			-1969.93**	99.158
47					2507.945**	128.481			549.475**	62.649
48					473.893**	47.307			-1494.06**	94.443
49					2629.600**	121.541			670.335**	61.654
50					2710.692**	150.994			755.365**	61.277
51					1735.096**	90.734			-226.729**	48.562
52					980.299**	75.968			-983.263**	58.666
53					2145.500**	130.640			187.275**	63.157
54					1988.935**	117.547			30.068	43.900
55					1384.849**	90.978			-576.587**	39.582
56					833.726**	69.962			-1130.07**	54.788
57					1768.808**	142.948			-187.146**	51.108
58					1589.263**	106.568			-371.775**	63.159
59					2559.506**	181.035			607.460**	84.238
60					1689.794**	86.265			-272.545**	51.135
61					2243.287**	161.721			288.485**	78.507
62					1449.436**	83.169			-514.683**	76.413
63					2592.339**	130.763			635.121**	42.757
64					1857.346**	83.225			-105.348*	53.134
PERIOD EFFECTS										
APR									-54.417**	18.163
MAY									-8.090	18.163
JUN									-1.319	18.163
JUL									26.158	18.163
AUG									-42.686**	18.163
SEP									-38.087**	18.163

OCT									28.087	18.163
NOV									3.301	18.163
DEC									62.043**	18.163
JAN									35.354	18.163
FEB									-144.350**	18.163
MAR									134.007**	18.163
R^2	0.61440		0.36475		0.93923		0.34465		0.94538	
Log-L	-11600.2650				-10203.4202				-10122.2491	
LM			13602.35 [0.00]				13608.01 [0.00]			
Hausman					89.86 [0.00]				92.55 [0.00]	
Autocorr	0.89410		0.557103		0.548860		0.665013		0.660235	
Hetero	0.00		563.73**		0.00		616.83**		0.00	

* significant at 10%

** significant at 5%

(where groups effects 1-64 correspond directly to providers Prov10-99 as ordered in Table A2.2)

Table A1.2.6 Lagged dependent variable

Variable	Lagged dependent variable	
	Coeff.	SE
FILT1993	-32.279**	12.000
FILT1994	-4.049	9.348
LAGEM	0.703**	0.018
CONSTANT	320.975**	32.541
GROUP EFFECTS		
1	354.540**	40.3398
2	696.692**	55.3057
3	962.251**	67.0482
4	170.354**	49.7756
5	834.086**	68.4040
6	274.745**	37.5010
7	76.8106**	33.9766
8	163.422**	49.8627
9	925.235**	74.1123
10	357.646**	53.1849
11	204.564**	39.7703
12	688.005**	64.6857
13	402.660**	41.9255
14	-8.4273	48.9076
15	151.028**	38.8404
16	38.2083	33.5825
17	287.011**	51.7578
18	205.214**	50.1631
19	463.057**	43.4966
20	289.340**	37.8676
21	292.372**	51.8041
22	60.3833	37.9785
23	170.688**	39.2682
24	752.587**	59.7233
25	146.328**	38.8269
26	307.582**	52.4557
27	751.166**	59.5708
28	454.110**	56.0042
29	820.543**	63.6976
30	456.167**	56.1908
31	193.608**	35.2409
32	639.837**	51.1423
33	847.529**	76.8715
34	295.587**	38.1936
35	394.506**	54.7958
36	209.256**	39.8064
37	769.011**	59.3691
38	259.911**	40.6138
39	244.274**	36.9301
40	738.352**	56.7075

41	289.260**	38.1138
42	413.011**	45.7203
43	376.392**	40.6509
44	394.529**	42.3251
45	404.505**	54.8709
46	-263.702**	51.69166
47	534.9693**	59.09236
48	-123.175**	38.6781
49	573.705**	60.2086
50	602.549**	53.6297
51	284.361**	41.6943
52	49.8991	37.9220
53	409.645**	55.6270
54	383.998**	44.3554
55	206.476**	35.2343
56	--	--
57	321.227**	39.0508
58	257.504**	51.1511
59	578.581**	49.2415
60	269.371**	41.2652
61	477.117**	56.9353
62	190.807**	50.1896
63	561.236**	48.5664
64	291.2373**	42.7889
PERIOD EFFECTS		
APR	-167.847**	21.995
MAY	-28.852	17.711
JUN	-55.351**	17.678
JUL	-32.635	17.677
AUG	-120.802**	17.682
SEP	-67.790**	17.698
OCT	-4.850	17.694
NOV	-76.170**	17.683
DEC	--	--
JAN	-67.996**	17.711
FEB	-228.9337**	17.687
MAR	175.797**	17.896
R^2	0.97348	
F	653.254**	
Durbin's h-test	34.42**	
Hetero	199.00**	

* significant at 10%

** significant at 5%

SPSS does not allow constant term to adjust for dummy variables that are collinear with constant term therefore two dummy variables are dropped from this equation (it was necessary to use SPSS as the lagged dependent variable model involved dropping the first observation for each provider).

Table A1.2.7 AR(1)

Variable	AR(1)	
	Coeff.	SE
FILT1993	-107.33**	34.540
FILT1994	-28.613	21.567
RHO	0.660691	
GROUP EFFECTS		
1	2145.644**	73.11833
2	3296.956**	73.11833
3	4146.568**	88.59267
4	1514.729**	125.5673
5	3725.586**	125.5673
6	1873.095**	73.11833
7	1214.543**	73.11833
8	1497.278**	125.5673
9	4046.285**	125.5673
10	2143.226**	125.5673
11	1641.515**	87.85266
12	3258.703**	125.5673
13	2305.640**	73.11833
14	913.1346**	125.5673
15	1460.848**	87.85266
16	1076.003**	73.11833
17	1908.935**	125.5673
18	1629.846**	125.5673
19	2500.474**	73.11833
20	1921.439**	73.11833
21	1925.664**	125.5673
22	1158.214**	87.85266
23	1525.675**	90.92934
24	3481.447**	87.85266
25	1446.601**	87.85266
26	1982.923**	125.5673
27	3470.806**	90.92934
28	2469.752**	125.5673
29	3722.271**	87.85266
30	2478.332**	125.5673
31	1594.470**	73.11833
32	3093.821**	73.11833
33	3841.279**	125.5673
34	1944.436**	73.11833
35	2277.059**	125.5673
36	1656.112**	87.85266
37	3526.811**	87.85266
38	1821.203**	87.85266
39	1774.550**	73.11833
40	3431.340**	73.11833
41	1924.680**	73.11833
42	2344.564**	87.85266
43	2213.726**	73.11833
44	2286.056**	73.11833
45	2307.627**	125.5673

46	61.34295	125.5673
47	2746.962**	125.5673
48	540.1586**	87.85266
49	2874.035**	125.5673
50	2984.911**	87.85266
51	1911.472**	87.85266
52	1122.825**	87.85266
53	2333.527**	125.5673
54	2242.380**	73.11833
55	1632.252**	73.11833
56	952.165**	73.11833
57	2031.471**	73.11833
58	1809.093**	125.5673
59	2896.347**	73.11833
60	1860.370**	87.85266
61	2549.882**	125.5673
62	1588.162**	125.567
63	2839.079**	73.118
64	1948.049**	87.852
PERIOD EFFECTS		
APR	-108.91**	27.992
MAY	-64.604**	24.667
JUN	-59.819**	23.042
JUL	-33.653	22.063
AUG	-103.36**	21.123
SEP	-99.338**	19.808
OCT	-33.541	17.655
NOV	-58.576**	13.724
DEC	--	--
JAN	-26.798*	13.721
FEB	-206.57**	17.636
MAR	71.736**	19.738
R^2	0.82276	
Log-Likelihood	-9176.5572	
F	83.74**	
Autocorr.	-0.080802	
White χ^2 (78 dof)	86.18	

* significant at 10%

** significant at 5%

A1.3: ESTIMATED RESIDUALS

Table A1.3.1: Estimated residuals

SITEID	YEAR	EMADM	RES (eq 4.18)	RES (eq 4.17)	RES (eq 4.19)
Prov10	1993	26821	1257	3323	3084
Prov10	1994	24048	1291	1604	1485
Prov10	1995	23541	761	2403	1484
Prov100	1993	37867	712	1960	741
Prov100	1994	39569	1175	2661	1592
Prov100	1995	37943	716	2910	1073
Prov103	1995	50914	1675	12400	2222
Prov104	1995	17258	553	7779	691
Prov106	1995	41703	1314	3520	2281
Prov107	1994	21808	728	1882	900
Prov107	1995	21174	492	769	642
Prov111	1994	15841	855	2801	2341
Prov112	1995	17469	315	6890	402
Prov114	1995	46866	1161	21107	1428
Prov115	1995	24619	689	1741	881
Prov117	1994	18202	561	1074	625
Prov117	1995	19430	531	1841	583
Prov120	1995	38110	1020	2729	900
Prov121	1995	26394	940	899	1039
Prov121	1994	27147	665	1703	731
Prov122	1995	10103	661	876	730
Prov124	1995	16981	603	3958	455
Prov130	1993	11429	316	13090	712
Prov130	1994	11800	745	13678	649
Prov131	1995	22094	548	1057	411
Prov133	1995	18533	620	1812	741
Prov15	1995	31637	2500	7695	3004
Prov16	1995	22292	871	3279	722
Prov16	1994	22015	916	1694	899
Prov17	1995	22204	748	1107	654
Prov20	1994	13171	691	6853	456
Prov20	1995	13139	583	10674	552
Prov22	1994	17034	729	1719	626
Prov22	1993	17017	948	1225	1153
Prov24	1994	37486	1137	7163	3224
Prov24	1995	43340	1768	11150	22286
Prov26	1994	16598	663	1850	663
Prov26	1995	16544	706	3274	503
Prov27	1995	23247	548	562	549
Prov3	1994	40955	1408	3584	1150
Prov30	1995	28684	791	2096	678
Prov31	1995	43669	1104	3504	1035
Prov31	1994	43830	1423	9014	1152
Prov32	1995	28816	633	8481	606
Prov37	1995	18937	593	974	637
Prov38	1995	37934	1789	3063	1702
Prov38	1994	35249	933	1500	1169
Prov39	1995	48527	2864	6067	4365

Prov4	1995	23080	622	779	724
Prov45	1995	18527	1111	2474	1690
Prov45	1994	19140	604	1909	1428
Prov48	1995	41354	994	2353	1159
Prov49	1995	21979	1433	2406	1118
Prov49	1994	19130	1097	1905	1218
Prov5	1994	20066	487	1632	327
Prov5	1995	19796	760	726	956
Prov50	1995	41651	904	5022	1428
Prov51	1993	22692	1058	2482	1728
Prov51	1994	22252	863	1214	545
Prov51	1995	21253	730	1295	1134
Prov54	1994	24769	778	1149	925
Prov54	1995	27165	629	1240	1350
Prov55	1994	24243	1934	10139	3645
Prov55	1995	23298	1048	12317	3411
Prov56	1995	27150	2144	2657	1953
Prov59	1995	443	662	3418	691
Prov6	1995	32289	820	6659	634
Prov60	1995	5862	580	1001	434
Prov60	1994	5766	561	1808	568
Prov62	1995	33595	1533	5466	1262
Prov64	1994	35546	1165	1306	1390
Prov66	1995	21919	319	4642	388
Prov67	1995	12593	559	441	547
Prov67	1994	12891	601	1610	615
Prov69	1995	27987	615	883	861
Prov75	1994	23878	791	1165	1964
Prov75	1995	27635	873	1650	1450
Prov78	1994	18850	3291	7411	7715
Prov78	1995	24071	1420	1240	5207
Prov82	1995	8854	958	6203	1848
Prov82	1993	11063	488	2424	1648
Prov82	1994	11673	864	2773	1314
Prov85	1995	23116	517	9345	922
Prov85	1994	23811	723	10353	790
Prov85	1993	22890	673	10666	1101
Prov86	1995	20767	499	2529	452
Prov88	1993	30026	3862	6771	4866
Prov88	1995	36030	1234	2984	2093
Prov88	1994	34500	2549	3902	2718
Prov89	1995	21934	803	960	733
Prov90	1995	29825	778	4252	584
Prov92	1995	18491	847	4662	650
Prov94	1993	33411	932	6720	1353
Prov94	1994	32518	806	2918	809
Prov94	1995	32800	776	5041	779
Prov99	1995	20342	1432	796	2311

A1.4: COST EQUATION ESTIMATES

Table A1.4.1 Grouping by bed size (1)

Variable	OLS		Random Effects (One way)		Fixed effects (One way)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE
EMADM	44.037	361.90	113.72	353.65	227.41	359.99
CASEMIX	0.46013E+08**	0.21752E+08	0.49968E+08**	0.21933E+08	0.60583E+08**	0.23112E+08
BEDSEM	30445**	13186	36627**	13456	46865**	14764
DVTEACH	0.47421E+08**	0.75997E+07	0.51299E+08**	0.77107E+07	0.52868E+08**	0.78330E+07
RES	1477.0	619.14	1440.5**	604.32	1526.8**	609.86
INVOCC	0.26354E+06	0.32349E+07	0.80786E+06	0.31499E+07	0.15557E+07	0.31788E+07
AEATT	0.20236E+06**	92654	0.19881E+06**	91558	0.21807E+06**	93262
DAYATT	344.18**	135.47	380.16**	135.35	450.59**	142.40
ELADM	648.99**	239.74	727.79**	240.87	880.50**	258.53
OPATT	94257**	28907	83815**	28598	75791**	28956
WAGE INDEX	-0.79573E+06	0.78476E+06	-0.97488E+06	0.76950E+06	-0.12097E+07**	0.78165E+06
CONSTANT	-0.20868E+08	0.20957E+08	-0.27444E+08	0.22060E+08		
GROUP EFFECTS						
1					-35149704	22760884
2					-39411341	23719115
3					-53169535**	25526800
4					-54048036	28794306
R^2	0.87189		0.867110		0.88478	
Log-Likelihood	-1509.7207				-1505.2154	
LM			0.00 [0.992853]			
Hausman					--	
Autocorr.	0.11319		0.004614		0.000497	
Hetero	0.00		4.10*		0.00	

* significant at 10%

** significant at 5%

Table A1.4.2 Grouping by bed size (2)

Variable	OLS		Random Effects (One way)		Fixed effects (One way)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE
EMADM	44.037	361.90	45.308	377.32	42.204	384.73
CASEMIX	0.46013E+08**	0.21752E+08	0.52026E+08**	0.22765E+08	0.53489E+08**	0.25328E+08
BEDSEM	30445	13186	30759**	16131	31621	21632
DVTEACH	0.47421E+08**	0.75997E+07	0.50371E+08**	0.79874E+07	0.50452E+08**	0.80913E+07
RES	1477.0**	619.14	1796.8**	635.95	1837.7**	642.12
INVOCC	0.26354E+06	0.32349E+07	0.55278E+06	0.36069E+07	0.39079E+06	0.37891E+07
AEATT	0.20236E+06**	92654	0.19121E+06**	91618	0.19114E+06**	92318
DAYATT	344.18**	135.47	408.62**	139.75	419.17**	148.07
ELADM	648.99**	239.74	796.40**	259.81	818.04**	286.86
OPATT	94257**	28907	73078**	29766	70723**	30045
WAGE INDEX	-0.79573E+06	0.78476E+06	-0.88915E+06	0.79252E+06	-0.87697E+06	0.79971E+06
CONSTANT	-0.20868E+08	0.20957E+08	-0.24995E+08	0.24947E+08		
GROUP EFFECTS						
1					-23461506	24004917
2					-29807444	27121481
3					-20856004	27735573
4					-21044218	28789019
5					-27140327	30512235
6					-32798320	31807612
7					-31159264	34772418
8					-37407748	36690608
9					-26809144	38593452
10					-16853644	41312613
R^2	0.87189		0.868153		0.89930	
Log-Likelihood	-1509.7207				-1499.4910	
LM			4.03 [0.144]			
Hausman					--	
Autocorr.	0.11319		0.00424		0.000039	
Hetero	0.00		3.80*		0.00	

* significant at 10%

** significant at 5%

Table A1.4.3 Grouping by demand size

Variable	OLS		Random Effects (One way)		Fixed effects (One way)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE
EMADM	44.037	361.90	61.433	394.30	77.323	407.11
CASEMIX	0.46013E+08**	0.21752E+08	0.46342E+08**	0.22834E+08	0.48126E+08**	0.23098E+08
BEDSEM	30445**	13186	30909**	13908	32125**	14072
DVTEACH	0.47421E+08**	0.75997E+07	0.49024E+08**	0.81177E+07	0.50090E+08**	0.83494E+07
RES	1477.0**	619.14	1474.4**	649.93	1500.8**	653.68
INVOCC	0.26354E+06	0.32349E+07	93799	0.34902E+07	10923	0.35209E+07
AEATT	0.20236E+06	92654	0.23545E+06**	99039	0.24444E+06**	0.10025E+06
DAYATT	344.18**	135.47	388.80**	160.03	377.17**	168.16
ELADM	648.99**	239.74	551.49**	273.93	501.29**	284.93
OPATT	94257**	28907	91783**	32013	89556**	32529
WAGE INDEX	-0.79573E+06	0.78476E+06	-0.90445E+06	0.83028E+06	-0.88753E+06	0.83912E+06
CONSTANT	-0.20868E+08	0.20957E+08	-0.20045E+08	0.23494E+08		
GROUP EFFECTS						
1					-21200173	25459961
2					-17167114	23024732
3					-25908716	23774947
4					-23040782	23653145
5					-25417271	25117506
6					-23912354	23313275
7					-11163648	25795771
8					-29116083	30635359
9					-21497398	25411867
10					-18316249	24669629
R^2	0.87189		0.870321		0.88071	
Log-Likelihood	-1509.7207				-1506.6911	
LM			0.44 [0.508064]			
Hausman					--	
Autocorr.	0.11319		0.00378		-0.023115	
Hetero	0.00		4.17*		0.00	

* significant at 10%

** significant at 5%

Table A1.4.4 Grouped by sitetype (have to drop the teaching filter for this one)

Variable	OLS		Random Effects (One way)		Fixed effects (One way)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE
EMADM	-769.37	415.22	-10.943	381.62	81.169	386.60
CASEMIX	0.55566E+08**	0.26687E+08	0.46960E+08**	0.23025E+08	0.45771E+08**	0.23205E+08
BEDSEM	25157	16184	31774**	14602	30956**	14856
RES	1172.0	759.10	1390.9**	658.79	1394.7**	661.50
INVOCC	-0.46355E+07	0.38596E+07	-0.72501E+06**	0.35117E+07	44574**	0.35751E+07
AEATT	0.25705E+06**	0.11344E+06	0.20198E+06**	96007	0.20132E+06**	96308
DAYATT	173.55	163.18	340.71**	139.39	357.26**	139.77
ELADM	976.04**	287.72	643.50**	262.17	603.04**	264.75
OPATT	0.16457E+06**	32740	99249**	29511	92463**	29769
WAGE INDEX	0.12216E+07**	0.87946E+06	-0.51073E+06	0.82299E+06	-0.73605E+06	0.84098E+06
CONSTANT	-0.52090E+08	0.25030E+08	-0.14596E+08	0.25322E+08		
GROUP EFFECTS						
1					-20515481	25881037
2					-17622133	22149179
3					-21793442	23650638
4					26747394	26928999
5					-18893880	25537966
R^2	0.80356		0.705186E		0.87266	
Log-Likelihood	-1527.8881				-1509.4642	
LM			0.73 [0.3922]			
Hausman					--	
Autocorr.	0.07538		0.0274		-0.034523	
Hetero	4.64*		3.98*		0.00	

* significant at 10%

** significant at 5%

Table A1.4.5 Translog cost function

Dependent Variable LNTVC

Variable	B	SE B	T
LNAEATT	.006851	.014583	.470
LNBDEMSQ	.763963	.576606	1.325
LNBEDEM	-3.480766	2.879319	-1.209
LNCASEMX	.080215	.24257	6.331**
LNDAYATT	.005274	.00993	0.531
LNELBDEM	8.79364E-04	6.9827E-04	1.259
LNELSQ	.171656	.239207	.718
LNEMBDEM	-9.95776E-04	7.2680E-04	-1.370
LNEMADM	-1.744234	2.876930	-.606
LNEMSQ	.159581	.156849	1.017
FILTEACH	.521721	.154617	3.374**
LNRESTOT	.015791	.038871	.406
LNWI	-.093079	6.458856	.014
LNWISQ	1.675865	1.976024	.848
LNOPATT	.101217	.342408	.296
LNEMOPAT	1.18778E-06	1.9863E-06	.598
LNELOPAT	-1.15513E-06	1.9285E-06	-.599
LNEMWI	.152897	.752124	.203
LNELWI	-.626549	.771399	-.812
LNINVOCC	.283866	.281415	1.009
LNELADM	2.650052	.067482	.537
LNELEM	5.168743	.079206	.631
LNOPATTSQ	-7.651705	-.072376	-.576
CONSTANT	31.753440	21.153743	1.501

R^2 0.85133
F 18.3242**
N 85

* significant at 10%
** significant at 5%

Table A1.4.6 Collinearity Diagnostics

Number	Eigenval	Cond Index	Variance Proportions Constant	LNAEATT	LNBDEMSQ	LNBEDEM	LNCASEMX	LNDAYATT
1	19.02329	1.000	.00000	.00005	.00000	.00000	.00013	.00016
2	1.06131	4.234	.00000	.00000	.00000	.00000	.00057	.00572
3	.42375	6.700	.00000	.00097	.00000	.00000	.00322	.00079
4	.17254	10.500	.00000	.00299	.00000	.00000	.00023	.00679
5	.11154	13.059	.00000	.00628	.00000	.00000	.01319	.54362
6	.10069	13.745	.00000	.01157	.00000	.00000	.37609	.00524
7	.05927	17.915	.00000	.02095	.00000	.00000	.06137	.00222
8	.02824	25.955	.00000	.52893	.00000	.00000	.16076	.01882
9	.00864	46.934	.00000	.00805	.00000	.00000	.00073	.00051
10	.00573	57.642	.00001	.02826	.00005	.00002	.00091	.02821
11	.00253	86.763	.00000	.00917	.00002	.00000	.15327	.08381
12	.00141	116.298	.00031	.05808	.00003	.00003	.00057	.02387
13	.00080	153.803	.00010	.02539	.00221	.00022	.02901	.15005
14	.00018	325.058	.00034	.01233	.00068	.00048	.01691	.02871
15	.00007	505.089	.00003	.02412	.00022	.00000	.04831	.03144
16	.00001	1287.074	.00000	.13247	.00006	.00106	.00475	.00007
17	.00001	1776.129	.00003	.05123	.00027	.00001	.00046	.00286
18	.00000	3105.983	.00047	.03509	.00098	.00190	.09602	.00572
19	.00000	4003.113	.00253	.00431	.48170	.48155	.02160	.00007
20	.00000	5045.030	.23977	.02347	.48972	.49980	.00244	.04884
21	.00000	8745.340	.75642	.01631	.02406	.01494	.00945	.01247

	LNELBDEM	LNELSQ	LNEMBDEM	LNEMADM	LNEMSQ	FILTEACH	LNRESTOT	LNWI
1	.00000	.00000	.00000	.00000	.00000	.00007	.00002	.00000
2	.00000	.00000	.00000	.00000	.00000	.19347	.00001	.00000
3	.00001	.00000	.00001	.00000	.00000	.05718	.00009	.00000
4	.00000	.00000	.00000	.00000	.00000	.00668	.00027	.00000
5	.00000	.00000	.00000	.00000	.00000	.14597	.00039	.00000
6	.00000	.00000	.00000	.00000	.00000	.02917	.00065	.00000
7	.00008	.00000	.00006	.00000	.00000	.04811	.00082	.00000
8	.00002	.00000	.00001	.00000	.00000	.01682	.00000	.00000
9	.00000	.00000	.00000	.00000	.00000	.05266	.62650	.00000
10	.00011	.00000	.00004	.00001	.00017	.00872	.06722	.00001
11	.00023	.00072	.00047	.00000	.00015	.00017	.05121	.00000
12	.00001	.00012	.00028	.00000	.00082	.01244	.10934	.00008
13	.00009	.00001	.00039	.00007	.00095	.07232	.00460	.00001
14	.00049	.00019	.00161	.00004	.00001	.00407	.03378	.00011
15	.06454	.00422	.04575	.00042	.00231	.06848	.01802	.00002
16	.31544	.00314	.29818	.00002	.00153	.13856	.00839	.00000
17	.08241	.00063	.08394	.00156	.33978	.00640	.00628	.00008
18	.20156	.08503	.19352	.07380	.36801	.07713	.01828	.19766
19	.02866	.12998	.00333	.01884	.09263	.02478	.03798	.22805
20	.27644	.29645	.35620	.00078	.06530	.03591	.00290	.31071
21	.02992	.47950	.01620	.90445	.12833	.00091	.01325	.26327

	LNWISQ	LNINVOCC	LNEMOPAT	LNELOPAT	LNEMWI	LNELWI	LNOPATTSQ
1	.00000	.00005	.00000	.00000	.00000	.00000	.00000
2	.00000	.00035	.00000	.00000	.00000	.00000	.00000
3	.00000	.00216	.00001	.00001	.00000	.00000	.00000
4	.00000	.11165	.00000	.00000	.00000	.00000	.00000
5	.00000	.00002	.00000	.00000	.00000	.00000	.00000
6	.00000	.00097	.00000	.00000	.00000	.00000	.00001
7	.00000	.00720	.00008	.00008	.00000	.00000	.00001
8	.00000	.01934	.00002	.00002	.00000	.00000	.00001
9	.00007	.00001	.00001	.00001	.00000	.00000	.00000
10	.00039	.01303	.00001	.00001	.00000	.00001	.00048

11	.00014	.01055	.00033	.00049	.00002	.00002	.00001
12	.00000	.16951	.00002	.00002	.00010	.00002	.00332
13	.00012	.09789	.00006	.00005	.00003	.00000	.00330
14	.00000	.01698	.00193	.00812	.00001	.00001	.40470
15	.00020	.17403	.04019	.03376	.00135	.00033	.10793
16	.00298	.00439	.67949	.69216	.00000	.00142	.10047
17	.19158	.15971	.11883	.11640	.05218	.01692	.04330
18	.64562	.05856	.00003	.00053	.00001	.04513	.06592
19	.01201	.02881	.00032	.00001	.00055	.11953	.05847
20	.03698	.10955	.05206	.05426	.02622	.24548	.00518
21	.10991	.01524	.10661	.09407	.91953	.57113	.20688

Table A1.4.7: Correlation Coefficients

- - Correlation Coefficients - -

	LNAEATT	LNBDEMSQ	LNBEDEM	LNCASEMX	LNDAYATT	LNELADM
LNAEATT	1.0000	.1843	.1970	-.4172**	-.0227	.1188
LNBDEMSQ	.1843	1.0000	.9969**	-.1013	.2350*	.7383**
LNBEDEM	.1970	.9969**	1.0000	-.1101	.2611*	.7294**
LNCASEMX	-.4172**	-.1013	-.1101	1.0000	-.1939	.1793
LNDAYATT	-.0227	.2350*	.2611*	-.1939	1.0000	-.1071
LNELADM	.1188	.7383**	.7294**	.1793	-.1071	1.0000
LNELBDEM	.1313	.9369**	.9081**	-.0467	.0713	.7682**
LNELOPAT	.0805	.6984**	.6755**	.0646	-.1725	.7181**
LNELSQ	.1152	.7287**	.7185**	.1844	-.1198	.9996**
LNELWI	.1523	.5314**	.5299**	.1429	-.1733	.8096**
LNEMBDEM	.1569	.9421**	.9142**	-.0835	.0944	.7375**
LNEMOPAT	.0921	.7047**	.6826**	.0379	-.1522	.6972**
LNEMSQ	.4256**	.8027**	.8150**	-.2185*	.1433	.7513**
LNEMADM	.4451**	.7890**	.8061**	-.2444*	.1687	.7175**
LNEMWI	.3917**	.6423**	.6626**	-.1904	.0686	.6550**
LNOPATT	.1590	.8154**	.8094**	-.0407	-.0329	.7592**
LNOPATTSQ	.1506	.8073**	.7996**	-.0318	-.0467	.7564**
LNRESTOT	.2004	.3350**	.3272**	-.1100	.0851	.3160**
LNWI	.0136	-.0511	-.0424	.0294	-.1381	.0707
LNWISQ	.0138	-.0565	-.0479	.0381	-.1469	.0666
LNELEM	.3072**	.8203**	.8228**	-.0344	.0222	.9284**
LNINVOCC	.4358**	.1257	.1603	-.1846	.0267	.3352**
FILTEACH	.1010	.0127	.0170	.1599	-.5367**	.2234*

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

- - Correlation Coefficients - -

	LNELBDEM	LNELOPAT	LNELSQ	LNELWI	LNEMBDEM	LNEMOPAT
LNAEATT	.1313	.0805	.1152	.1523	.1569	.0921
LNBDEMSQ	.9369**	.6984**	.7287**	.5314**	.9421**	.7047**
LNBEDEM	.9081**	.6755**	.7185**	.5299**	.9142**	.6826**
LNCASEMX	-.0467	.0646	.1844	.1429	-.0835	.0379
LNDAYATT	.0713	-.1725	-.1198	-.1733	.0944	-.1522
LNELADM	.7682**	.7181**	.9996**	.8096**	.7375**	.6972**
LNELBDEM	1.0000	.7676**	.7660**	.5362**	.9969**	.7673**
LNELOPAT	.7676**	1.0000	.7187**	.6233**	.7591**	.9982**
LNELSQ	.7660**	.7187**	1.0000	.8105**	.7346**	.6972**
LNELWI	.5362**	.6233**	.8105**	1.0000	.5069**	.6020**
LNEMBDEM	.9969**	.7591**	.7346**	.5069**	1.0000	.7634**
LNEMOPAT	.7673**	.9982**	.6972**	.6020**	.7634**	1.0000
LNEMSQ	.7157**	.6280**	.7426**	.5761**	.7347**	.6445**
LNEMADM	.6822**	.5937**	.7075**	.5415**	.7015**	.6107**
LNEMWI	.5294**	.5604**	.6471**	.8148**	.5362**	.5682**
LNOPATT	.7929**	.9343**	.7526**	.6843**	.7922**	.9391**
LNOPATTSQ	.7935**	.9450**	.7504**	.6805**	.7924**	.9494**
LNRESTOT	.3539**	.4072**	.3159**	.2430*	.3615**	.4140**
LNWI	-.0822	.1054	.0723	.6417**	-.0931	.0953
LNWISQ	-.0864	.1074	.0684	.6386**	-.0976	.0969
LNELEM	.7881**	.7154**	.9235**	.7278**	.7826**	.7137**
LNINVOCC	.0156	.1698	.3290**	.2875**	.0372	.1863
FILTEACH	.0186	.3478**	.2259*	.4278**	-.0021	.3274**

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

- - Correlation Coefficients - -

	LNEMSQ5	LNEMADM	LNEMWI	LNOPATT	LNOPATTSQ	LNRESTOT
LNAEATT	.4256**	.4451**	.3917**	.1590	.1506	.2004
LNBDEMSQ	.8027**	.7890**	.6423**	.8154**	.8073**	.3350**
LNBEDEMT	.8150**	.8061**	.6626**	.8094**	.7996**	.3272**
LNCASEMX	-.2185*	-.2444*	-.1904	-.0407	-.0318	-.1100
LNDAYATT	.1433	.1687	.0686	-.0329	-.0467	.0851
LNELADM	.7513**	.7175**	.6550**	.7592**	.7564**	.3160**
LNELBDEM	.7157**	.6822**	.5294**	.7929**	.7935**	.3539**
LNELOPAT	.6280**	.5937**	.5604**	**.9343****	**.9450****	.4072**
LNELSQ	.7426**	.7075**	.6471**	.7526**	.7504**	.3159**
LNELWI	.5761**	.5415**	.8148**	.6843**	.6805**	.2430*
LNEMBDEM	.7347**	.7015**	.5362**	.7922**	.7924**	.3615**
LNEMOPAT	.6445**	.6107**	.5682**	**.9391****	**.9494****	.4140**
LNEMSQ	1.0000	**.9947****	.8283**	.7631**	.7522**	.3521**
LNEMADM	.9947**	1.0000	.8281**	.7419**	.7295**	.3287**
LNEMWI	.8283**	.8281**	1.0000	.7219**	.7090**	.2540*
LNOPATT	.7631**	.7419**	.7219**	1.0000	**.9995****	.4227**
LNOPATT5	.7522**	.7295**	.7090**	**.9995****	1.0000	.4239**
LNRESTOT	.3521**	.3287**	.2540*	.4227**	.4239**	1.0000
LNSALPC	-.0267	-.0396	.5261**	.1629	.1592	-.0486
LNSALPCSQ	-.0341	-.0468	.5197**	.1621	.1587	-.0513
LNELEM	**.9423****	.9236**	.7934**	.8101**	.8027**	.3556**
LNINVOCC	.6472**	.6612**	.5669**	.2685**	.2581*	.1452
FILTEACH	.0706	.0705	.2930**	.3075**	.3126**	-.0305

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

- - Correlation Coefficients - -

	LNWI	LNWISQ	LNELEM	LNINVOCE	FILTEACH
LNAEATT	.0136	.0138	.3072**	.4358**	.1010
LNBDEMSQ	-.0511	-.0565	.8203**	.1257	.0127
LNBEDEM	-.0424	-.0479	.8228**	.1603	.0170
LNCASEMX	.0294	.0381	-.0344	-.1846	.1599
LNDAYATT	-.1381	-.1469	.0222	.0267	-.5367**
LNELADM	.0707	.0666	**.9284****	.3352**	.2234*
LNELBDEM	-.0822	-.0864	.7881**	.0156	.0186
LNELOPAT	.1054	.1074	.7154**	.1698	.3478**
LNELSQ	.0723	.0684	.9235**	.3290**	.2259*
LNELWI	.6417**	.6386**	.7278**	.2875**	.4278**
LNEMBDEM	-.0931	-.0976	.7826**	.0372	-.0021
LNEMOPAT	.0953	.0969	.7137**	.1863	.3274**
LNEMSQ	-.0267	-.0341	**.9423****	.6472**	.0706
LNEMTADM	-.0396	-.0468	**.9236****	.6612**	.0705
LNEMWI	.5261**	.5197**	.7934**	.5669**	.2930**
LNOPATT	.1629	.1621	.8101**	.2685**	.3075**
LNOPATTSQ	.1592	.1587	.8027**	.2581*	.3126**
LNRESTOT	-.0486	-.0513	.3556**	.1452	-.0305
LNWI	1.0000	**.9992****	.0172	-.0094	.4039**
LNWISQ	**.9992****	1.0000	.0111	-.0145	.4195**
LNELEM	.0172	.0111	1.0000	.5366**	.1557
LNINVOCC	-.0094	-.0145	.5366**	1.0000	.1532
FILTEACH	.4039**	.4195**	.1557	.1532	1.0000

* - Signif. LE .05 ** - Signif. LE .01 (2-tailed)

Table A1.4.8: Cobb-Douglas cost function

Variable	B	SE	t-stat
EMADM	-0.14183	0.16046	-0.884
CASEMIX	0.11644	0.20641	0.564
BEDSEM	0.40048	0.17285	2.317**
RES	0.16757E-01	0.34614E-01	0.484
INVOCC	-0.73827E-01	0.20458	-0.361
AEATT	0.14318E-01	0.28752E-01	0.498
DAYATT	-0.23879E-03	0.86347E-02	-0.028
ELADM	0.30349	0.82392E-01	3.683**
OPATT	0.25094	0.10202	2.460**
WAGE INDEX	0.91879E-01	0.20101	0.457
DVTEACH	0.39759	0.11199	3.550**
CONSTANT	-1.6600	1.0578	-1.569
Log-Likelihood	-335.402		
R^2	0.87189		
Hetero.	0.00		
N	85		

* significant at 10%

** significant at 5%

Table A1.4.9: Log-likelihood grid search values

	λ										
θ	**-1.0**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**1.0**	**1.25**	**1.5**	**1.75**
-1.0	-335.18	-353.13	-351.95	-352.03	-354.30	-357.86	-360.69	-364.36	-368.45	-372.51	-376.32	-379.79
-0.75	-350.87	-348.67	-346.86	-346.06	-347.21	-349.11	-350.65	-353.44	-356.99	-360.79	-364.52	-368.04
-0.5	-348.43	-345.83	-343.45	-341.82	-341.80	-342.05	-342.22	-343.97	-346.81	-350.22	-353.80	-351.32
-0.25	-347.63	-344.72	-341.84	-339.46	-338.31	-336.95	-335.72	-336.26	-338.91	-341.02	-344.32	-347.76
0	-348.55	-345.41	-342.16	-339.14	**-336.94**	-334.14	-331.51	-330.70	-331.50	-333.54	-336.37	-339.57
0.25	-351.29	-348.03	-344.53	-341.03	-337.95	-333.92	-330.02	-327.80	-327.27	**-328.25**	-330.34	-333.11
0.5	-355.94	-352.67	-349.06	-345.25	-341.53	-336.58	-331.65	-328.09	**-326.13**	**-325.80**	**-326.86**	-328.89
0.75	-362.57	-359.39	-355.81	-351.89	-347.78	-342.30	-336.67	-332.04	-328.71	**-326.95**	-326.68	-327.62
1.0	-	-368.21	-364.79	-360.93	-356.70	-351.07	-345.14	-339.80	**-335.40**	-332.28	-330.56	-330.11

Table A1.4.10: Box-Cox cost function

Variable	Box-Cox model with teaching dummy (λ=1.2112, θ=0.53812)		
	Coeff.	SE	t-stat
EMADM	0.17291E-05	0.47135E-05	0.367
CASEMIX	7.1242	2.8144	2.531**
BEDSEM	0.15571E-02	0.37484E-03	4.154**
RES	0.28872E-04	0.11057E-04	2.611**
INVOCC	-0.39541	0.29344	-1.348
AEATT	0.88775E-02	0.49121E-02	1.807*
DAYATT	0.71681E-05	0.20272E-05	3.536**
ELADM	0.86015E-05	0.32872E-05	2.617**
OPATT	0.26963E-02	0.10556E-02	2.554**
WAGE INDEX	0.41711E-02	0.50607E-01	0.082
DVTEACH	5.7278	0.90725	6.313**
CONSTANT	9.9175	1.6035	6.185**
R^2	0.99789		
Log-Likelihood	-325.71		
Hetero	0.44		

* significant at 10%

** significant at 5%

Table A1.4.11: Elasticities and marginal costs

Variable	Coeff.	Mean	Elasticity	Marginal cost
EMADM	0.17291E-05	25026	0.0381	£110.97
CASEMIX	7.1242	0.6937	0.4726	£49,660,433.30
BEDSEM	0.15571E-02	645	0.4084	£45,617.05
RES	0.28872E-04	1536	0.0217	£1,029.81
INVOCC	-0.39541	1.49	-0.0797	-£3,358,154.71
AEATT*	0.88775E-02	59137	0.1290	£159.01
DAYATT	0.71681E-05	11111	0.0591	£387.72
OPATT*	0.26963E-02	190243	0.1612	£61.77
WAGE INDEX*	0.41711E-02	19061	0.0154	£58.89
ELADM	0.86015E-05	24672	0.1863	£550.42
TVC		72.89345		

All calculations are re-adjusted by scaling factor for TVC of 1,000,000

* re-adjusted for further scaling factor of 1,000

Table A1.4.12: Linear and Log-linear estimates

	Linear model (λ=1, θ=1)			Log-linear model (λ=0, θ=0)		
	Coeff.	SE	t-stat	Coeff.	SE	t-stat
EMADM	0.44037E-04	0.33538E-03	0.131	-0.14183	0.16046	-0.884
CASEMIX	46.013	20.158	2.283**	0.11644	0.20641	0.564
BEDSEM	0.30445E-01	0.12220E-01	2.491**	0.40048	0.17285	2.317**
RES	0.14770E-02	0.57377E-03	2.574**	0.16757E-01	0.34614E-01	0.484
INVOCC	0.26354	2.9978	0.088	-0.73827E-01	0.20458	-0.361
AEATT	0.20236	0.85865E-01	2.357**	0.14318E-01	0.28752E-01	0.498
DAYATT	0.34418E-03	0.12554E-03	2.742**	-0.23879E-03	0.86347E-02	-0.028
ELADM	0.64899E-03	0.22217E-03	2.921**	0.30349	0.82392E-01	3.683**
OPATT	0.94257E-01	0.26788E-01	3.519**	0.25094	0.10202	2.460**
WAGE INDEX	-0.79573	0.72726	-1.094	0.91879E-01	0.20101	0.457
DVTEACH	47.421	7.0428	6.733**	0.39759	0.11199	3.550**
CONSTANT	23.942	14.020	1.708	-1.6600	1.0578	-1.569
Log-Likelihood	-335.40228			-336.94		
Hetero.	0.00			0.00		

* significant at 10%

** significant at 5%

Table A1.4.13: Instrumental variables estimate

Variable	Box-Cox model with teaching dummy (λ = 1.2112, θ = 0.53812)		
	Coeff.	SE	t-stat
EMADM	0.162902E-05	0.26808E-05	0.607
CASEMIX	7.4767	3.4460	2.170**
BEDSEM	0.14723E-02	0.43017E-03	3.423**
RES	0.28535E-04	0.11666E-04	2.446**
INVOCC	-0.37517	0.30932	-1.213
AEATT	0.10438E-01	0.50212E-02	2.079**
DAYATT	0.72369E-05	0.20920E-05	3.459**
ELADM (IV)	0.75891E-05	0.58973E-05	1.287
OPATT	0.25942E-02	0.10865E-02	2.388**
WAGE INDEX	0.10945E-01	0.52237E-01	0.210
DVTEACH	5.7134	1.0033	5.695**
CONSTANT	9.7692	1.7602	5.550**
Log-Likelihood	-325.94		

* significant at 10%

** significant at 5%

Table A1.4.14: Box-Cox estimates dropping RES and INVOCC

	Box-Cox model without RES (λ = 1.2112, θ = 0.53812)			Box-Cox model without RES & INVOCC (λ = 1.2112, θ = 0.53812)		
	Coeff.	SE	t-stat	Coeff.	SE	t-stat
EMADM	0.40087E-05	0.48141E-05	0.833	0.61617E-05	0.44441E-05	1.387
CASEMIX	6.3165	2.9074	2.173**	4.8854	2.6293	1.858
BEDSEM	0.14151E-02	0.38546E-03	3.671**	0.12508E-02	0.35895E-03	3.485**
RES	--	--	--	--	--	--
INVOCC	-0.33992	0.30418	-1.117	--	--	--
AEATT	0.72072E-02	0.50618E-02	1.424	0.67521E-02	0.50823E-02	1.329
DAYATT	0.71678E-05	0.21070E-05	3.402**	0.71155E-05	0.21219E-05	3.353**
ELADM	0.95980E-05	0.33934E-05	2.828**	0.99292E-05	0.34052E-05	2.916**
OPATT	0.28294E-02	0.10959E-02	2.582**	0.29130E-02	0.11013E-02	2.645**
WAGE INDEX	0.12569E-01	0.52491E-01	0.239	-0.58307E-02	0.50206E-01	-0.116
DVTEACH	5.5564	0.94046	5.908**	5.7994	0.92167	6.292**
CONSTANT	9.4557	1.6564	5.709**	9.1853	1.6506	5.565**
Log-L	-328.98			-329.61		
R^2	0.9977			0.9977		
Hetero	0.426			0.485		

* significant at 10%
** significant at 5%

Table A1.4.15: Elasticities and marginal costs without RES and INVOCC

Variable	Elasticity (-RES)	Marginal cost (-RES)	Elasticity (-RES & INVOCC)	Marginal cost (-RES & INVOCC)
EMADM	0.0882	£257.99	0.1357	£396.94
CASEMIX	0.4185	£43,975,650.61	0.3239	£3,4035,157.06
BEDSEM	0.3707	£41,406.07	0.3278	£36,614.27
RES	--	--	--	--
INVOCC	-0.0685	-£2,868,007.65	--	--
AEATT	0.1046	£128.51	0.0981	£120.53
DAYATT	0.0590	£386.06	0.0586	£383.44
ELADM	0.2077	£614.20	0.2149	£635.49
OPATT	0.1689	£64.72	0.1740	£66.67
WAGE INDEX	0.0463	£177.06	-0.0215	-£82.22

Table A1.4.16: Box-Cox estimates using SERES

Variable	Box-Cox model with site dummies (λ=0.91040,θ=0.39150)		
	Coeff.	SE	t-stat
EMADM	0.36948E-04	0.60974E-04	0.606
CASEMIX	3.2830	1.4044	2.338**
BEDSEM	0.54487E-02	0.15828E-02	3.442**
SERES	0.36112E-02	0.20608E-02	1.752*
INVOCC	-0.20203	0.24663	-0.819
AEATT	0.10907E-01	0.85315E-02	1.278
DAYATT	0.69228E-04	0.21351E-04	3.242**
ELADM	0.11524E-03	0.39346E-04	2.929**
OPATT	0.66233E-02	0.32181E-02	2.058**
WAGE INDEX	0.33322E-01	0.68249E-01	0.488
DVTEACH	2.9313	0.50071	5.854**
CONSTANT	5.4992	1.0588	5.194**
Log-Likelihood	-327.09958		
R^2	0.999		
Hetero	0.614		

* significant at 10%

** significant at 5%

Table A4.1.17: Elasticities and marginal costs using SERES

Variable	Elasticity	Marginal cost
EMADM	0.0710	£206.80
CASEMIX	0.4466	£46,928,373.91
BEDSEM	0.3747	£42,346.01
RES	0.0466	£2,211.48
INVOCC	-0.0634	-£2,671,355.19
AEATT	0.0852	£105.02
DAYATT	0.0635	£416.59
ELADM	0.2187	£646.15
OPATT	0.1499	£57.44
WAGE INDEX	0.0928	£354.89

APPENDIX 2: Appendix to Chapter 5

APPENDIX A2.1: DEMAND ESTIMATION

Table A2.1 Estimate of ADC using AR(1)regression

Variable	Coefficient	Standard Error	t-stat
DVJAN	14.188**	3.5398	4.008
DVFEB	12.364**	4.4314	2.790
DVMAR	27.337**	4.8613	5.623
DVAPR	-8.3545	6.5516	-1.275
DVMAY	-14.677**	5.6876	-2.580
DVJUN	11.133*	5.3789	2.070
DVJUL	-13.566**	5.2362	-2.591
DVAUG	-24.850**	5.0958	-4.877
DVSEP	-2.9483	4.8655	-0.606
DVOCT	-7.4433	4.4323	-1.679
DVNOV	10.907**	3.5399	3.081
FILT1993	-20.229**	6.8532	-2.952
FILT1994	-1.8713	4.5948	-0.407
GROUP EFFECTS			
1	408.36630**	14.58774	
2	688.50641**	14.58774	
3	659.30807**	17.63003	
4	275.95857**	24.86976	
5	731.05859**	24.86976	
6	351.22111**	14.58774	
7	259.82505**	14.58774	
8	312.68822**	24.86976	
9	747.04048**	24.86976	
10	403.68652**	24.86976	
11	328.55727**	17.47994	
12	612.56033**	24.86976	
13	393.10210**	14.58774	
14	161.58887**	24.86976	
15	239.73477**	17.47994	
16	254.74832**	14.58774	
17	345.46465**	24.86976	
18	301.40867**	24.86976	
19	414.30561**	14.58774	
20	421.13559**	14.58774	
21	337.02874**	24.86976	
22	217.81015**	17.47994	
23	320.67433**	18.08912	
24	536.71913**	17.47994	
25	271.95527**	17.47994	
26	373.17560**	24.86976	
27	610.38905**	18.08912	
28	425.16747**	24.86976	
29	636.80810**	17.47994	
30	432.42916**	24.86976	
31	303.38804**	14.58774	
32	618.19259**	14.58774	
33	548.49187**	24.86976	

34	405.06521**	14.58774
35	430.85940**	24.86976
36	265.87081**	17.47994
37	552.31076**	17.47994
38	289.66410**	17.47994
39	335.56431**	14.58774
40	574.65451**	14.58774
41	337.61038**	14.58774
42	373.33090**	17.47994
43	450.69191**	14.58774
44	414.22225**	14.58774
45	433.25038**	24.86976
46	9.18333	24.86976
47	476.42260**	24.86976
48	49.26510**	17.47994
49	444.58813**	24.86976
50	655.38108**	17.47994
51	355.04663**	17.47994
52	236.70757**	17.47994
53	450.41951**	24.86976
54	449.20119**	17.47994
55	322.39819**	14.58774
56	195.49531**	14.58774
57	443.90057**	14.58774
58	327.83393**	24.86976
59	625.16770**	14.58774
60	338.90679**	17.47994
61	499.90229**	24.86976
62	260.04885**	24.86976
63	485.37237**	14.58774
64	373.06814**	17.47994

RHO = 0.569637
R-squared = 0.77879
F = 63.51**
Autocorrelation: -0.097
Heteroskedasticity: 77.35

** significant at 5%
* significant at 10%

APPENDIX A2.2: PROBABILITY OF TURNAWAY ESTIMATES

Table A2.2: Probability of turnaway

SITEID	BEDS	ADC est	S.D. of ADC residual	Probability of turnaway (annual)	Probability of turnaway (monthly)
Prov10	689	433	28	0.000E+00	2.35E-11
Prov100	1085	699	26	0.000E+00	1.32E-12
Prov103	934	670	75	2.158E-04	1.72E-3
Prov104	513	285	14	0.000E+00	0.000E+00
Prov106	1012	745	94	2.253E-03	6.89E-03
Prov107	524	368	20	3.109E-15	1.58E-03
Prov111	368	261	45	8.709E-03	2.06E-02
Prov112	548	330	11	0.000E+00	0.000E+00
Prov114	781	538	31	2.331E-15	9.99E-01
Prov115	403	333	22	7.318E-04	9.869E-01
Prov117	534	339	17	0.000E+00	1.277E-15
Prov120	988	658	16	0.000E+00	0.000E+00
Prov121	587	411	23	9.992E-15	6.65E-05
Prov122	292	173	13	0.000E+00	5.857E-12
Prov124	484	268	15	0.000E+00	0.000E+00
Prov130	684	261	18	0.000E+00	0.000E+00
Prov131	512	379	13	0.000E+00	0.000E+00
Prov133	527	317	9	0.000E+00	0.000E+00
Prov15	498	415	95	1.911E-01	2.07E-01
Prov16	822	440	22	0.000E+00	0.000E+00
Prov17	580	386	20	0.000E+00	4.47E-15
Prov20	697	231	17	0.000E+00	0.000E+00
Prov22	431	324	18	1.392E-09	1.28E-05
Prov24	653	580	29	5.914E-03	4.31E-02
Prov26	404	282	19	6.799E-11	4.27E-06
Prov27	666	393	15	0.000E+00	0.000E+00
Prov3	1001	637	29	0.000E+00	0.000E+00
Prov30	613	454	19	0.000E+00	1.63E-10
Prov31	968	680	40	3.032E-13	6.29E-11
Prov32	679	460	16	0.000E+00	0.000E+00
Prov37	438	307	32	2.123E-05	1.08E-03
Prov38	972	610	49	7.527E-14	1.90E-07
Prov39	1274	611	171	5.285E-05	1.42E-04
Prov4	614	421	49	4.097E-05	9.12E-03
Prov43	737	453	23	0.000E+00	0.000E+00
Prov45	491	282	71	1.622E-03	2.58E-03
Prov48	849	590	29	0.000E+00	2.58E-12
Prov49	462	318	23	1.922E-10	2.57E-6
Prov5	467	344	32	6.061E-05	1.07E-03
Prov50	676	603	42	4.110E-02	1.09E-01
Prov51	570	382	23	0.000E+00	2.49E-12
Prov53	656	409	19	0.000E+00	0.000E+00
Prov54	655	467	30	1.852E-10	2.61E-07
Prov55	834	423	47	0.000E+00	5.99E-15
Prov56	641	490	45	3.961E-04	2.73E-03
Prov59	92	11	17	9.468E-07	8.97E-04
Prov6	721	544	12	0.000E+00	0.000E+00

Prov60	111	50	12	1.858E-07	2.63E-03
Prov62	592	473	17	1.288E-12	1.23E-06
Prov64	933	670	19	0.000E+00	0.000E+00
Prov66	492	362	17	1.033E-14	4.28E-02
Prov67	386	249	14	0.000E+00	2.47E-15
Prov69	684	477	23	0.000E+00	1.94E-06
Prov75	763	464	30	0.000E+00	0.000E+00
Prov78	614	426	124	6.474E-02	8.31E-02
Prov82	393	196	28	9.978E-13	3.24E-05
Prov85	852	478	23	0.000E+00	0.000E+00
Prov86	595	335	11	0.000E+00	0.000E+00
Prov88	1112	646	95	4.672E-07	6.02E-06
Prov89	530	356	34	1.550E-07	1.77E-04
Prov90	628	521	26	1.934E-05	1.71E-02
Prov92	374	272	16	9.190E-11	3.56E-07
Prov94	704	491	26	0.000E+00	1.70E-07
Prov99	479	390	84	1.447E-01	1.58E-01

Annual average = 0.007214
Monthly average = 0.040215

APPENDIX A2.3: COST FUNCTION ESTIMATES

Table A2.3.1 Log-likelihood values of lambda and theta for grid search (Sample = non-zero PT)

	λ				
θ	**0**	**0.25**	**0.5**	**0.75**	**1.0**
0	8478.90	8482.10	8484.27	8485.00	8484.37
0.25	7834.12	7836.80	7838.63	7839.62	7839.93
0.5	5860.38	5862.99	5864.75	5865.64	5865.84
0.75	3644.42	3647.98	3650.59	3652.14	3652.77
1.0	1322.46	1327.05	1330.67	1333.17	1334.58

Optimal values ($\lambda = 0.76969$, $\theta = 0.065794$) Log-likelihood = 8644.50

Table A2.3.2 Box-Cox analysis (sample PT= nonzero)

Variable	Coefficient	Standard Error	t-stat
DEL	0.52119E-02	0.68089E-03	7.655**
DEM	0.70107E-02	0.34797E-02	2.014*
P10	-6.0146	0.61084	-9.846**
P100	-10.307	1.1211	-9.194**
P103	-2.8170	0.63718	-4.421**
P106	-3.7705	0.89927	-4.193**
P107	-4.4496	0.43137	-10.315**
P111	3.7666	0.37357	10.083**
P114	4.7026	0.70994	6.624**
P115	5.9680	0.56251	10.610**
P117	-6.1095	1.1531	-5.298**
P121	-5.6027	0.57111	-9.810**
P122	-4.8276	1.1612	-4.158**
P15	3.5920	0.46957	7.649**
P17	-7.0204	1.1619	-6.042**
P22	-3.4432	0.42262	-8.147**
P24	-1.5631	0.68316	-2.288**
P26	-3.6117	0.42751	-8.448**
P30	-7.6497	1.0357	-7.386**
P31	-8.9271	0.65060	-13.721**
P38	-8.2061	0.67547	-12.149**
P39	-3.6663	0.72189	-5.079**
P4	-2.6446	0.50490	-5.238**
P45	1.9705	0.42445	4.642**
P48	-10.224	0.90448	-11.303**
P49	-3.9184	0.41707	-9.395**
P5	0.26298	0.37426	0.703
P50	-0.12281	0.65379	-0.188
P51	-6.8638	0.53889	-12.737**
P54	-6.8927	0.52053	-13.242**
P55	-7.4254	0.72881	-10.188**
P56	-0.65130	0.59280	-1.099
P59	-1.0836	0.66086	-1.640
P60	0.34057	0.53264	0.639
P62	-7.1763	0.69454	-10.333**
P66	1.9204	0.54906	3.498**
P69	-6.9052	1.0348	-6.673**
P78	5.2545	0.40348	13.023**
P82	-2.6432	0.49478	-5.342**
P88	-4.3945	0.53557	-8.205**
P89	-2.5818	0.42620	-6.058**
P90	-5.2971	0.82560	-6.416**
P92	-4.0854	0.55332	-7.383**
P94	-7.8723	0.64563	-12.193**
P99	5.3970	0.42824	12.603**
Constant	-10.467	0.51655	-20.263**
Lambda	0.76969		
Theta	0.65794E-01		

Log-likelihood = 8644.50
N = 814
White test for heteorskedasticity: 0.1599
Autocorrelation: 0.0006

* significant at 5%
* siginificant at 10%

Since the Box-Cox analysis using the full sample required a manual transformation of the data the identification of the optimal transform became more difficult since we could not take advantage of LIMDEP's MLE search once appropriate starting values had been identified. Table A2.3.3 indicates that the optimal transformation occurred when lambda and theta both equalled one, i.e. a linear transform. This allowed the TOBIT and PROBIT models to be interpreted in a straightforward manner without applying transforms to the data. In particular the calculation of the implied trade-off between waiting lists and turnaway is uncomplicated. Consequently, subsequent models using the full sample size were estimated using a linear specification.

Table A2.3.3 Log-likelihood values of lambda and theta for grid search (Sample = All PT)

	λ				
θ	**0**	**0.25**	**0.5**	**0.75**	**1.0**
0	-4829.10	-4836.33	-4835.23	-4834.30	-4833.68
0.25	-9156.10	-10533.93	-10564.52	11184.45	1183.02
0.5	1147.89	1187.66	1186.92	1184.45	1183.02
0.75	2100.48	2141.16	2141.82	2140.06	2139.18
1.0	2858.42	2894.57	2896.87	2897.05	2897.33

Table A2.3.4 Box-Cox analysis (sample PT= all)

Variable	Coefficient	Standard Error	t-stat
DEL	0.15744E-04	0.18202E-05	8.649**
DEM	0.10749E-03	0.21829E-04	4.924**
P10	-0.29680E-01	0.90377E-02	-3.284**
P100	-0.12693	0.13839E-01	-9.172**
P103	-0.10634	0.13341E-01	-7.971**
P106	-0.15915	0.18735E-01	-8.495**
P107	-0.19115E-01	0.87292E-02	-2.190**
P111	0.24210E-01	0.86268E-02	2.806**
P114	0.90619	0.15639E-01	57.943**
P115	0.94695	0.12595E-01	75.182**
P117	-0.10718E-01	0.96341E-02	-1.112
P121	-0.57448E-01	0.99521E-02	-5.772**
P122	0.15599E-01	0.12471E-01	1.251
P15	0.15027	0.99179E-02	15.151**
P17	-0.27325E-01	0.12350E-01	-2.212**
P22	-0.82538E-02	0.96170E-02	-0.858
P24	-0.75485E-01	0.13982E-01	-5.399**
P26	-0.10952E-02	0.96124E-02	-0.114
P30	-0.58937E-01	0.13120E-01	-4.492**
P31	-0.10663	0.13413E-01	-7.950**
P38	-0.10205	0.12196E-01	-8.368**
P39	-0.11069	0.15461E-01	-7.160**
P4	-0.58970E-01	0.10480E-01	-5.627**
P45	-0.64890E-02	0.96997E-02	-0.669
P48	-0.11626	0.13808E-01	-8.419**
P49	-0.78697E-02	0.96171E-02	-0.818
P5	-0.86648E-02	0.86294E-02	-1.004
P50	-0.10514E-01	0.13288E-01	-0.791
P51	-0.30468E-01	0.89455E-02	-3.406**
P54	-0.74503E-01	0.10710E-01	-6.956**
P55	-0.43979E-01	0.93372E-02	-4.710**
P56	-0.55316E-01	0.13105E-01	-4.221**
P59	0.40881E-01	0.13598E-01	3.006**
P60	0.45346E-01	0.10970E-01	4.134**
P62	-0.74678E-01	0.13716E-01	-5.444**
P66	0.14196E-01	0.98951E-02	1.435
P69	-0.75435E-01	0.13729E-01	-5.495**
P78	0.72698E-01	0.90410E-02	8.041**
P82	0.84591E-03	0.90482E-02	0.093
P88	-0.81185E-01	0.11497E-01	-7.061**
P89	-0.19487E-01	0.97283E-02	-2.003**
P90	-0.11538	0.17076E-01	-6.757**
P92	-0.73306E-02	0.12244E-01	-0.599
P94	-0.97208E-01	0.12092E-01	-8.039**
P99	0.14092	0.98079E-02	14.368**
P104	-0.82970E-02	0.12208E-01	-0.680
P112	-0.50373E-01	0.13235E-01	-3.806**
P120	-0.77103E-01	0.14340E-01	-5.377**
P124	-0.29817E-02	0.96769E-02	-0.308
P130	-0.18518E-01	0.91010E-02	-2.035*
P131	-0.99730E-04	0.12320E-01	-0.008
P133	-0.10068E-01	0.12171E-01	-0.827

P16	-0.28693E-01	0.90530E-02	-3.169**
P20	-0.53560E-03	0.98000E-02	-0.055
P27	-0.36268E-01	0.12509E-01	-2.899**
P3	-0.77948E-01	0.12096E-01	-6.444**
P32	-0.38210E-01	0.12647E-01	-3.021**
P43	-0.10252	0.15341E-01	-6.683**
P53	-0.38053E-01	0.10106E-01	-3.765**
P6	-0.70977E-01	0.13595E-01	-5.221**
P64	-0.97526E-01	0.12958E-01	-7.527**
P67	-0.13331E-02	0.97298E-02	-0.137
P75	-0.32666E-01	0.10156E-01	-3.216**
P85	-0.64423E-01	0.10155E-01	-6.344**
P86	-0.24221E-01	0.12337E-01	-1.963*
Lambda	1.0		
Theta	1.0		

Log-L = 2897.05
N = 1512
White test for heteorskedasticity: 0.65843
Autocorrelation: 0.30260

* significant at 5% level
** significant at 10% level

Table A2.3.5 Tobit model

Variable	Coefficient	Standard Error	t-stat
DEL	0.27803E-04	0.28562E-05	9.734**
DEM	0.11374E-03	0.30886E-04	3.683**
Constant	-0.78870E-01	0.12026E-01	-6.558**
P10	-0.87545E-01	0.14119E-01	-6.201**
P100	-0.27254	0.24351E-01	-11.192**
P103	-0.16204	0.18489E-01	-8.764**
P106	-0.25633	0.26825E-01	-9.556**
P107	-0.44669E-01	0.11971E-01	-3.731**
P111	0.24780E-01	0.11250E-01	2.203**
P114	0.87043	0.21037E-01	41.376**
P115	0.92417	0.16617E-01	55.616**
P117	-0.88520E-01	0.19965E-01	-4.434**
P121	-0.12310	0.15044E-01	-8.183**
P122	-0.38819E-01	0.22978E-01	-1.689
P15	0.11549	0.13552E-01	8.522**
P17	-0.99289E-01	0.23141E-01	-4.291**
P22	-0.14053E-01	0.12608E-01	-1.115
P24	-0.14478	0.19871E-01	-7.286**
P26	-0.61336E-02	0.12667E-01	-0.484
P30	-0.15759	0.25411E-01	-6.202**
P31	-0.16437	0.18667E-01	-8.805**
P38	-0.18610	0.18253E-01	-10.196**
P39	-0.16678	0.21128E-01	-7.894**
P4	-0.10108	0.14524E-01	-6.960**
P45	-0.14250E-01	0.12695E-01	-1.123
P48	-0.23366	0.22401E-01	-10.431**
P49	-0.12207E-01	0.12542E-01	-0.973
P5	-0.12172E-01	0.11255E-01	-1.082
P50	-0.78504E-01	0.18962E-01	-4.140**
P51	-0.83601E-01	0.13420E-01	-6.229**
P54	-0.12065	0.14890E-01	-8.102**
P55	-0.12422	0.15583E-01	-7.971**
P56	-0.84941E-01	0.17402E-01	-4.881**
P59	0.50149E-01	0.18000E-01	2.786**
P60	0.57925E-01	0.14598E-01	3.968**
P62	-0.13003	0.19291E-01	-6.740**
P66	-0.17646E-01	0.13611E-01	-1.296
P69	-0.16875	0.23501E-01	-7.180**
P78	0.74858E-01	0.11901E-01	6.290**
P82	-0.31697E-01	0.12953E-01	-2.447**
P88	-0.11658	0.15610E-01	-7.468**
P89	-0.29887E-01	0.12731E-01	-2.348**
P90	-0.19977	0.24401E-01	-8.187**
P92	-0.18524E-01	0.16233E-01	-1.141
P94	-0.18255	0.17813E-01	-10.248**
P99	0.13607	0.12822E-01	10.613**
P104	-0.25979	4.3321	-0.060
P112	-0.33262	4.2181	-0.079
P120	-0.36729	3.9459	-0.093
P124	-0.25218	3.0072	-0.084
P130	-0.27995	2.4725	-0.113
P131	-0.23818	4.2908	-0.056

P133	-0.26167	4.1744	-0.063
P16	-0.28513	2.4284	-0.117
P20	-0.25122	2.9321	-0.086
P27	-0.30278	4.0965	-0.074
P3	-0.36522	2.6780	-0.136
P32	-0.30007	4.2708	-0.070
P43	-0.42866	3.7195	-0.115
P53	-0.30467	3.0110	-0.101
P6	-0.35379	4.0685	-0.087
P64	-0.39440	2.7831	-0.142
P67	-0.93979E-01	0.24495E-01	-3.837**
P75	-0.28997	3.0308	-0.096
P85	-0.35183	2.2818	-0.154
P86	-0.29011	3.9775	-0.073

Threshold values for the model: Lower= 0.0000
Upper= +Infinity

Log-likelihood = 1219.827
N = 1512
Heteorskedasticity: 0.66
Autocorrelation: 0.0324

* significant at 5% level
** significant at 10% level

Heckman two-stage model

```
Binomial Probit Model
Maximum Likelihood Estimates
Dependent variable                    Y1   (binary y = 0 or not zero)
Number of observations              1512
Iterations completed                   3
Log likelihood function        -1040.030
Restricted log likelihood      -1043.584
Chi-squared                     7.108457
```

Table A2.3.6 Binomial probit model

Variable	Coefficient	Standard Error	t-stat
DEL	0.65048E-04	0.24507E-04	2.654**
DEM	-0.57269E-03	0.31013E-03	-1.847
Constant	0.10104	0.94626E-01	1.068

* significant at 5% level
** significant at 10% level

Frequencies of actual & predicted outcomes
Predicted outcome has maximum probability.

	Predicted		
Actual	0	1	TOTAL
0	55	643	698
1	78	736	814
TOTAL	36	1379	1512

```
Sample Selection Model
Probit selection equation based on Y1
Selection rule is: Observations with Y1      =  1

Results of selection:
                   Data points     Sum of weights
Data set               1512            1512.0
Selected sample         814             814.0
```

Table A2.3.7 Two stage least squares regression

Variable	Coefficient	Standard Error	t-stat
DEL	0.26255E-04	0.45156E-05	5.814**
DEM	0.12461E-03	0.59275E-03	0.210
P10	-0.41463E-01	0.18445E-01	-2.248**
P100	-0.19292	0.32754E-01	-5.890**
P103	-0.15066	0.18744E-01	-8.038**
P106	-0.23958	0.27481E-01	-8.718**
P107	-0.27225E-01	0.13018E-01	-2.091*
P111	0.24577E-01	0.11341E-01	2.167*
P114	0.88127	0.20873E-01	42.220**
P115	0.92826	0.16843E-01	55.114**
P117	-0.16804E-01	0.34387E-01	-0.489
P121	-0.10210	0.17060E-01	-5.985**
P122	0.10246E-01	0.35839E-01	0.286
P15	0.12114	0.13871E-01	8.734**
P17	-0.47797E-01	0.34591E-01	-1.382
P22	-0.11690E-01	0.12659E-01	-0.923
P24	-0.13289	0.20795E-01	-6.391**
P26	-0.30608E-02	0.12813E-01	-0.239
P30	-0.12622	0.30939E-01	-4.080**
P31	-0.14905	0.19012E-01	-7.840**
P38	-0.15870	0.19964E-01	-7.949**
P39	-0.15493	0.21126E-01	-7.334**
P4	-0.94364E-01	0.15173E-01	-6.219**
P45	-0.13247E-01	0.12735E-01	-1.040
P48	-0.18774	0.27180E-01	-6.907**
P49	-0.11253E-01	0.12476E-01	-0.902
P5	-0.10965E-01	0.11294E-01	-0.971
P50	-0.66454E-01	0.19816E-01	-3.354**
P51	-0.50627E-01	0.16179E-01	-3.129**
P54	-0.11083	0.15739E-01	-7.041**
P55	-0.62778E-01	0.21801E-01	-2.880**
P56	-0.78865E-01	0.17797E-01	-4.431**
P59	0.46437E-01	0.29382E-01	1.580
P60	0.55565E-01	0.23746E-01	2.340**
P62	-0.11380	0.20812E-01	-5.468**
P66	0.40261E-01	0.16403E-01	2.454**
P69	-0.12674	0.30775E-01	-4.118**
P78	0.76415E-01	0.12416E-01	6.155**
P82	-0.10910E-01	0.16110E-01	-0.677
P88	-0.10761	0.15681E-01	-6.862**
P89	-0.27755E-01	0.12787E-01	-2.171*
P90	-0.18693	0.25510E-01	-7.328**
P92	-0.15053E-01	0.16627E-01	-0.905
P94	-0.13789	0.19543E-01	-7.056**
P99	0.13816	0.12959E-01	10.661**
Constant	-0.54440E-01	0.42826	-0.127
LAMBDA	-0.42021E-01	0.99266	-0.042

Log-likelihood = 1358.79
N = 1512
Heteorskedasticity: 0.957
Autocorrelation: 0.302

* significant at 5% level
** significant at 10% level

The Heckman two-stage model is not, without its problems. Most notably, the PROBIT analysis, which is the first stage of the procedure, cannot deal with dummy variables where the dependent variable does not change. That is, if the probability of turnaway remains at zero for a single provider for all observations then the PROBIT analysis cannot deal with these dummy variables. As this condition is likely to be violated by a number of our providers we opted to undertake a simple bivariate regression for the probit analysis. (Another option would have been to identify all the providers who violated the above condition and take out these dummy variables, either of these approaches is less than ideal).

The model above is very similar to the Box-Cox linear model; the coefficients are almost identical for all variables (see table A2.3.8 below). This indicates that the exclusion of non-zero values may not lead to the introduction of bias in the estimation. Furthermore, the insignificant values of lambda supports this supposition. Therefore we can use the sampled data to estimate the trade-off.

Table A2.3.8 Box-Cox linear specification

Variable	Coefficient	Standard Error	t-stat
DEL	0.26393E-04	0.31518E-05	8.374**
DEM	0.99588E-04	0.32500E-04	3.064**
P10	-0.41292E-01	0.18528E-01	-2.229**
P100	-0.19288	0.34302E-01	-5.623**
P103	-0.15058	0.19277E-01	-7.811**
P106	-0.23951	0.28437E-01	-8.422**
P107	-0.27122E-01	0.13085E-01	-2.073*
P111	0.24499E-01	0.11382E-01	2.153*
P114	0.88116	0.21653E-01	40.694**
P115	0.92839	0.16970E-01	54.707**
P117	-0.16691E-01	0.35162E-01	-0.475
P121	-0.10196	0.17187E-01	-5.932**
P122	0.99380E-02	0.35339E-01	0.281
P15	0.12124	0.14041E-01	8.635**
P17	-0.47669E-01	0.35374E-01	-1.348
P22	-0.11645E-01	0.12877E-01	-0.904
P24	-0.13270	0.20977E-01	-6.326**
P26	-0.30827E-02	0.13036E-01	-0.236
P30	-0.12605	0.31593E-01	-3.990**
P31	-0.14898	0.19651E-01	-7.581**
P38	-0.15858	0.20468E-01	-7.748**
P39	-0.15497	0.21915E-01	-7.071**
P4	-0.94209E-01	0.15152E-01	-6.218**
P45	-0.13311E-01	0.12882E-01	-1.033
P48	-0.18755	0.27815E-01	-6.743**
P49	-0.11226E-01	0.12710E-01	-0.883
P5	-0.10891E-01	0.11396E-01	-0.956
P50	-0.66283E-01	0.20035E-01	-3.308**
P51	-0.50492E-01	0.16277E-01	-3.102**
P54	-0.11065	0.15599E-01	-7.093**
P55	-0.62611E-01	0.22090E-01	-2.834**
P56	-0.78686E-01	0.17823E-01	-4.415**
P59	0.45473E-01	0.18374E-01	2.475**
P60	0.54787E-01	0.14971E-01	3.660**
P62	-0.11361	0.20969E-01	-5.418**
P66	0.40366E-01	0.16610E-01	2.430**
P69	-0.12655	0.31453E-01	-4.023**
P78	0.76571E-01	0.12139E-01	6.308**
P82	-0.11191E-01	0.14821E-01	-0.755
P88	-0.10754	0.16106E-01	-6.677**
P89	-0.27674E-01	0.12923E-01	-2.141*
P90	-0.18672	0.25877E-01	-7.216**
P92	-0.15122E-01	0.16819E-01	-0.899
P94	-0.13769	0.19573E-01	-7.035**
P99	0.13826	0.13021E-01	10.618**
Constant	-0.72570E-01	0.12329E-01	-5.886**

Log-likelihood = 1334.58
N = 814
Heteorskedasticity: 0.882
Autocorrelation: 0.362

* significant at 5% level
** significant at 10% level

APPENDIX 3: DATA

This appendix is separated into two main sections. The first deals with data definitions and sources, the second deals with descriptive statistics for the data used in the analyses presented in the main text. These descriptive data will be separated into three parts, referring to the three different analyses undertaken in Chapters 4 and 5.

A3.1: Data sources

The data were collected from three sources; CHKS Ltd, CIPFA and Ivan Csaba, Central European University, Budapest.

The primary data were collected for activity variables, such as number of admissions, length of stay, number of beds, etc. These data were obtained from CHKS Ltd.

CHKS was founded as a transatlantic joint venture between CPHA (Commission on Professional and Hospital Activity) in the US and CASPE research in the UK. CHKS is a subsidiary of HCIA Inc. based in the US which holds the world's largest patient-based hospital database (350 million records). CHKS is the biggest supplier of comparative hospital data in the UK and a major supplier in Europe.

CHKS collects anonymised copies of the Contract Minimum Data Sets (CMDS) returns that hospital have a mandatory duty to submit to purchasers. These contain information for each finished consultant episode. The data include detailed patient-based information as well as information on the type of admission, length of stay, etc. CHKS validates these data and then complies reports summarising the information received. They produce a publication entitled Acute Care which are a series of annual handbooks providing information on hospital performance at an HRG level.

CHKS are overseen by a board of trustees chaired by the King's Fund with representatives from the Royal College of Nurses, the Royal College of Midwives and other professional organisations. One of the aims of CHKS is to improve the quality of routinely collected patient-based information.

For our purposes data were aggregated to trust level and all data were anonymised by CHKS. All activity data provided by CHKS were on a monthly basis.

The initial data set provided by CHKS were obtained for 47, 83 and 119 hospitals in England, Wales, Scotland and Northern Ireland for financial years 1992/3, 1993/4 and 1994/5 respectively, giving a total of 2,988 observations. Hospitals with incomplete observations within any (financial) year were excluded. This restricted the sample to 124 hospitals. However, since the analysis required extrapolation between the different estimated equations we only considered data for those hospitals for which we had both cost and activity data. The minimum data set was therefore restricted by the availability of cost data. Whilst this reduced our sample size we wanted to avoid potential problems of bias that might have occurred had we used different hospitals to estimate the different relationships. This restricted our sample to English hospitals only since cost data were only available for English trusts

A complete data set, where cost data were also available, was obtained for 64 provider units. The activity data provided 1512 observational points. Of these 288 were from 1992/93, 480 from 1993/94, and 744 from 1994/95.

A list of variables and their definitions and means and standard deviations are presented in Table A3.1 below.

Table A3.1: Variable definitions

Variable name	**Definition**	**Mean (S.D.)**
ADMS	Total number of admissions	4,153.58 (1753.71)
EMERGADM	Number of admissions categorised as emergency (by provider)	1,390.69 (538.03)
ELADM	Number of admissions categorised as elective (by provider)	2,063.92 (1046.79)
OTHADM	maternity care and transfers from other providers	697.92 (365.99)
BEDS	Total number of available staffed beds	975.38 (365.42)
ALOS	Average length of stay (days) for all admissions	4.75 (1.69)
ELALOS	Average length of stay (days) for elective admissions	4.76 (.74)
EMERGALOS	Average length of stay (days) for emergency admissions	6.90 (1.10)
OTHALOS	Average length of stay (days) for admissions categorised as 'other'	4.33 (2.05)
AVCMIX	Casemix weight based on DRG weights	.69 (.09)
VALIDWAIT	Total number of elective patients waiting for admission	1,617.66 (1069.51)
PROV	Provider type (teaching London, teaching provincial, children's, non-teaching, 'other')	

Each trust (provider unit) was categorised by CHKS as teaching London (TL), teaching provincial (TP), children's (C), non-teaching (N) or other (--). Of the sample of 64 hospitals the breakdown is listed in Table A3.2 below.

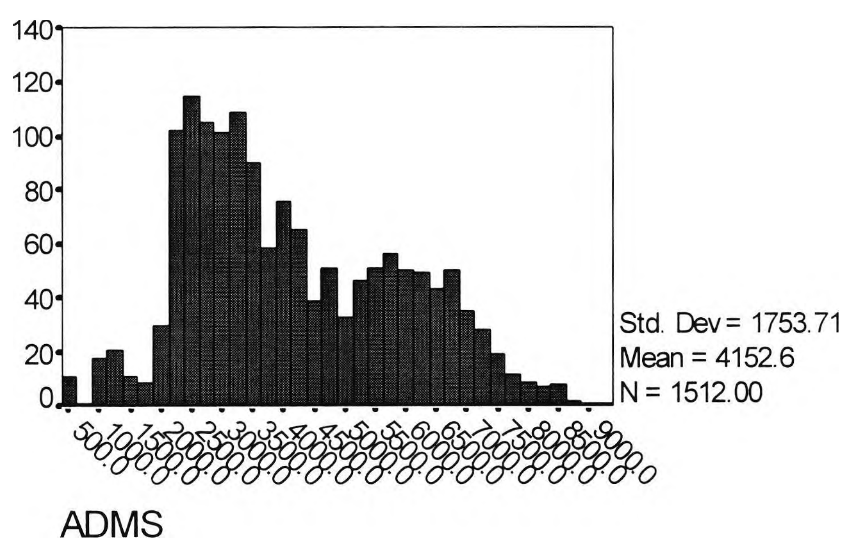
Table A3.2: Hospital categories

Hospital type	Number in sample
TL	6
TP	8
N	45
C	1
--	4

The sample contained 14 (22%) teaching hospitals. The majority of hospitals were, however, non-teaching hospitals 50 (78%).

The sample provided a total of 6,278,701 inpatient admissions to hospitals in England over the three year sample period. 1,161,573 from 1992/93, 1,892,293 from 1993/94, and 3,224,835 from 1994/95. The number of monthly admissions ranged from 494 to 9,210. Figure A3.1 provides a histogram of monthly admissions.

Figure A3.1: Admissions



Of these admissions, 3,120,653 were categorised as elective, 2,102,722 were emergency and 1,055,256 were 'other', giving a total of 3,157,981 non-elective admissions in total.

In addition to the monthly activity data provided by CHKS we also collected data on some additional variables that were not held within the CHKS dataset. These variables were provided by Ivan Csaba (based on the Department of Health dataset) and were provided on an annual basis and are defined in Table A3.3 below. The data provided by Ivan Csaba were tied into the CHKS data by CHKS.

Table A3.3: Annual activity data

Variable name	Definition	Mean (S.D.)
DAYATT	Total number of day attendances	11,110.58 (12766.99)
AEATT	Total number of non-inpatient Accident and Emergency attendances	59,137.13 (26781.11)
OPATT	Total number of non-A&E outpatient attendances	190,243.08 (93857.06)
CATCHSH	% share of beds in total area (DoH (1989)) DP41	.83 (.28)
CATCHD	Catchment population of provider (DoH (1989))	356.44 (222.83)

Cost data were taken from CIPFA Healthcare data sets (CIPFA, 1995-1996) for 1994 and 1995. Cost data refer to those data reported in the annual accounts of NHS trust hospitals, including operating costs, capital costs and operating surplus and losses, and are reported in an annual form. Data for financial year 1992/93 were also provided by Ivan Csaba (due to the unavailability of these data from CIPFA). The variables and their definitions are presented in Table A3.4 below. These data were tied into the CHKS data set by CHKS to retain the anonymity of the hospitals. All costs were converted into 1995 prices using HCHS inflation indices.

A complete data set were obtained for 23, 48 and 66 hospitals for the financial years 1992/3, 1993/4 and 1994/5 respectively. However, the full sample used in analyses refers to the same 64 hospitals identified in the activity data. The number of observations for the cost data are fewer than for activity data due to lack of data availability over the three years. The sample therefore refers to 85 annual observational points.

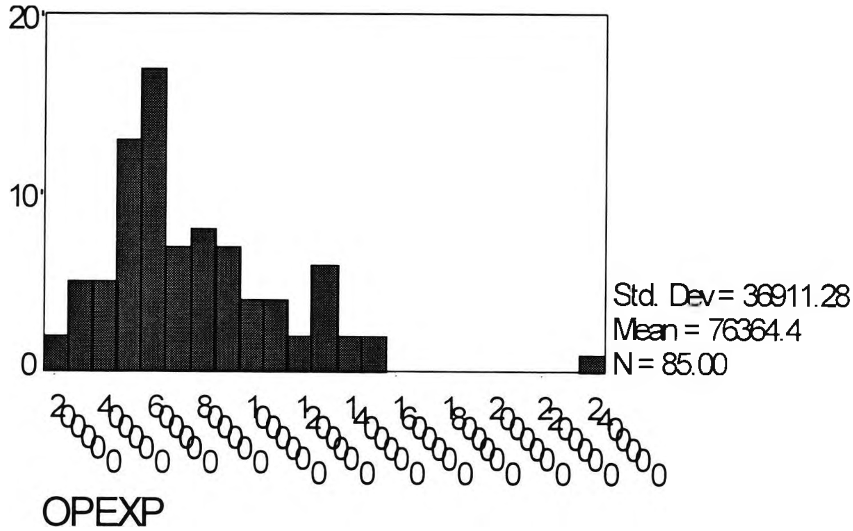
Table A3.4: Variable definitions

Variable name	**Definition**	**Mean (000) (S.D.)**
OPEXP	Operating expenses	76,364.40 (36,911.28)
OPERATING EXPENSES		
BOARD	Board members' fees	254.34 (164.14)
SAL	Staff costs	48805.02 (22049.00)
SUPCLIN	Supplies and services: Clinical	10567.62 (8191.61)
SUPGEN	Supplies and services: General	1715.91 (783.32)
ESTAB	Establishment	1774.09 (737.77)
TRANSEXP	Transport	208.83 (236.36)
PREM	Premises	5136.06 (2812.11)
BADDEBT	Bad debts	81.46 (194.71)
DEPREC	Depreciation and amorisation	3318.34 (2006.16)
AUDIT	Audit (including fees and other auditors' remuneration	98.27 (39.74)
OTHER	Other expenses including miscellaneous services from other NHS providers	4431.56 (2476.06)
TOTAL VARIABLE COST	OPEXP-(DEPREC+BADDEBT)	72,893.44 (35,174.58)
NUMBER OF EMPLOYEES		
MEDNUM	Medical and dental	216.84 (116.08)
NURSNUM	Nursing and midwifery	1149.55 (468.88)
PROFNUM	Professions allied to medicine	146.68 (66.23)
ANCILNUM	Ancillaries	315.58 (226.17)
ADMINNUM	Admin and clerical	446.20 (217.86)
WORKNUM	Works	45.20 (27.46)
TECHNUM	Other professional and technical	221.79 (148.36)

OTHNUM	All other staff	37.79 (46.74)
TOTNUM	Total number of employees	2579.64 (1185.02)

The sample represents total operating expenditure of £6,490,971,000 over the three year period. the range of operating expenses (in £000) covered in the sample is presented in Figure A2 below. The average operating expense is £76,364,400

Figure A3.2: Operating expenses



A3.2: Descriptive statistics

This section provides descriptive data for the variables included in the three analyses presented in the paper.

A3.2.1 Demand estimation (Chapters 4 and 5)

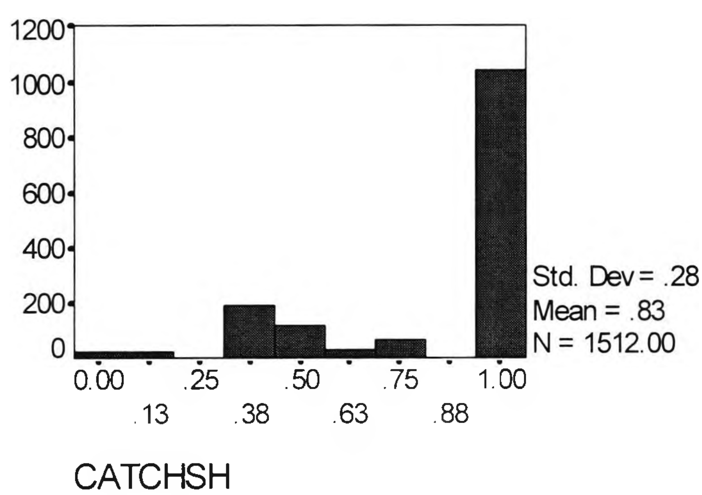
The following descriptive data refer to the demand equations estimated in Chapters 4 and 5.

Table A3.5: Demand equation(s)

Variable	Mean	Std Dev	N
CATCHSH	.83	.28	1512
CATCHD	356.44	222.83	1512
BEDS	975.00	365.42	1512
EMADM	2088.61	837.15	1512
ADC	425.22	155.51	1512

As Figure A3.3, below indicates, the majority of observations are taken from providers that operate in monopoly situations.

Figure A3.3: Catchment share



The distribution of emergency admissions is indicated in Figure A3.4 below. The maximum value is 4,903 the minimum 26.

Figure A3.4: Emergency admissions

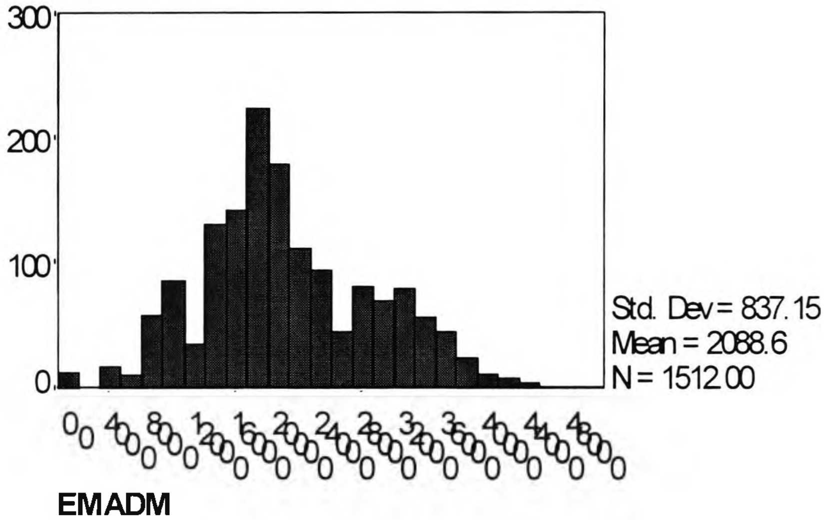


Figure A3.5: ADC

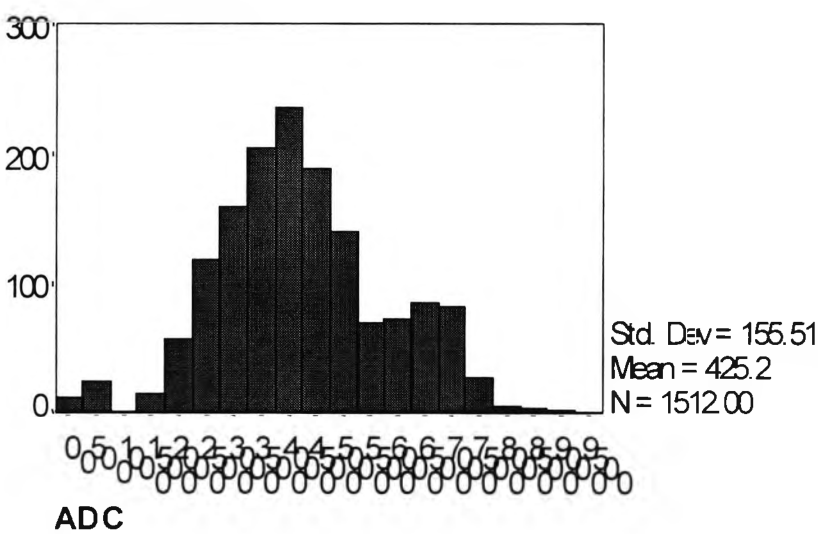
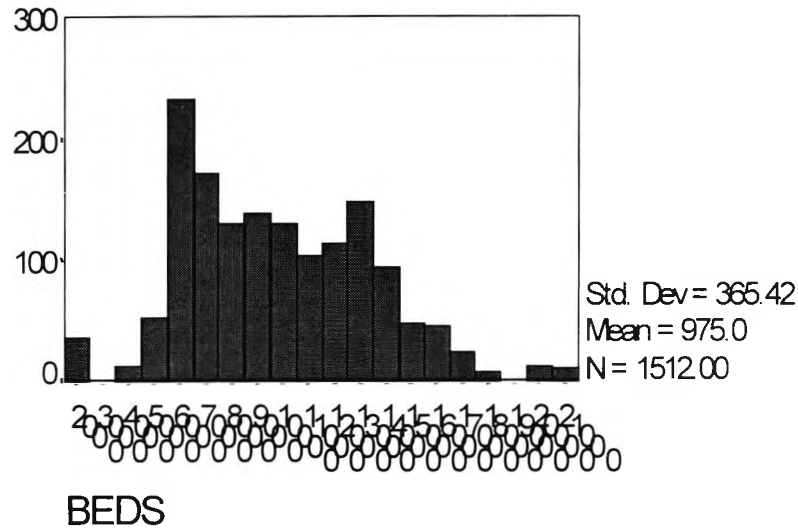


Figure A3.6: Beds



A 3.2.2: Cost estimation (Chapter 4)

The data in this section refer to data used in the cost function estimate presented in Chapter 4.

Table A3.6: Cost equation

Variable	Mean	Std Dev	N
CASEMIX	.69	.09	85
INVOCC	1.49	.65	85
WI	19061	2320.00	85
BEDEM	645.02	246.77	85
DAYATT	11110.58	12766.99	85
ELADM	24672.08	12916.20	85
EMADM	25026.98	10451.89	85
AEATT	59137.13	26781.11	85
OPATT	190243.08	93857.06	85
RES	1535.72	2551.63	85
SERES	102.78	89.93	85
TVC	72893446	35174582.40	85

The total number of emergency admissions that were unexpected was 130,536 out of a 2,127,293 total emergency admissions. The distribution of which is presented in Figure A3.7 below. The minimum residual in any one month was 327 with a maximum of 22,286, reflecting the different size of hospital and the relative accuracy of the forecasts.

Figure A3.7: RES

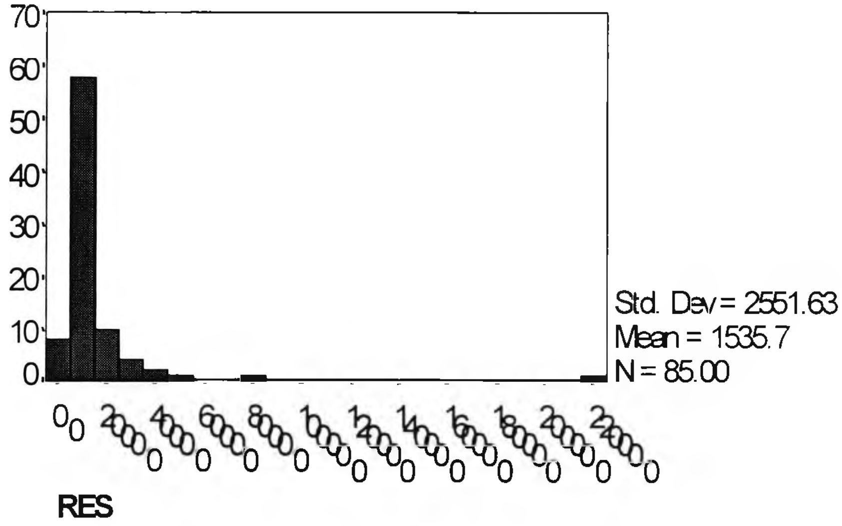


Figure A3.8: Beds allocated to emergency sector (BEDSEM)

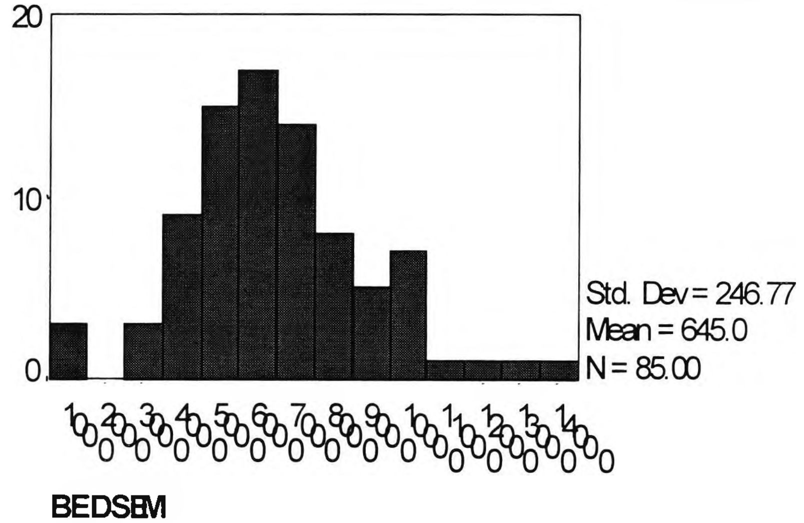


Figure A3.9: Emergency admissions

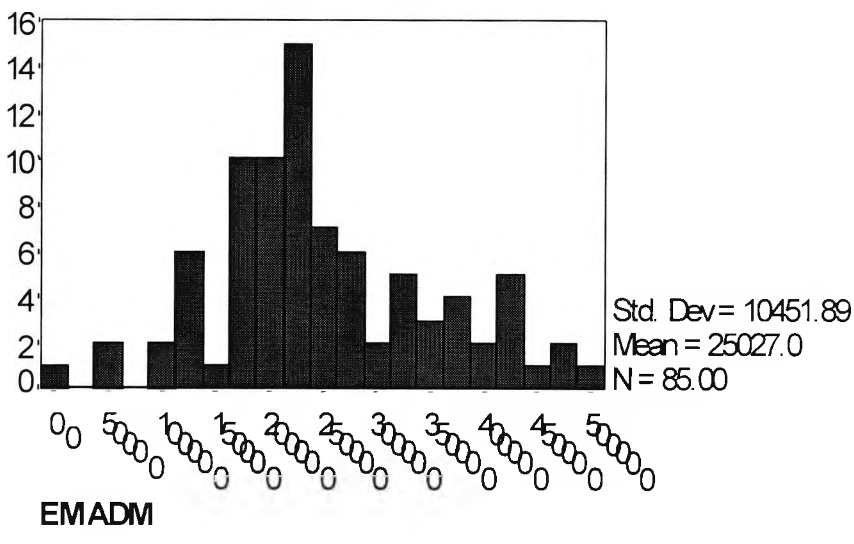


Figure A3.10: Elective admissions

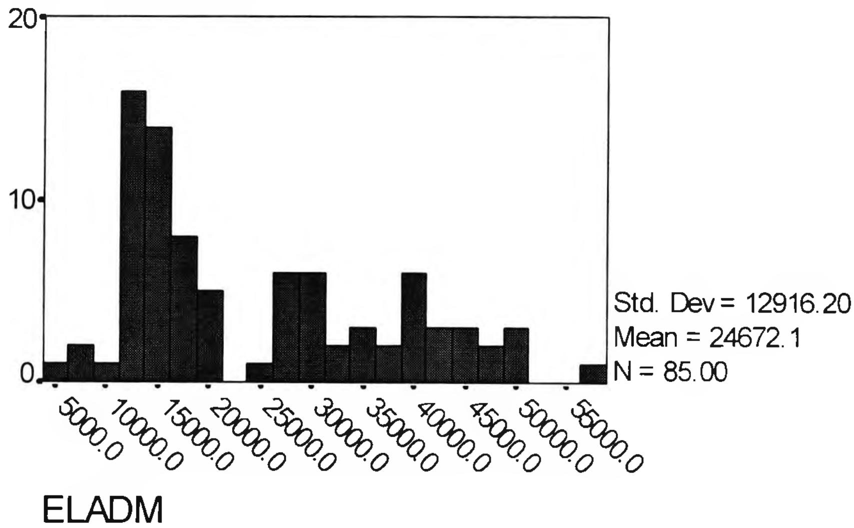
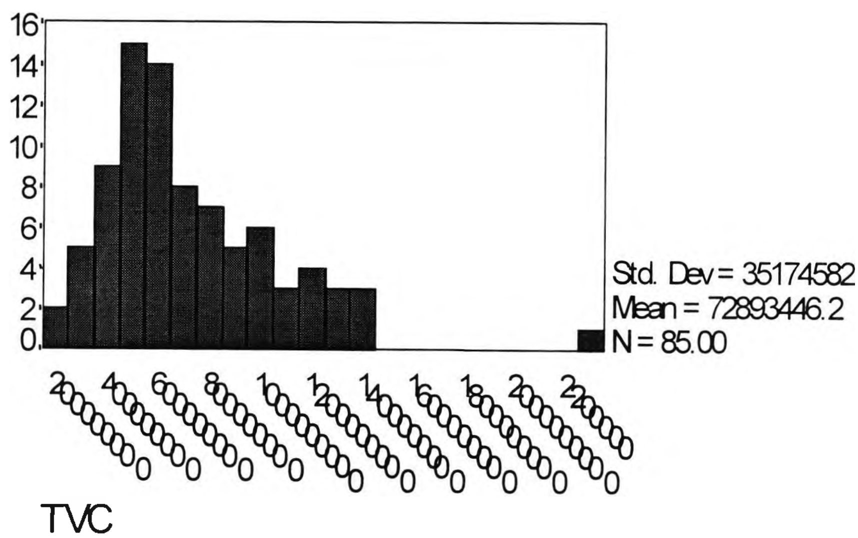


Figure A3.11: Total variable cost (TVC)



A 3.2.3: Social cost estimation (Chapter 5)

The data in this section refer to the partial analysis of probability of turnaway and waiting lists estimated in Chapter 5.

A3.2.3a Probit, Box-Cox and Heckman analyses (full sample)

Table A3.7: Social cost estimation (full sample)

Variable	Mean	Std Dev	N
DEM	425.22	155.51	1512
DEL	3681.58	1982.57	1512
PT	.04	.13	1512

Figure A3.12: Probability of turnaway

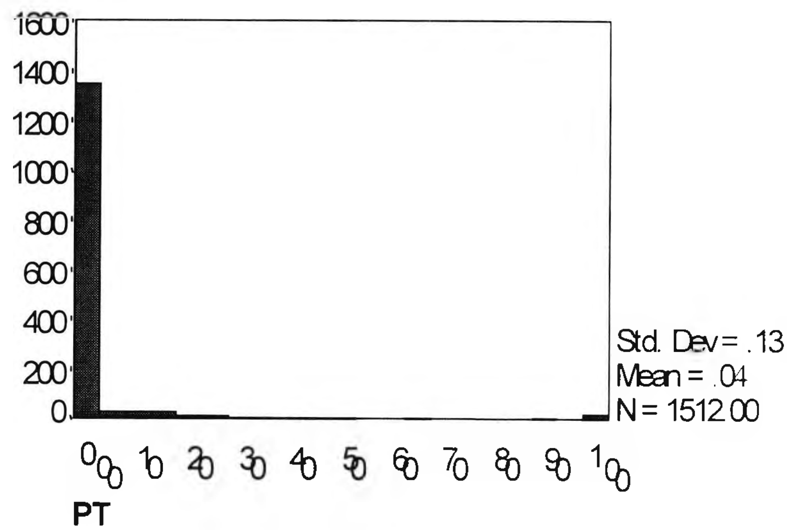


Figure A3.13: Emergency demand

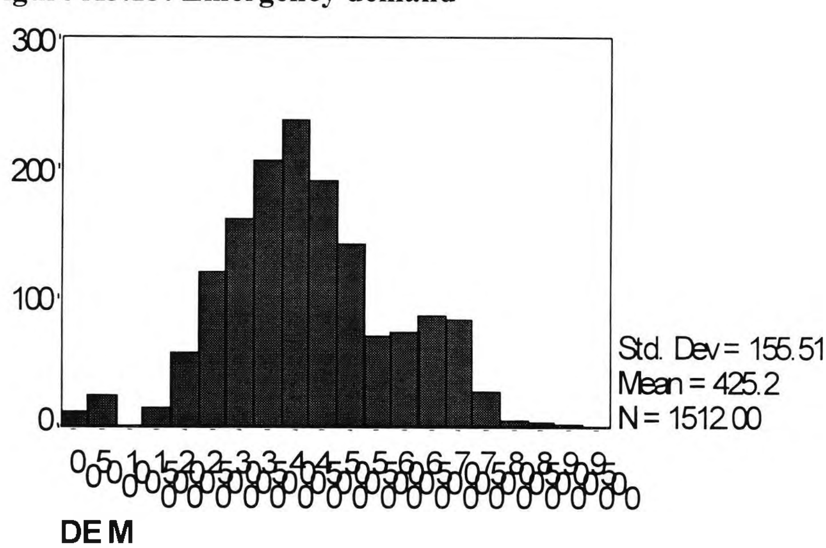
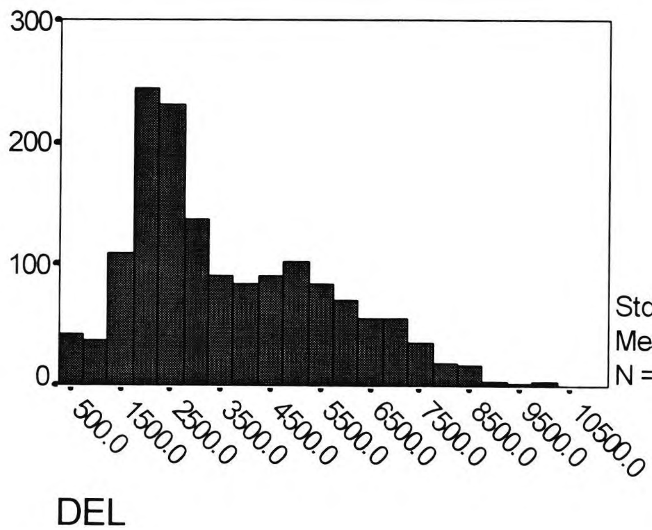


Figure A3.14: Total elective demand



A3.2.3b Box-Cox analyses (sample PT > 0)

Table A3.8: Social cost estimation (sample PT > 0)

Variable	Mean	Std Dev	N
PT	.06	.18	814
DEM	425.91	168.75	814
DEL	3771.81	2221.60	814

Figure A3.15: Probability of turnaway

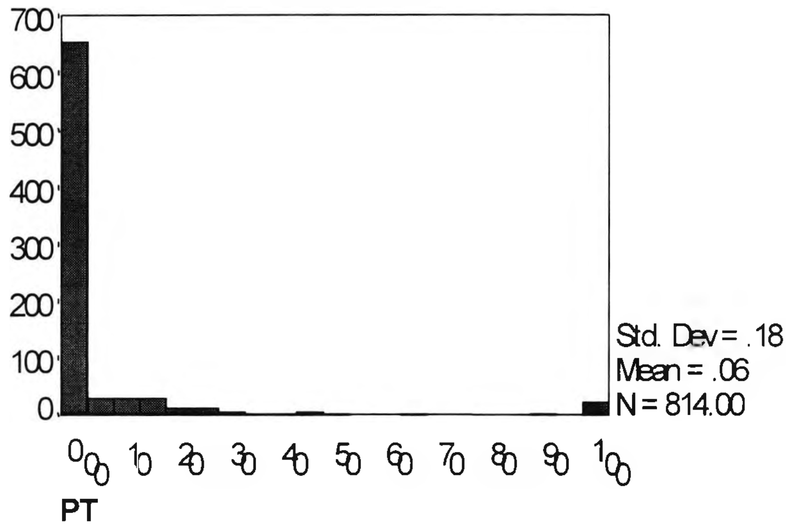


Figure A3.16: Total elective demand

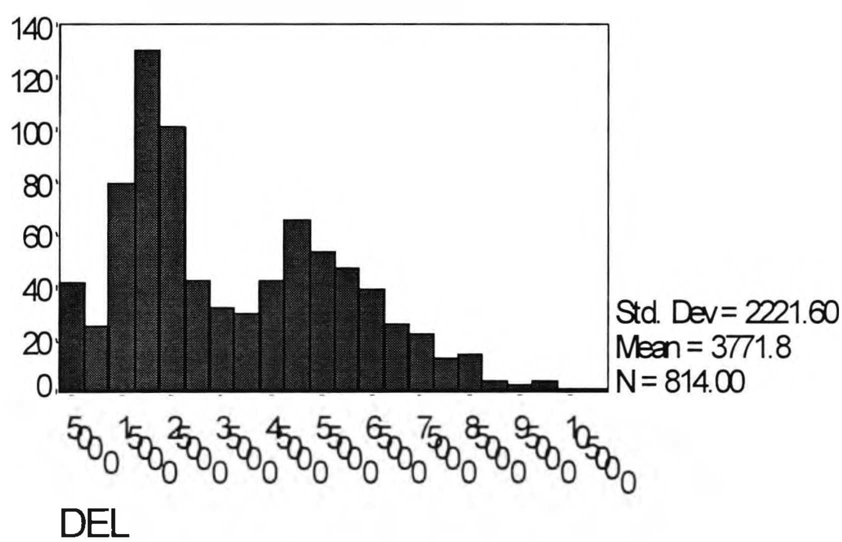
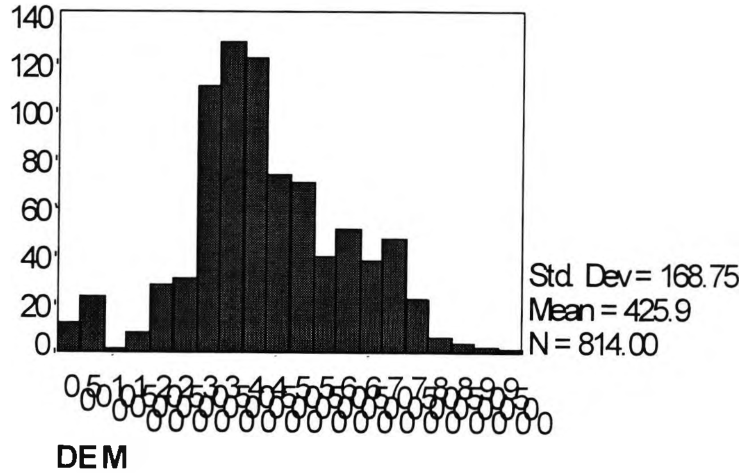


Figure A3.17: Emergency demand



REFERENCES

Aiginger K (1987). Production and decision theory under uncertainty. Blackwell Oxford.

Ben-Ner A (1986). Nonprofit organistaions: why do they exist in market economies? In Rose-Ackerman ed. The nonprofit sector: Economic theory and public policy. 94-113. OUP, Oxford.

Berki SE (1983). The design of case-based hospital payment systems. Medical Care XXI, 1:1-13.

Blumberg MS (1961). Distinctive patient facilities concept helps predict patient bed needs. Modern Hospital, 97:75-81.

Breusch T and Pagan A (1980). The LM test and its applications to model specification in econometrics. Review of Economic Studies, 47: 239-254.

Breyer F (1987). The specification of a hospital cost function: A comment on the recent literature. Journal of Health Economics, 6:147-157.

Carey K (1996). Stochastic demand for hospitals and optimising "excess" bed capacity. Discussion Paper, US Department of Veterans Affairs.

Caves D, Christensen L and Tretheway M, (1980). Flexible cost functions for multiproduct firms. Review of Economics and Statistics, 62:477-481.

Chiswick BR (1976). Hospital utilisation: An analysis of SMSA differences in occupancy rates, admissions rates, and bed rates in Expolorations in economic Research. Occassional Papers of the National Bureau of Economic Research vol3 (3) Summer 1976.

CHKS (1995). Acute Care 1994 Healthcare Resource Group National Statistics. CHKS Alcester.

CIPFA (1995 & 1996). Health Database 1994 & 1995. CIPFA, London.

Clarkson K (1972). Some implications of property rights in hospital management. Journal of Law and Economics, 15:363- 384.

Conrad R and Strauss R (1983). A multiple-output multiple-input model of the hospital industry in North Carolina. Applied Economics, 15(3):341-352.

Cowing T and Holtman A (1983). Multiproduct short-run hospital cost functions: Empirical evidence and policy implications from cross-section data. Southern Economic Journal, 49: 637-653.

Cowing T, Holtman A and Powers S (1983). Hospital cost analysis: A survey and evaluation of recent studies in R. Schleffler, ed. Advances in health economics and health service research, vol 4 JAI press, Greenwich, CT.

Csaba I (1997). Hospital costs, organsiational slack and quasi-markets – an empirical analysis of the British health reforms. Paper presented to HESG, Liverpool.

Cyert RM and March CG (1963). A behavioural theory of the firm. Prentice Hall USA.

Danzon PM (1982). Hospital 'profits': The effect of reimbursements policies, Journal of Health Economics, 1:29-52.

Davis K (1972). Economic theories of behaviour in non-profit private hospitals Economic and Business Bulletin, 24:1-13.

De Alessi A (1983). Property rights transaction costs and X-inefficiency: An essay in economic theory. American Economic Review, 73: 64-81.

Dor A and Farley D (1996). Payment source and the cost of hospital care: Evidence from a multiproduct cost function with multiple payers. Journal of Health Economics. 15:1-21.

Duncan GM (1990). The effect of probabilistic demands on the structure of hospital costs. Journal of Risk and Uncertainty, 3:211-220

Eby C and Cohodes DR (1985). What do we know about rate-setting? Journal of Health Politics, Policy and Law, Vol 10(2):299-326.

Ellis RP and McGuire TG (1986). Provider behavior under prospective reimbursement: cost sharing and supply. Journal of Health Economics, 5:129-152.

Ellis RP and McGuire TG (1990). Optimal payments systems for health services. Journal of Health Economics, 9:375-396.

Evans RG (1971). Behavioural cost functions. Canadian Journal of Economics, 4:198-215.

Farrell MJ (1957). The measurements of productive efficiency. Journal of the Royal Statistical Society, series A, 120:253-281.

Feldstein MS (1961). An empirical investigation of the marginal costs of hopsital services. Chicage: Graduate Program in Hospital Administration, University of Chicago.

Feldstein MS (1973). Planning Health Care. In Cooper MH and Culyer AJ eds. Health Economics, Penguin Books, London.

Feldstein MS (1967). Economic analysis for health service efficiency: Econometric studies of the British NHS. North-Holland, Amsterdam.

Freidman B and Pauly M (1983). A new approach to hospital cost functions and some issues in revenue regulation Health Care Financing Review, 4:105-114.

Gaynor M and Anderson G (1995). Uncertain demand, the structure of hospital costs and the cost of an empty bed, Journal of Health Economics, 14:291-317.

Ginsburg P (1970). 'Capital in non-profit hospitals' unpublished doctoral dissertation. Harvard Univ.

Granneman T, Brown R and Pauly M (1986). Estimating hospital costs: A multiple-output analysis. Journal of Health Economics 5:107-127.

Gray A (1982). The production of dental care in the British National Health Service. Scottish journal of Political Economy 29, 59-74.

Greene WH (1993). Econometric Analysis. Prentice-Hall, New Jersey.

Hansmann H (1980). The role of nonprofit enterprise. The Yale Law Journal, 89:835-98.

Hansmann H (1981). Nonprofit enterprise in the performing arts, The Bell Journal of Economics 12, 341-361.

Harris J (1977). The internal organisations of hospitals: Some economic implications, Bell Journal of Economics, 8:467-482.

Hausman J (1978). Specification tests in econometrics, Econometrica, 46:1251-1271.

Heckman J (1976). The common structure of statitsical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement, 5: 475-492.

Heckman J (1979). Sample selection bias as a specification error. Econometrica, 47:153-161.

Holtman AG (1988). Theories of non-proft institutions. Journal of Economic Surveys, 2(1):29-45.

Hornbrook M and Rafferty J (1982). The economics of hospital reimbursement. Advances in Health Economics and Health Services Research, vol 3 pp79-115. JAI press, Greenwich, CT.

Hsiao C (1986). Analysis of panel data. Econometric Society Monographs. CUP Cambridge.

Hughes D and McGuire A (1997). Production responses to uncertainty in the hospital sector: Some preliminary investigations. Paper presented to the HESG, Liverpool.

Jenkins (1980). Multiple cost analysis; Service and case-type cost equations for Ontario hospitals. Applied Economics, 12: 103-113.

Joseph H and Folland S (1972). Uncertainty and hospital costs. Southern Economic Journal, October, 267-273.

Joskow PJ (1980). The effects of competition and regulation on hospital bed supply and the reservation quality of the hospital. Bell Journal of Economics, 6:421-447.

Koutsoyiannis A (1979). Modern Microeconomics. Macmillan Press, London.

Lavers RJ and Whynes DK (1978). A production function analysis of English maternity hospitals. Socioeconomic Planning Sciences 12, 85-93.

Lee ML (1971). A conspicuous production theory of hospital behaviour. Southern Economic Journal, 38:48-58.

Liebenstein H (1966). Allocative efficiency versus X-efficiency. American Economic Review, 56:392-415.

Long (1964). Efficient use of hopsitals in Mushkin S J The economics of health and medical care. 211-226. AnnArbor Univ of Michigan Press.

Maddala GS (1983). Limited-dependent and qualitative variables in econometrics. CUP, Cambridge.

Maddala GS (1992). Introduction to Econometrics. Maxwell, New York.

McGuire A (1985). The theory of the hospital: A review of the models. Social Science and Medicine 20(11):1177-1184.

McGuire and Westoby (1983). A prodcution function analysis of acute hospitals. Discussion paper 04/83 Health Economics Research Unit, University of Aberdeen.

McGuire and Williams (1986) Information theory and Scottish hospital cost functions. Discussion paper 01/86 Health Economics Research Unit, University of Aberdeen.

Morrisey MA, Conrad DA, Shortell SM and KS Cook (1984). Hospital rate review a theory and empirical review, Journal of Health Economics, 3:25-47.

Mulligan JG (1985). The stochastic determinants of hospital bed supply. Journal of Health Economics, 4:177-185.

Newhouse JP (1970). Towards a theory of non-profit institutions: An economic model of a hospital. American Economic Review, 60:64-74.

Pauly MV (1987). Nonprofit firms in medical markets, Lesson From Health Economics. American Economic Association papers and proceedings, May. 257-262.

Pauly M and Redisch M (1973). The not-for-profit hospital as a physicians' Co-operative. American Economic Review, 63:87-100.

Pauly M and Wilson P (1986). Hospital output forecasts and the cost of empty beds. Health Services Research 21(3):403-428.

Quester A and Greene W (1982). A stae preference approach to wives labour supply. Social Science Quarterly, 63:16-27.

Rafferty JA (1971). Patterns of hospital use: An analysis of short-run variations. Journal of Political Economy, 79:154-165.

Reder M (1965). 'Economic theory and non-profit enterprise: Some problems in the economics of hospitals' American Economic Review Proceedings and Papers May 1965, 44:472-480.

Rice R (1966). An analysis of the hospital as an economic organisation. The Modern Hospital, 106:87-91.

Roemer MI (1960). How medical judgement affects hospital admissions. Modern Hospital, 94.

Shonick W (1970). A stochastic model for occupancy-related random variables in general-acute hospitals. Journal of the American Statistical Association, 65:1474-1500.

Shonick W (1972). Understanding the nature of the random fluctuations of hospital daily census: an important health planning tool. Medical Care, 10(2):118-142.

Shonick W and J R Jackson (1973). An improved stochastic model for occupancy-related variables in general-acute hospitals. Operations Research, 21:952-965.

Thorpe KE and Phelps CE (1990). Regulatory intensity and hospital cost growth. Journal of Health Economics, 9(2):143-166.

Vita M (1990). Exploring hospital production functions with flexible functional forms. Journal of Health Economics, 9:1-21.

Wagstaff (1989). Econometric studies in health economics: A survey of the British literature. Journal of Health Economics, 8:1-51.

Watt JM, Derzon RA, Renn SC, Scramm CJ, Hahn JS and Pillari GD (1986). The comparative economic performance of investor-owned chain and not-for-profit hospitals, New England Journal of Medicine, January 9, 314:89-96.

Weisbrod B A (1975). Towards a theory of the voluntary non-profit sector in a three sector economy. In Phelps E ed Altruism, Morality and Economic Theory, 171-195, New York Russell Sage Foundation.

White H (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity, Econometrica, 48:817-838.

Worthington DJ (1987). Queueing model for hospital waiting lists. Journal of Operational Research Society, 38(5):413-422.